Published: 17 March 2023
Online AM: 31 January 2023
Accepted: 28 November 2022
Revised: 25 November 2022
Received: 13 November 2021

RESEARCH-ARTICLE

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

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

DANA LI, Copenhagen University Hospital, Copenhagen, Hovedstaden, Denmark

XIANG DAI, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

JONATHAN FREDERIK CARLSEN, Copenhagen University Hospital, Copenhagen, Hovedstaden, Denmark

FINN KENSING, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

TARIQ OSMAN ANDERSEN, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

Open Access Support provided by:

University of Copenhagen

Copenhagen University Hospital

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

HUBERT D. ZAJĄC, University of Copenhagen, Denmark
DANA LI, Copenhagen University Hospital, Denmark
XIANG DAI, University of Copenhagen, Denmark
JONATHAN F. CARLSEN, Copenhagen University Hospital, Denmark
FINN KENSING and TARIQ O. ANDERSEN, University of Copenhagen, Denmark

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.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models;**

Additional Key Words and Phrases: Systematic review, machine learning, artificial intelligence, clinician-facing systems, health, real-world, implementation, conceptual framework

ACM Reference format:
Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 33 (March 2023), 39 pages.
https://doi.org/10.1145/3582430

1 INTRODUCTION

Artificial Intelligence (AI) is receiving growing attention from policymakers, scientists, and society at large, as well as massive public and private investments, suggesting a new "AI spring" [80].

Authors' addresses: H. D. Zając, X. Dai, F. Kensing, and T. O. Andersen, University of Copenhagen, Universitetsparken 1, Copenhagen, Denmark, 2100; emails: {hdz, xiang.dai, kensing, tariq}@di.ku.dk; D. Li and J. F. Carlsen, Copenhagen University Hospital, Blegdamsvej 9, Copenhagen, Denmark, 2100; emails: {dana.li, jonathan.frederik.carlsen}@regionh.dk.

1073-0516/2023/03-ART33
https://doi.org/10.1145/3582430

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 [23]. 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 [21] without being explicitly programmed [115]. Such algorithms, like deep learning, random forest, and others, have been shown to outperform statistical models when applied to health data (see, e.g., [94]). Multiple ML models for the detection and prediction of clinical outcomes have been validated in many disease areas, including diabetes [54], cancer [66], mental health [115], and heart disease [94]. 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 [90], reducing diagnostic and therapeutic errors [36, 48, 77, 95], and obviating repetitive clinical tasks [43].

Despite the promising outlook, only few medical devices based on ML have been regulatory approved [87], indicating that clinician-facing ML-based systems are particularly challenging to realise in clinical practice [32, 68]. Going the "last mile" of medical AI innovation is afflicted by many challenges compared to conventional systems [32]. Unlike rule-based algorithms, whose decision-making processes can be inspected, ML models are often criticised for the "black-box" problem [27], i.e., obscuring the reasoning behind a model output [43]. ML requires also large amounts of labelled training data [21, 118], 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 [60].

In **Human-Computer Interaction (HCI)**, Human-AI interaction has been of key interest for more than 20 years [3]. 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 [105, 108]. HCI contributions have centred on tackling Human-AI interaction issues by developing conceptual frameworks [117], models and principles [30, 39, 75], methods and techniques [29, 71, 120], and design guidelines [5] for ML-based systems, as well as undertaking experiments with users to empirically explore the challenges of human and AI engagements [61].

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., Reference [124]) in comparison to traditional systems, especially in safety critical and complex settings like healthcare [6, 49, 87]. Challenges arise on various levels, for example, designers, or HCI researchers and practitioners, find it hard to understand the capabilities of ML [38, 124]; interdisciplinary collaborations between designers, ML engineers, and clinical domain experts require more work than usual [123]; early feedback for iteration is often unavailable in the development phase [29, 125]; and the unpredictable outcomes of ML models or "imperfect algorithms" make it difficult to obtain purposeful and trustful interfaces for prototyping ML [71, 124]. 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 [15, 50, 119, 124] and the lack of mutual understanding, shared conceptual frameworks, and well-established interdisciplinary methods and techniques [15, 50, 119, 124].

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* [6, 9, 32, 41, 61, 82, 105,

115]. In their systematic review of HCI literature on ML in mental health [115], 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" [9] and how an ML model with high accuracy in the lab was heavily impacted in practice by "socio-environmental" factors such as clinical workflows and patient experience [9]. 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"* [41].

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 article, 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" [100]. We posed six **research questions (RQs)** 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. Last, 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.

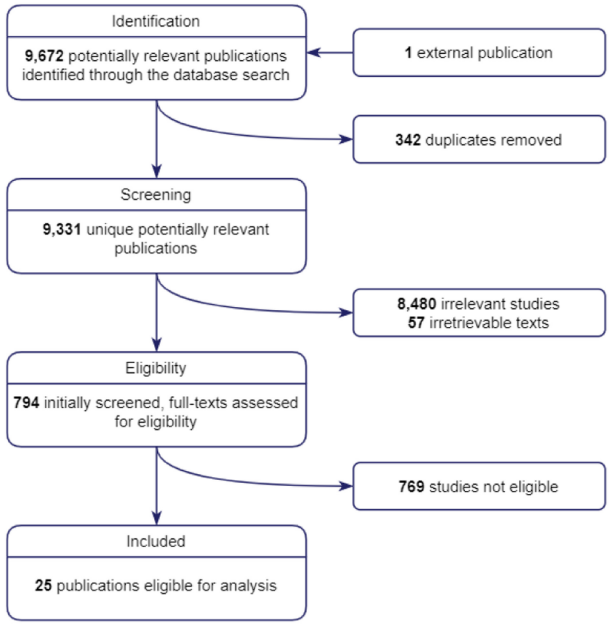

Fig. 1. A PRISMA flow diagram of the literature search and study selection.

2 MATERIALS AND METHODS

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 = 4,109, PubMed = 5,561, arXiv = 37). We included one external publication that was not returned by the queries but was known to the authors of this article. It was submitted to JMIR in January 2021 and accepted in October 2021.

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 [92] to conduct the screening process, which comprised three phases. A flow diagram of **Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)** [93] of the process can be seen in Figure 1. The selected articles had to meet six inclusion criteria described in Table 1. Ultimately, 25 publications were selected for the review.

2.3 Data Extraction and Synthesis

In our analysis, we strove to follow qualitative systematic review principles [20, 37]. 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

Table 1. 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 [74, 115]. 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."* [115]
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 [116], perspectives on interaction design in the context of clinical use [24], and investigation of barriers to successful clinical use [96].	Studies that focused on an ML-based system outside of the scope of clinical use, e.g., solely in a patient's home [35].
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.

analysis of several relevant articles using the constructing grounded theory approach [28] 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 such as "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

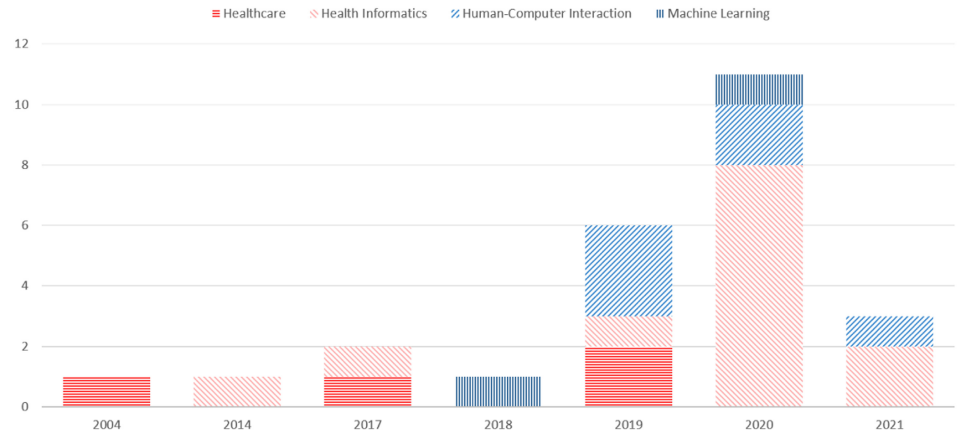

Fig. 2. A distribution of the included articles grouped by publish year and highlighted domains.

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.

3 RESULTS

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 2.

The 25 included articles describe 23 distinct clinician-facing ML-based systems, as shown in Table 2. The only system described in more than one publication that met our criteria was Sepsis Watch [103, 106, 107]. 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 five systems were classified as targeting interdisciplinary settings.

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. Four 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 three main *intended uses* of ML in a medical

Table 2. Overview of Included Studies

Author	Medical speciality	Intended use	ML output	System status	Authors' evaluation
Barda et al. [7]	Paediatrics	Prioritisation Decision support	Prediction (Overall mortality risk)	Pre-deployment Prototype	Positive responses
Baxter et al. [8]	Internal medicine	Prioritisation	Prediction (Overall unplanned readmission risk)	In clinical use Operational system	Ineffective adoption
Beede et al. [9]	Ophthalmology	Automation (Issuing referrals)	Detection and severity rating (diabetic retinopathy)	Deployment Operational system	Ineffective adoption
Benda et al. [10]	Interdisciplinary	Prioritisation	Prediction (potential cost and care needs)	Pre-deployment Mock-up	N/A
Brennan et al. [19]	Surgery	Decision support (Risk assessment)	Prediction (Eight types of postoperative complications)	Deployment Prototype	Mixed responses
Cai et al. [25]	Laboratory medicine Oncology	Decision support (Diagnosis)	Detection and severity rating (Prostate cancer)	Pre-deployment Prototype	N/A
Cai et al. [24]	Laboratory medicine Oncology	Decision support (Case-based reasoning)	Similar historical cases (Prostate cancer)	Pre-deployment Prototype	Positive responses
Cho et al. [31]	Clinical care	Prioritisation Decision support (Prevention of adverse effects)	Prediction (Falling risk) Intervention recommendations	In clinical use Operational system	Mixed adoption
Gastounioti et al. [45]	Vascular medicine	Decision support (Diagnosis)	Prediction (Atherosclerosis risk) Similar historical cases	Deployment Operational system	N/A
Ginestra et al. [46]	Emergency medicine	Prioritisation	Prediction (Sepsis risk)	In clinical use Operational system	Ineffective adoption
Gu et al. [52]	Oncology	Decision support (Case-based reasoning)	Similar historical cases (Breast cancer)	In clinical use Operational system	Positive responses
Hollander et al. [59]	Cardiology	Decision support (Diagnosis)	Prediction (Acute coronary syndrome risk and acute myocardial infraction)	In clinical use Operational system	Ineffective adoption
Jauk et al. [63]	Interdisciplinary	Prioritisation	Prediction (Delirium risk)	In clinical use Operational system	Ineffective adoption
Jin et al. [64]	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. [82]	Cardiology	Prioritisation Decision support	Prediction (Ventricular tachycardia and ventricular fibrillation)	Pre-deployment Mock-up	Mixed responses
McCoy et al. [83]	Emergency medicine	Prioritisation	Prediction (Sepsis risk)	Deployment Operational system	Successful adoption
Morrison et al. [86]	Neurology	Decision support (Diagnosis)	Detection and severity rating (Multiple Sclerosis)	Pre-deployment Prototype	Positive responses
Petitgand et al. [96]	Emergency medicine	Decision support	Compilation of relevant medical information	Deployment Operational system	Ineffective adoption
Romero-Brufau et al. [102]	Interdisciplinary	Prioritisation	Prediction (Overall unplanned readmission risk)	Deployment Operational system	Successful adoption

(Continued)

Table 2. Continued

Author	Medical speciality	Intended use	ML output	System status	Authors' evaluation
Romero-Brufau et al. [101]	Internal medicine	Prioritisation Decision support	Prediction (Diabetes risk) Intervention recommendations	In clinical use Operational system	Ineffective adoption
Sandhu et al. [103]	Emergency medicine	Prioritisation Decision support	Prediction (Sepsis risk)	In clinical use Operational system	Successful adoption
Sendak et al. [106]	Emergency medicine	Prioritisation Decision support	Prediction (Sepsis risk)	Deployment Operational system	Successful adoption
Sendak et al. [107]	Emergency medicine	Prioritisation Decision support	Prediction (Sepsis risk)	Deployment Operational system	Successful adoption
Wang et al. [116]	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. [125]	Cardiology Surgery	Decision support	Prediction (Postoperative lifespan)	Pre-deployment Mock-up	Mixed responses

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.

setting: (1) decision support, (2) prioritisation, and (3) automation of tasks. Many of the systems supported more than one *intended use*. Eighteen systems focused on decision support, 9 systems utilised ML for prioritisation, and 1 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. [62] outlined several steps to increase adoption. Pre-deployment evaluations received better responses, on average. No system received solely negative responses. Five out of 10 systems received positive responses, and 2 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 [52], time spent on a task [45], sensitivity and specificity [52, 102], area under the receiver operating characteristic curve [19, 52], and **Positive Predictive Value (PPV)** [102], i.e., the probability that disease prediction corresponds to a patient having that disease.

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"* [64] and augmented their current clinical skills while providing previously unavailable resources [24, 86]. In some cases, the lack of such additional information reduced trust in the predictions [116, 125]. Similarly, Morrison et al. [86] 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 [106, 107] or other technical metrics, e.g., with high PPV, even a relatively low sensitivity (67%) was deemed satisfactory [102]. Baxter et al. [8] 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. [82], 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 [106, 107].

Last, *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 [9]. However, in contrast to *prioritisation* and *decision support*, the cost of low negative predictive value could have been even higher. After the pilot, if the system missed a case of retinopathy, the patient would miss a potentially sight-saving referral.

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 [124]. 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 3.

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. [98] 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"* [64]. A final challenge when using structured EHR data is availability. In comparison to other data types such as image or text, available training data is sparse, which hinders progress in developing the ML algorithms that rely on it.

Six systems utilised image data ranging from single retina images [9] to videos from a depth camera [86]. This data type is characterised by objectivity and format consistency between patients, however, the quality of the input may vary [9]. Thanks to the wide availability of training data, ML algorithms that use image data are relatively advanced and offer high performance.

Table 3. Overview of Machine Learning Algorithms, Used Data, and ML Development Approach

Data	Structured (EHR data)	Barda et al. [7], Baxter et al. [8], Cho et al. [31], Gastounioti et al. [45], Ginestra et al. [46], Gu et al. [52], Hollander et al. [59], Jauk et al. [63], Jin et al. [64], McCoy et al. [83], Romero-Brufau et al. [101], Romero-Brufau et al. [102], Sandhu et al. [103], Sendak et al. [106], Sendak et al. [107], Wang et al. [116], Yang et al. [125]
	Image (retina image, ultrasound image, mammogram, physical examination recording, biopsy image)	Beede et al. [9], Cai et al. [24], Cai et al. [25], Gastounioti et al. [45], Gu et al. [52], Morrison et al. [86]
	Text (sociodemographic data, insurance claims, patient's self-reported data)	Romero-Brufau et al. [101], Romero-Brufau et al. [102], Benda et al. [10], Petitgand et al. [96]
	Time series (defibrillator implants transmissions)	Matthiesen et al. [82]
Machine learning	Knowledge-driven (random forest, SVM)	Barda et al. [7], Ginestra et al. [46], Jauk et al. [63], Matthiesen et al. [82], Morrison et al. [86], Romero-Brufau et al. [101], Gastounioti et al. [45]
	Data-driven (deep learning, neural network)	Beede et al. [9], Cai et al. [25], Cai et al. [24], Jin et al. [64], Petitgand et al. [96], Sandhu et al. [103], Sendak et al. [106], Sendak et al. [107], Hollander et al. [59]
	Implementation dependent (Bayesian network)	Cho et al. [31]
	Other (genetic algorithm, custom classifier, n/a)	Gu et al. [52], Baxter et al. [8], Brennan et al. [19], Romero-Brufau et al. [102], Wang et al. [116], Yang et al. [125], Benda et al. [10], McCoy et al. [83]
Integration method	Standalone application (Web application, PC program)	Beede et al. [9], Brennan et al. [19], Cai et al. [25], Gaastounioti et al. [45], Jin et al. [64], Morrison et al. [86], Romero-Brufau et al. [101], Romero-Brufau et al. [102], Sandhu et al. [103], Sendak et al. [106], Sendak et al. [107]
	EHR	Barda et al. [7], Baxter et al. [8], Cho et al. [31], Ginestra et al. [46], Jauk et al. [63], Wang et al. [116]
	Other (Printouts, phone call)	Hollander et al. [59], Matthiesen et al. [82], McCoy et al. [83], Petitgand et al. [96], Yang et al. [125]
	N/A	Benda et al. [10], Cai et al. [24], Gu et al. [52]
ML development approach	Cohesive	Barda et al. [7], Ginestra et al. [46], Gu et al. [52], Jin et al. [64], Matthiesen et al. [82], Sandhu et al. [103], Sendak et al. [106], Sendak et al. [107]
	Discrete	Beede et al. [9], Benda et al. [10], Brennan et al. [19], Cai et al. [25], Hollander et al. [59], Jauk et al. [63], Morrison et al. [86]
	Third-party	Baxter et al. [8], McCoy et al. [83], Romero-Brufau et al. [101], Romero-Brufau et al. [102]
	N/A	Cai et al. [24], Cho et al. [31], Petitgand et al. [96], Wang et al. [116], Yang et al. [125]

Five 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 [58]. 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) [10, 101, 102]. 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. [96], who analysed patients' self-reported data.

Last, 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 [82].

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 [44, 121]. 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 [111], 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"* [82].

Seven 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 [76]. Due to this, they require much larger datasets 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 [110]. 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. [106], *"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."*

Nine 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 [31]) or the lack of technical description beyond claims about using machine learning.

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"* [10]. 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. Nine 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. [96] and Yang et al. [125] 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 [83]. In a few instances, the authors did not report in detail on the integration method.

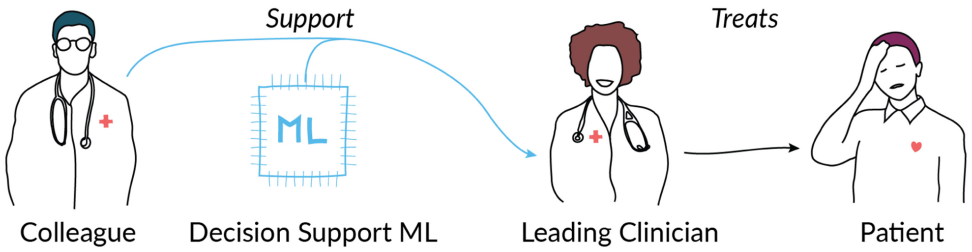

Fig. 3. A clinician-facing ML-based system supports clinicians at work.

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.

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 [87]. 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.

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 3). 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.

Alternative outlook denotes use cases where ML adds to the understanding of a patient's condition. We distinguished five different sub-types 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 4.

3.3.2 Prioritisation. An ML-based system designed for prioritisation assists healthcare providers within the context of multiple patients (Figure 4). It conducts an individual assessment of all the relevant patients according to predefined criteria. The outcome of such assessment is

Table 4. 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. [24] 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. [45], Hollander et al. [59], and Morrison et al. [86]. In their scenarios, physicians received the probability of specific diagnoses the systems were designed to detect. Jin et al. [64] and Wang et al. [116] 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 [64].
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 [25, 52, 64]. Second, Cai et al. [24] 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. [86] 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. [19], Jin et al. [64], Yang et al. [125] 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 [19, 125] or could be triggered by selecting an intervention to see its predicted impact on a patient [64].
ML provides intervention recommendations
Among the reviewed applications, we discovered two types of recommendations: preventive interventions and treatments. Benda et al. [10], Cho and Jin [31], and Romero-Brufau et al. [101] 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 [10]. Similarly, Jin et al. [64] and Wang et al. [116] evaluated systems that suggested treatments to described conditions.
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 [64, 116]. 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 [46]) or a high degree of subjectivity in the existing ones (multiple sclerosis [86]). 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 [64]. 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 [24], and pre-screening of patients to highlight historical information relevant to the current condition [96].

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 5).

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 5). Among the reviewed articles, only one system was described as being delegated such autonomy. Beede et al. [9] 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

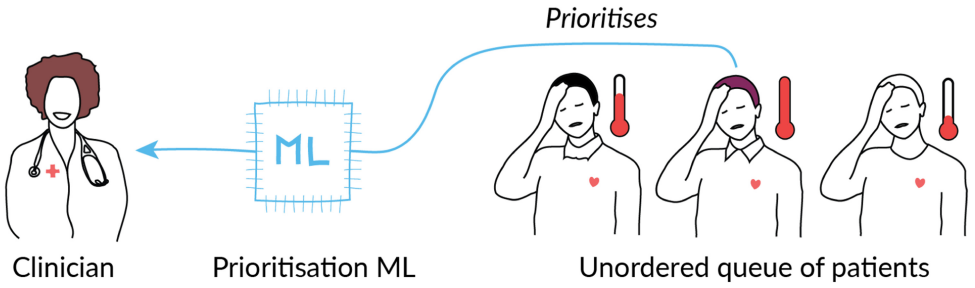

Fig. 4. A clinician-facing ML-based system conducts prioritisation of patients requiring attention.

Table 5. Descriptions and Examples of the Intended Use: Prioritisation

Prioritisation
Only prioritisation
Baxter et al. [8] 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 [46, 83]. In both cases, the attending staff received a notification and proceeded with standard examinations and care. Benda et al. [10] 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 [7, 62, 82, 103, 106, 107]. 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 [31] and Romero-Brufau et al. [101] 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 [102].

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.

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

First, ML served as a discussion catalyst and communication booster [46], 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 [10]. 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 [52, 101, 102]. Benda et al. [10] 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 [9, 46, 125] or when trying to convince a patient [9, 82].

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-

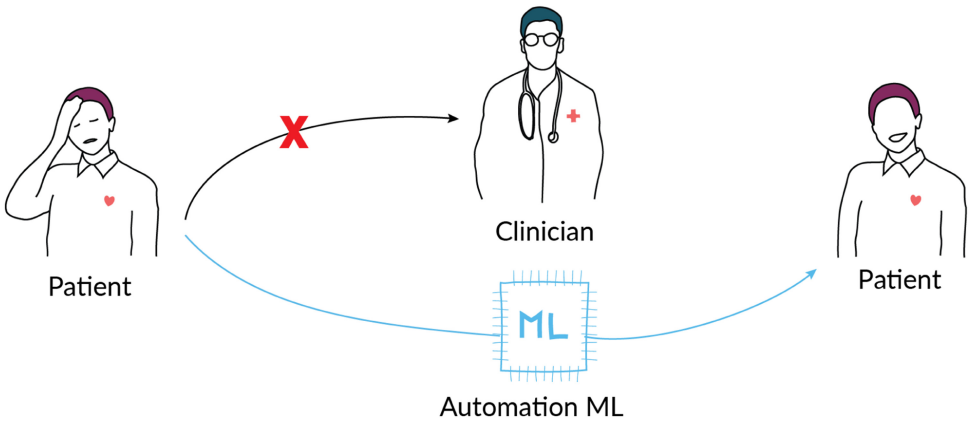

Fig. 5. A clinician-facing ML-based system is responsible for a single task of the healthcare delivery process.

experienced clinicians [52, 64, 116]. Additionally, inexperienced medical professionals used ML's prioritisation and contributing factors to develop expertise and learn about the focal conditions [7, 31, 82, 107]. Clinicians also hypothesised that exploiting new forms of data visualisation provided by the ML-based system could have research applications [86].

Third, the authors reported increased vigilance of medical professionals. In two of the studies targeting sepsis risk prediction [46, 103], clinicians argued that the presence of an ML-based prioritisation system made them more aware of the risk [106] and increased their monitoring [46].

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 6) of the studies' reported goals and stakeholders (divided into six professional groups). We specified evaluations conducted in the wild as a distinct study goal 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 [10] or gathering pre-deployment information needs [25]. Second, six studies described an innovation process, i.e., covering some steps of design, development, and deployment and bringing their system into the wild [9, 62, 83, 102, 106, 107].

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.

Table 6. Overview of Study Goals (Divided between Confined (C) and Innovation (I) Types) and Involved Stakeholders

Author		Goal of the study	P	N	M	NM	O	IT
Barda et al. [8]	C	Designing AI explanations	12	8				
Baxter et al. [9]	C	Identifying barriers to AI utilisation	7	3	3	2		
Beede et al. [10]	I	System deployment and observation		18		1		
Benda et al. [11]	C	Identifying facilitators, challenges, and recommendations		15		25		7
Brennan et al. [20]	C	Prospective evaluation of usability and accuracy	20					
Cai et al. [26]	C	Gathering pre-deployment information needs	12					
Cai et al. [25]	C	Designing interactions	21					
Cho et al. [33]	C	Prospective qualitative evaluation		18				
Gastounioti et al. [47]	C	Design, development, deployment	5					
Ginestra et al. [48]	C	Post deployment evaluation	44	43				
Gu et al. [53]	C	Development	12					
Hollander et al. [60]	C	Pre-post deployment evaluation	27					
Jauk et al. [63]	I	Development and deployment	5	5				5
Jin et al. [65]	C	Design and development	11				1	
Matthiesen et al. [84]	C	Near-live feasibility and qualitative evaluation	8					
McCoy et al. [85]	I	Development and deployment	x			x		
Morrison et al. [88]	C	Visualising AI output	5					
Petitgand et al. [98]	C	Identifying post-deployment adoption barriers	7	3		5		5
Romero-Brufau et al. [104]	C	Pre-post deployment evaluation	x	x	x			
Romero-Brufau et al. [103]	I	Design, development, and deployment	7	21	3		6	
Sandhu et al. [105]	C	Investigating factors influencing integration	7	8				
Sendak et al. [108]	I	Design, development, and deployment	x	x		x	x	x
Sendak et al. [109]	I	Design, development, and deployment	x	x		x	x	x
Wang et al. [119]	C	Post deployment evaluation	22					
Yang et al. [128]	C	Design	23	x				

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.

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 [9, 101, 102]. Non-medical professionals served as administrative workers—secretaries, clerks, administration [83], managers and board members [10]. Other stakeholders were people who did not fit into any other category, e.g., statisticians [107], designers [64, 96, 106, 107], or self-reported as "other position" [101]. Last, some of the studies involved IT specialists—hospital's information officers [10], ML engineers [62], other professionals [62, 106], developers [96, 106], data scientists, and engineers [107]. The diversity of involved stakeholders spotlights the breadth of medical ML innovation projects and foreshadows potential challenges that may arise from it.

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 7.

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"* [102]. 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"* [24] vs. *"requirements ... specified in collaboration"* [45], which were reported in two confined studies, in the *defining needs* and *requirements* activity. Interviews were the most used technique—reported in seven 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.

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 6). 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.

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*

Table 7. Ten Synthesised Activities of Medical ML Innovation Processes Derived from Both Innovation and Confined Studies

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" [10] "discussed challenges ... when making diagnosis" [64] "understand complexities in addressing the problem" [107] *In-situ observation:* "Observe frontline staff in clinical settings where the problem occurs" [106] *Undefined collaboration:* "assembling [a] team around the problem," "front-line physicians [...] work with a local innovation team to improve detection and treatment of sepsis" [106] *Data analysis:* "curating the data to better characterize the problem" [106, 107]
Involvement	*End-users:* Physicians [10, 64, 106, 107], Nurses [106] *Other stakeholders:* IT specialists [10, 106, 107], Non-medical Professionals [10]
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" [8], "15 hours of interviews" [9], "understand how the decision making process ... unfolds" [125] (in Reference [126]) *In-situ observation:* " to ... understand the eye screening process" [9], "to understand the clinical workload" [82], "user-centered observations, workflow-mapping ... were conducted to determine how best to incorporate the tool into existing workflows" [101], "how an implant decision is reached across many clinician roles and contexts" [125] (described in Reference [126]) *Undefined collaboration:* "collaborated ... to map local workflows" [102], "working with front-line clinicians to understand ... the care delivery process" [107] *Undefined work:* "assessments ... to determine how best to incorporate the tool into existing workflows" [101]
Involvement	*End-users:* Physicians [8, 82, 101, 107, 125], Nurses [8, 9, 101, 107], Other Medical Professionals [8, 101], Non-medical Professionals [8] *Other stakeholders:* Undefined [102]
Defining needs and requirements	
Goal	Devising high-level requirements for the new ML system. Clarifying assumptions and shaping the future work direction.
Techniques	*Interview:* "to understand pathologists' needs" [24] "they expressed a desire for ... they wished the tool to..." [64] *Mock-up experimenting:* "paper prototypes ... to understand pathologists' needs" [24] *Workshop:* "to understand the information ... they ... use in their decision-making.," "[rapidly sketch their ideas] ... to discuss and clarify requirements." [86] *Undefined collaboration:* "requirements ... specified in collaboration" [45], "invested ... effort into gathering requirements" [106]
Involvement	*End-users:* Physicians [24, 45, 64, 86], Nurses [106] *Other stakeholders:* Physicians, Other Medical Professionals, Other, IT specialists [106]
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" [7], "repeatedly met with ... nurses to iterate on functions, information, control, and visual components of the design" [106], "co-design ... focusing on ... sketching the user interface" [82] *Prototyping:* "created ... functional prototypes and iterated on [them] with further feedback" [24], "an interactive ... prototype ... was demonstrated to ... refine the initial requirements" [64], "presenting the design ... opportunity to try the app" [86] *Design feedback:* "present a set of visualisations" [86], "iterated on the design based on feedback" [125] *Interview:* "we informally interviewed nine neurologists and asked them to discuss three potential visualisations" [86] *Survey:* "a questionnaire to indicate preferred design options" [7]
Involvement	*End-users:* Physicians [7, 24, 64, 82, 86, 125], Nurses [7, 106]

(Continued)

Table 7. Continued

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" [103], "collaboration ... to finalize workflow decisions" [106], "collaborated ... to identify how to ... integrate the tool" [102], "a transdisciplinary team ... designed the ... workflow" [107] *Team decision:* "To situate a DST into the current VAD decision-making routine ... we chose the ... meetings" [125]
Involvement	*End-users:* Physicians [102, 125], Nurses [103, 106, 107] *Other stakeholders:* Physicians, Non-medical professionals, IT specialists [103, 107]
ML model development	
Goal	Sensitising development team to domain knowledge and informing the ML model development.
Techniques	*Interview:* "probes regarding ... the model including external rules and regulations, internal organization, clinical content" [10], "to enumerate diagnostically important concepts" [24], *Workshop:* "Feature engineering was carried out ... during five co-design workshops" [82] *Focus groups:* "feedback ... reported to the ... developers during focus groups" [96] *Previous implementations:* "sensitivity ... was determined a priori and informed by our experience with EWS 1.0" [46] *Delphi method:* "attribute weights elicited from experts using a Delphi method" [52] *Data-focused collaboration:* "[data-generating examinations] were chosen by our MS expert clinicians" [86] (in Reference [72]), "Clinical experts specified and reviewed ... all [the used] data" [107], "the department of nursing [requested removing a data point]" [31]
Involvement	*End-users:* Physicians [10, 24, 52, 82, 86, 96, 107], Nurses [31, 96, 107] *Other stakeholders:* Non-medical professionals, IT specialists [10]
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" [19], "comments ... during ... expert group meetings before and during the pilot", "the expert group suggested improvements ... and new functionalities ..." [62] *Iterative development:* "cycles to evaluate processes and incorporate clinical feedback" [83], "iterations that explore the best way to communicate visually sensed data" [86], "feedback loops were crucial to improving Sepsis Watch" [107], "reviewed multiple versions of the user interface", "clinicians specified the most relevant information to accompany the risk level" [106] *Case study:* "to validate ... the refinement mechanisms", "how SMILY ... affects user experience and search practices during a medical task" [25], "provide evidence [of] the usefulness of the system" [64] *Interview:* "usefulness, ease of use, general pros and cons of the prototype system, visualization designs, and insights" [64] *Survey:* "regarding the usability of the algorithm and web interface" [19]
Involvement	*End-users:* Physicians [19, 25, 62, 64, 83, 86], Nurses [62, 107] *Other stakeholders:* Non-medical professionals [83]
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.
Techniques	*Training session:* "training necessary for effective human-AI collaboration" [25], "guidance on how to use [the system]" [31], "various training sessions" [62], "education sessions to train ... on the proper uses" [83], "in the program workflow and application" [103], "one-on-one training sessions" [106] *Training materials:* "preparing training materials" [102], "collaboration ... to develop training material", "A 'Model Facts' sheet" [106] *Promotion:* "via emails" [46], "promoted the application throughout all participating departments" [62], "in faculty meetings and via email" [103, 106] *Undisclosed training:* "training staff" [102], "instructed to consider the list of recommendations" [101], "educated on the model's aggregate performance measures" [103]
Involvement	*End-users:* Nurses [25, 31, 46, 62, 101, 103, 106], Physicians [46, 62, 83, 101, 103]

(Continued)

Table 7. Continued

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" [9], "quality improvement team met regularly" [83], "formal and informal lines of communication" [106] *Gradual deployment:* "some oncologists have begun to use the system on trial" [52], "During this cycle ... implemented into ... ED" [83], "rolled out to a small group of RRT nurses" [106] Parallel workflows: "use of the [ML] scores ... [and] continuing standard procedure" [83], "A 3-month silent period" [107]
Involvement	*End-users:* Physicians [52], Nurses [9, 83, 106] *Other stakeholders:* IT specialists [106]
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" [31], "user acceptance ... using questionnaires seven months after implementation" [62], "end-user satisfaction ... questionnaire" [45], "web-based questionnaires to assess clinician perceptions" [46], "a survey of the users" [52], "[a survey] after completion of the trial" [59], " user acceptance ... using questionnaires seven months after implementation" [62], "the post intervention survey" [101] *In-situ observation:* "Observations focused on ... the use of systems" [116] *Interview:* "The interviews focused on ... user experience and perceptions of [the ML-based system]" [116]
Involvement	*End-users:* Physicians [45, 46, 52, 59, 62, 101, 116], Nurses [31, 46, 62, 101] *Other stakeholders:* Other [101]

Each activity includes its goal, employed techniques, and involved stakeholders.

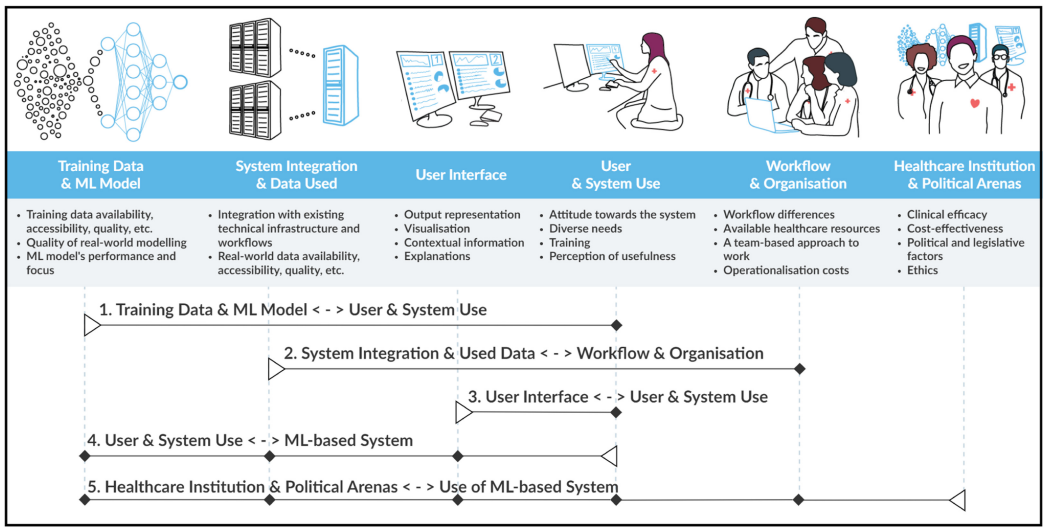

Fig. 6. Five sociotechnical interdependencies of clinician-facing ML-based systems' innovation.

Chinese hospitals were much worse than those of the MIMIC dataset [an open data set]" [64]. 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. [102]. Last, **challenges arose when some of the data were not considered during modelling** e.g., patient individual information [116], or social stressors such as unemployment or loss of a family member [8, 101].

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.)."* [96] or returning not useful information like *"a high-risk prediction for a sedated and paralysed patient"* [7]. In such cases, the ML output not only led to no change in clinical action [46] but also connoted that ML does not "understand" the clinician's job [101] and undermined its perception of usefulness [7]. 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 [101], and low accuracy in sepsis detection led to significant alarm fatigue [106, 107].

These findings indicate that technical factors such as 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.

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** [31, 45, 52, 96, 101, 101, 116, 125]. Second, **integration issues affected quality of the available data**, e.g., a system investigated by Wang et al. [116] lacked a connection to the local pharmaceutical system, which resulted in the use of outdated prescriptions. Similarly, Matthiesen et al. [82] 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 [7, 9, 101, 106, 107], e.g., algorithm using real-world data significantly increased the number of false positives [106, 107] or real-world images could not be graded in 21% of all cases [9]. Last, **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" [10, 59], or pertaining to an already diagnosed patient [46], which was perceived as irrelevant "in-the-moment" of the current workflow [24].

The above-mentioned **integration issues had several repercussions on the flow of work at the clinical sites.** In the study by Beede et al. [9], 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 [103] 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 [101, 116], overthinking [24] or outright pursuing wrong diagnoses suggested by the system [96], waiting for ML output [45], or filtering through a sea of false-positive alerts [106, 107] 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** [10] and painted a picture where not all of the required bedside assessments would be possible [83]. The additional costs of complying with recommended clinical actions were of concern to nurses in the study by Cho et al. [31], 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 such as clinical staff, build environments, and monetary resources.

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 [86], triggering ML-based alerts of severe sepsis and not providing reasons for the result [46], or presenting predictions of diagnosis without revealing the impact of historic data [64] 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 [7, 24, 46, 64], however, some clinicians cautioned against *"going too deep down a tangential rabbit hole"* [24]. 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 [7] or when the ML-output alone was not enough to support clinical action [103]. 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 [8]; risk predictions in exact percentages instead of broader risk categories [82]; or too many confounding features [24]. Providing clear, understandable, clinically relevant, and actionable output was regarded as essential for utilising the system's output [10, 24, 62, 82, 86].

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"* [64]. 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"* [64]. Similarly, nurses that received a sepsis alert were uncomfortable making decisions based on the minimal patient information available [103], and electrophysiologists' interpretation of risk predictions were, in some cases, dependent on contacting the patient for more information [82]. In paediatrics, clinicians lacked patient information such as current disease state or baseline risk to assess the clinical relevance of the predictions [7]. 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 [46]. **Clinical accountability was also found to be affected by insufficient explainability**, e.g., Sandhu et al. [103] 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. [31] 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 [24, 64] 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" [8], muddy the results [24], or make it "hard to translate" into clinically relevance [82].

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.

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"[103], and "sceptical attitudes" [96]. In some cases, physicians felt their knowledge of patients' histories and circumstances was more accurate than risk score estimations generated by the system [8, 46]. 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 [82, 125]. 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 [82].

Second, **diverse clinical roles and diverse needs** affected the perceived usefulness of the overall ML-based systems. Several studies highlighted that systems that engaged multiple clinical specialities had to conform to mixed preferences. For example, using overall mortality risk predictions [7] in paediatrics, nurses wanted minimal, actionable information and generally preferred simple and static explanations. However, 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 one-fifth of "providers" [46]. 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 [31].

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 [63]. "Misperception" [46], "misunderstanding" [103], and "confusion" [64] 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"* [64]. Lack of knowledge about the capabilities and limitations of the ML tool could also lead to degradation of trust among clinical end-users [25, 101]. When providers had little knowledge about ML and predictive modelling, they found themselves unable to assess and verify the ML model information and credibility [7]. While the absence of training was reported as problematic, some physicians considered trust in the ML-based system as something that emerges 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 [...]"* [82].

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.

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 [86]. Similarly, issues emerged when, e.g., there was weak consensus on definitions of

clinical diagnosis [45, 83]; ways of applying clinical guidelines differed across clinical sites [24], or when the unpredictability of the disease in current medical practice was high [86]. 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. [125] reported that due to low clinical outcomes, merely 7% of the physicians used the neural network score in clinical decision-making. Hollander et al. [59] 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 [101]. Clinicians considered the ML-based prediction tool as "nice to have" rather than a "need to have" [82, 116], or only half of the physicians deemed the new technology helpful [19]. Moreover, if the use of the ML-based system required actions but insufficient resources were available for carrying out the intervention, then it led to frustration and growing distrust [10].

Third, **political and legislative factors** were decisive for the success of the complete ML-based system. Benda et al. [10] 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. [9], nurses were burdened with weighing the tradeoffs 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. 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.

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 article. 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.

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 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 [73, 124]. 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 3.3). While a clear definition of the *intended use* is required to obtain regulatory approval [87], research points out that conceptualising functionalities and future use of ML-based systems is a nontrivial task [38, 124]. 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 [79]. 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 10 activities and related techniques that were employed in confined studies and during medical ML innovation processes (see Section 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 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.

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. [15] 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 [50] argued that the HCI and ML paradigms differ so much that they are, historically, contradicting [50]. Similar perspectives are raised in the respective communities [47, 67, 122] 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, 4, 29, 73].

Health research and evidence-based medicine are committed to sequential development processes and randomised control trials [15]. 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 [26]. 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 [15]. Although new, more flexible approaches have been proposed [33] and new guidelines that suggest iterative development are in place [34], the sequential nature is deeply rooted in healthcare development processes [15].

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"* [119]. 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. [3], 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 [55, 65, 99]. 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 [3]. 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 [47].

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 [51]. Several methodologies aiming to incorporate HCI perspectives into and support the development process have been proposed throughout the years [104]. HCI also offers a closer look into collaboration during such processes. Piorkowski et al. [97] 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.

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 3.2). Knowledge-driven algorithms required in-depth collaboration on feature engineering [82]. By contrast, data-driven algorithms that derived features from annotated data [76] 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. [9] used open datasets to *discretely* create their ML model, and the collaboration with domain experts was

limited to providing annotations [54]. Moreover, ML-based medical systems delivered by third-party providers limited the opportunities for mutual learning, especially during the initial activities, e.g., [83, 101].

Innovation studies applied additional informal techniques and engaged a diverse set of stakeholders. Innovation studies applied well-cited HCI techniques such as user-centred observations [101] and interviews [107], however, the same studies also applied unspecified, informal techniques using formulations such as *"working with frontline clinicians"* [107], *"assessments [...] to determine"* [101], and *"cycles to evaluate processes and incorporate clinical feedback"* [83]. On the contrary, confined studies tended to apply only well-established techniques such as interviews [8, 10, 64] and prototyping [24, 64, 86]. 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.

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 [69, 91]. 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, such as 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 [70, 109].

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 [8, 10, 19, 82, 96]. 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. [127] concentrated on increasing designers' and end-users' understanding of the tradeoffs between design objectives and helping them navigate the model selection. The developed method employed a visualisation tool that translated the tradeoffs into the end-user's practice. Another group has focused on alleviating the challenge of using AI as a design material [38, 67]. Subramonyam et al. [113] 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., References [73, 81]). 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. [17], 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. [14], 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" [88], 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 [18]. 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.

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 [6, 9, 32, 41, 61, 82, 105, 115]. 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 [11, 13], 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"* [11, 89]. They proposed to undertake a sociotechnical approach [89] 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 such as 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 such as 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. Last, 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.

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., [3, 30, 71, 78, 117]. Lim et al. [78], for example, recommended generating reasoning explanations to novice users for improvement of understanding and trust in the system. Amershi et al. [5] 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 [30, 39, 75]. 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. [124] 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. [85] 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" [11, 12] between the ML-based system and the corresponding workflows, organisation, healthcare institution, and political arenas. There are several opportunities in attending to the interrelations [11, 12] or the "sociotechnical configurations" [16, 114], which 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" [40] and "configuring" working relations [16] 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" [40] 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 [103, 106, 107]. This extended view of the innovation process, as one of configuration [112], 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., [82]) and actual deployments (see, e.g., [9, 41]) 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 [42]. 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 that can help navigate the uncertainties of AI/ML design and real-world deployment [53, 61, 125]. 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 [24, 82, 102]. Researchers also highlight the need for minimising the extra effort incurred by the operationalisation of the ML output. Yang et al. [125] 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. [61] 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 [41] 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 [114]. Other research has similarly proposed processes of "co-realisation" [57], "bricolage" [22], and "bootstrapping" [56]. 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., [84]) and regulatory oversight to manage patient risk and to ensure final approval of what essentially will be classified as a medical device [87]. This can be achieved by committing to robust clinical evaluations that aim at the high quality of care and attractive patient outcomes [49, 68].

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 article, 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 and two were covering computer science. We acknowledge that this decision may have resulted in some publications missing from this review.

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 outperforms statistical methods and is capable of supporting the work of healthcare professionals and improving 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 requires 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. 1–18. DOI: https://doi.org/10.1145/3173574.3174156

[2] George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-Kains, and Anna Goldenberg. 2020. 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. PMLR, 710–731. https://proceedings.mlr.press/v126/adam20a.html

[3] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *Proceedings of the IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP'19)*. 291–300. DOI: https://doi.org/10.1109/ICSE-SEIP.2019.00042

[4] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Mag.* 35, 4 (12 2014), 105–120. DOI: https://doi.org/10.1609/AIMAG.V35I4.2513

[5] 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. 2019. Guidelines for human-AI interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. DOI: https://doi.org/10.1145/3290605.3300233

[6] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in healthcare: Challenges appearing in the wild. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–5. DOI: https://doi.org/10.1145/3411763.3441347

[7] Amie J. Barda, Christopher M. Horvat, and Harry Hochheiser. 2020. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med. Inform. Decis. Mak.* 20, 1 (Oct 2020). DOI: https://doi.org/10.1186/s12911-020-01276-x

[8] Sally L. Baxter, Jeremy S. Bass, and Amy M. Sitapati. 2020. Barriers to implementing an artificial intelligence model for unplanned readmissions. *ACI Open* 4, 2 (7 2020), e108–e113. DOI: https://doi.org/10.1055/s-0040-1716748

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'20)*. ACM, New York, NY, 1–12. DOI: https://doi.org/10.1145/3313831.3376718

[10] Natalie C. Benda, Lala Tanmoy Das, Erika L. Abramson, Katherine Blackburn, Amy Thoman, Rainu Kaushal, Yongkang Zhang, and Jessica S. Ancker. 2020. "How did you get to this number?" Stakeholder needs for implementing predictive analytics: A pre-implementation qualitative study. *J. Amer. Med. Inform. Assoc.* 27, 5 (5 2020), 709–716. DOI: https://doi.org/10.1093/jamia/ocaa021

[11] Marc Berg. 1999. Patient care information systems and health care work: A sociotechnical approach. *Int. J. Med. Inform.* 55, 2 (1999), 87–101.

[12] Marc Berg, Jos Aarts, and Johan van der Lei. 2003. ICT in health care: Sociotechnical approaches. *Meth. Inform. Med.* 42, 4 (2003), 297–301.

[13] Marc Berg, Chris Langenberg, Ignas V. D. Berg, and Jan Kwakkernaat. 1998. Considerations for sociotechnical design: Experiences with an electronic patient record in a clinical context. *Int. J. Med. Inform.* 52, 1–3 (1998), 243–51.

[14] Pernille Bjørn and Nina Boulus-Rødje. 2015. The multiple intersecting sites of design in CSCW research. *Comput. Support. Cooper. Work* 24, 4 (8 2015), 319–351. DOI: https://doi.org/10.1007/S10606-015-9227-4

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

[16] Jeanette L. Blomberg, Lucy A. Suchman, and Randall H. Trigg. 1996. Reflections on a work-oriented design project. *Hum. Comput. Interact.* 11 (1996), 237–265.

[17] Susanne Bødker, Christian Dindler, and Ole Sejer Iversen. 2017. Tying knots: Participatory infrastructuring at work. *Comput. Support. Cooper. Work* 26, 1–2 (4 2017), 245–273. DOI: https://doi.org/10.1007/s10606-017-9268-y

[18] Susanne Bødker, Pelle Ehn, Joergen Knudsen, Morten Kyng, and Kim Madsen. 1988. Computer support for cooperative design. In *Proceedings of the ACM Conference on Computer-supported Cooperative Work (CSCW'88)*. ACM Press, New York, New York, 377–394. DOI: https://doi.org/10.1145/62266.62296

[19] Meghan Brennan, Sahil Puri, Tezcan Ozrazgat-Baslanti, Zheng Feng, Matthew Ruppert, Haleh Hashemighouchani, Petar Momcilovic, Xiaolin Li, Daisy Zhe Wang, and Azra Bihorac. 2019. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. *Surgery (United States)* 165, 5 (5 2019), 1035–1045. DOI: https://doi.org/10.1016/j.surg.2019.01.002

[20] Nicky Britten, Rona Campbell, Catherine Pope, Jenny Donovan, Myfanwy Morgan, and Roisin Pill. 2002. Using meta ethnography to synthesise qualitative research: A worked example. *J. Health Serv. Res. Polic.* 7, 4 (2002), 209–215. DOI: https://doi.org/10.1258/135581902320432732

[21] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. 2018. Points of significance: Statistics versus machine learning. *Nature Methods* 15, 4 (apr 2018), 233–234. DOI: https://doi.org/10.1038/nmeth.4642

[22] Monika Büscher, Satinder Gill, Preben Mogensen, and Dan Shapiro. 2001–03. Landscapes of practice: Bricolage as a method for situated design. *Comput. Support. Cooper. Work* 10, 1 (2001–03), 1–28. DOI: https://doi.org/10.1023/a:1011293210539

[23] Federico Cabitza, Andrea Campagner, and Clara Balsano. 2020. Bridging the "last mile" gap between AI implementation and operation: "Data awareness" that matters. *Ann. Trans. Med.* 8, 7 (2020), 501. DOI: https://doi.org/10.21037/atm.2020.03.63

[24] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'19)*. ACM, New York, 1–14. DOI: https://doi.org/10.1145/3290605.3300234

[25] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–24. DOI: https://doi.org/10.1145/3359206

[26] Michelle Campbell, Ray Fitzpatrick, Andrew Haines, Ann Louise Kinmonth, Peter Sandercock, David Spiegelhalter, and Peter Tyrer. 2000. Framework for design and evaluation of complex interventions to improve health. *BMJ* 321, 7262 (9 2000), 694–696. DOI: https://doi.org/10.1136/bmj.321.7262.694

[27] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature* 538, 7623 (10 2016), 20–23. DOI: https://doi.org/10.1038/538020a

[28] K. Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

[29] Quan Ze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi. 2022. HINT: Integration testing for AI-based features with humans in the loop. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 549–565. DOI: https://doi.org/10.1145/3490099.3511141

[30] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'19)*. 1–12. DOI: https://doi.org/10.1145/3290605.3300789

[31] Insook Cho and Insun Jin. 2019. 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, 1650–1651. DOI: https://doi.org/10.3233/SHTI190579

[32] Enrico Coiera. 2019. The last mile: Where artificial intelligence meets reality. *Journal of Medical Internet Research* 21, 11 (11 2019), e16323 DOI: https://doi.org/10.2196/16323

[33] Linda M. Collins, Susan A. Murphy, Vijay N. Nair, and Victor J. Strecher. 2005. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine* 30, 1 (8 2005), 65–73. https://doi.org/10.1207/s15324796abm3001_8

[34] Peter Craig, Paul Dieppe, Sally Macintyre, Susan Mitchie, Irwin Nazareth, and Mark Petticrew. 2008. Developing and evaluating complex interventions: The new medical research council guidance. *BMJ* 337 (2008), 1655. DOI: https://doi.org/10.1136/bmj.a1655

[35] Tom Diethe, Miquel Perello Nieto, Emma Tonkin, Mike Holmes, Kacper Sokol, Niall Twomey, Meelis Kull, Hao Song, and Peter Flach. 2018. 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, 243–252. DOI: https://doi.org/10.1145/3219819.3219883

[36] Steven E. Dilsizian and Eliot L. Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr. Cardiol. Rep.* 16, 1 (1 2014), 1–8. DOI: https://doi.org/10.1007/s11886-013-0441-8

[37] 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. 2006. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology* 6, 1 (12 2006), 35. DOI: https://doi.org/10.1186/1471-2288-6-35

[38] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. (2017). DOI: https://doi.org/10.1145/3025453

[39] John J. Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Trans. Inter. Intell. Sys. (TiiS)* 8, 2 (2018), 8. DOI: https://doi.org/10.1145/3185517

[40] Paul N. Edwards, Steven J. Jackson, Geoffrey C. Bowker, and Cory P. Knobel. 2007. Understanding infrastructure: Dynamics, tensions, and design. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/Understand?sequence=3.

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

[42] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)* 22, 4–6 (8 2013), 609–665. DOI: https://doi.org/10.1007/s10606-012-9168-0

[43] Alexander L. Fogel and Joseph C. Kvedar. 2018. Artificial intelligence powers digital medicine. *NPJ Digit. Med.* 1, 1 (2018), 5. DOI: https://doi.org/10.1038/s41746-017-0012-2

[44] Vijay N. Garla and Cynthia Brandt. 2012. Ontology-guided feature engineering for clinical text classification. *J. Biomed. Inform.* 45, 5 (10 2012), 992–998. DOI: https://doi.org/10.1016/j.jbi.2012.04.010

[45] Aimilia Gastounioti, Vasileios Kolias, Spyretta Golemati, Nikolaos N. Tsiaparas, Aikaterini Matsakou, John S. Stoitsis, Nikolaos P. E. Kadoglou, Christos Gkekas, John D. Kakisis, Christos D. Liapis, Petros Karakitsos, Ioannis Sarafis, Pantelis Angelidis, and Konstantina S. Nikita. 2014. CAROTID - A web-based platform for optimal personalized management of atherosclerotic patients. *Comput. Meth. Prog. Biomed.* 114, 2 (4 2014), 183–193. DOI: https://doi.org/10.1016/j.cmpb.2014.02.006

[46] Jennifer C. Ginestra, Heather M. Giannini, William D. Schweickert, Laurie Meadows, Michael J. Lynch, Kimberly Pavan, Corey J. Chivers, Michael Draugelis, Patrick J. Donnelly, Barry D. Fuchs, and Craig A. Umscheid. 2019. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit. Care Med.* 47, 11 (11 2019), 1477–1484. DOI: https://doi.org/10.1097/CCM.0000000000003803

[47] Fabien Girardin and Neal Lathia. 2017. When user experience designers partner with data scientists. In *AAAI Spring Symposium*, Vol. SS-17-01 -. 376–381. https://www.girardin.org/fabien/publications/girardin-lathia-aaai-symposiumspring-2017-nal.pdf.

[48] Mark L. Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Arch. Internal Med.* 165, 13 (7 2005), 1493–1499. DOI: https://doi.org/10.1001/archinte.165.13.1493

[49] The DECIDE-AI Steering Group. 2021. DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine* 27, 2 (2021), 186–187. https://doi.org/10.1038/s41591-021-01229-5

[50] Jonathan Grudin. 2009. AI and HCI: Two fields divided by a common focus. *AI Mag.* 30, 4 (9 2009), 48–57. DOI: https://doi.org/10.1609/aimag.v30i4.2271

[51] Jonathan Grudin and Steven Poltrock. 1995. 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, 93–112. DOI: https://doi.org/10.1007/bfb0035809

[52] Dongxiao Gu, Changyong Liang, and Huimin Zhao. 2017. A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. *Artif. Intell. Med.* 77, C (3 2017), 31–47. DOI: https://doi.org/10.1016/j.artmed.2017.02.003

[53] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang "Anthony" Chen. 2021. Lessons learned from designing an AI-enabled diagnosis tool for pathologists. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (4 2021), 1–25. DOI: https://doi.org/10.1145/3449084

[54] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Amer. Med. Assoc.* 316, 22 (12 2016), 2402–2410. DOI: https://doi.org/10.1001/jama.2016.17216

[55] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer. 2011. 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*.

[56] Ole Hanseth and Margunn Aanestad. 2003. Design as bootstrapping. On the evolution of ICT networks in health care. *Meth. Inform. Med.* 42, 4 (2003), 385–91.

[57] Mark Hartswood, Rob Procter, Roger Slack, Alex Voß, Monika Büscher, Mark Rounceeld, and Philippe Rouchy. 2002. Co-Realisation: Towards a Principled Synthesis of Ethnomethodology and Participatory Design. *Scand. J. Inf. Syst.* 14, 2 (sep 2002), 9–30.

[58] Hamed Hassanzadeh, Anthony Nguyen, Sarvnaz Karimi, and Kevin Chu. 2018. Transferability of artificial neural networks for clinical document classification across hospitals: A case study on abnormality detection from radiology reports. *J. Biomed. Inform.* 85 (9 2018), 68–79. DOI: https://doi.org/10.1016/j.jbi.2018.07.017

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

[60] Andreas Holzinger. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[61] Maia Jacobs, Jeffrey He, and Melanie F. Pradier. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In *Proceedings of the Conference on Human Factors in Computing Systems*. DOI: https://doi.org/10.1145/3411764.3445385

[62] Stefanie Jauk, Diether Kramer, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. 2021. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: A mixed-methods study. *J. Med. Syst.* 45, 4 (4 2021). DOI: https://doi.org/10.1007/s10916-021-01727-6

[63] Stefanie Jauk, Diether Kramer, Franz Quehenberger, Sai Pavan Kumar Veeranki, Dieter Hayn, Günter Schreier, and Werner Leodolter. 2019. Information adapted machine learning models for prediction in clinical workflow. In *Studies in Health Technology and Informatics*, Vol. 260. IOS Press, 65–72. DOI: https://doi.org/10.3233/978-1-61499-971-3-65

[64] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An intelligent clinical decision assistance system. *ACM Trans. Comput. Healthc.* 1, 1 (3 2020). DOI: https://doi.org/10.1145/3344258

[65] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the Conference on Human Factors in Computing Systems*. 3363–3372. DOI: https://doi.org/10.1145/1978942.1979444

[66] John Kang, Olivier Morin, and Julian C. Hong. 2020. Closing the gap between machine learning and clinical cancer care—first steps into a larger world. *JAMA Oncol.* 6, 11 (2020), 1731–1732. DOI: https://doi.org/10.1001/jamaoncol.2020.4314

[67] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the intersections: User experience + research scientist collaboration in a generative machine learning interface. In *Conference on Human Factors in Computing Systems*. DOI: https://doi.org/10.1145/3290607.3299059

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

[69] Finn Kensing and Joan Greenbaum. 2012. Heritage: Having a say. In *Routledge International Handbook of Participatory Design*. Routledge, 41–56. DOI: https://doi.org/10.4324/9780203108543-9

[70] Finn Kensing, Jesper Simonsen, and Keld Bødker. 1998. MUST: A method for participatory design. *Hum.-Comput. Interact.* 13, 2 (1998), 167–198. DOI: https://doi.org/10.1207/s15327051hci1302_3

[71] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI'19*. 1–14. DOI: https://doi.org/10.1145/3290605.3300641

[72] Peter Kontschieder, Jonas F. Dorn, Cecily Morrison, Robert Corish, Darko Zikic, Abigail Sellen, Marcus D'Souza, Christian P. Kamm, Jessica Burggraaff, Prejaas Tewarie, Thomas Vogel, Michela Azzarito, Ben Glocker, Peter Chin, Frank Dahlke, Chris Polman, Ludwig Kappos, Bernard Uitdehaag, and Antonio Criminisi. 2014. Quantifying progression of multiple sclerosis via classification of depth videos. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 17 (2014), 429–437.

[73] Ilpo Koskinen, Youn-kyung Lim, Teresa Cerratto-Pargman, Kenny Chow, William Odom, Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the Designing Interactive Systems Conference*, 585–596. DOI: https://doi.org/10.1145/3196709.3196730

[74] John R. Koza, Forrest H. Bennett, David Andre, and Martin A. Keane. 1996. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Artificial Intelligence in Design'96*. Springer, Dordrecht, 151–170. DOI: https://doi.org/10.1007/978-94-009-0279-4_9

[75] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. Association for Computing Machinery, New York, NY, 5686–5697. https://doi.org/10.1145/2858036.2858529

[76] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (5 2015), 436–444. DOI: https://doi.org/10.1038/nature14539

[77] Cindy S. Lee, Paul G. Nagy, Sallie J. Weaver, and David E. Newman-Toker. 2013. Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology* 201, 3 (9 2013), 611–617. DOI: https://doi.org/10.2214/AJR.12.10375

[78] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. 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 (CHI'09)*. Association for Computing Machinery, New York, NY, 2119–2128. https://doi.org/10.1145/1518701.1519023

[79] Christopher J. Lindsell, William W. Stead, and Kevin B. Johnson. 2020. Action-informed artificial intelligence—matching the algorithm to the problem. *JAMA* 323, 21 (06 2020), 2141–2142. DOI: https://doi.org/10.1001/jama.2020.5035 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2765667/jama_lindsell_2020_vp_200067.pdf.

[80] Jocelyn Maclure. 2019. The new AI spring: A deflationary view. *AI & SOCIETY* 35 (2019), 747–750.

[81] Regan Mandryk, Mark Hancock, Mark Perry, Anna Cox, Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping machine learning advances from HCI research to reveal starting places for design innovation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–11. DOI: https://doi.org/10.1145/3173574.3173704

[82] Stina Matthiesen, Søren Zöga Diederichsen, Mikkel Klitzing Hartmann Hansen, Christina Villumsen, Mats Christian Højbjerg Lassen, Peter Karl Jacobsen, Niels Risum, Bo Gregers Winkel, Berit T. Philbert, Jesper Hastrup Svendsen, and Tariq Osman Andersen. 2021. Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: Near-live feasibility and qualitative study. *JMIR Hum. Fact.* 8, 4 (11 2021), e26964. DOI: https://doi.org/10.2196/26964

[83] Andrea McCoy and Ritankar Das. 2017. 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. *BMJ Open Qual.* 6, 2 (10 2017), e000158. DOI: https://doi.org/10.1136/bmjoq-2017-000158

[84] Stuart McLennan, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin, and Alena Buyx. 2020. An embedded ethics approach for AI development. *Nat. Mach. Intell.* 2, 9 (2020), 488–490.

[85] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4 (12 2021), 1–45. DOI: https://doi.org/10.1145/3387166

[86] Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and Sân Lindley. 2018. Visualizing ubiquitously sensed measures of motor ability in multiple sclerosis: Reflections on communicating machine learning in practice. *ACM Trans. Interact. Intell. Syst.* 8, 2 (7 2018). DOI: https://doi.org/10.1145/3181670

[87] Urs J. Muehlematter, Paola Daniore, and Kerstin N. Vokinger. 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. *The Lancet Digital Health* 3, 3 (2021), e195–e203.

[88] Michael J. Muller and Allison Druin. 2012. Participatory design. In *The Human–Computer Interaction Handbook*. CRC Press, 1125–1153. DOI: https://doi.org/10.1201/b11963-ch-49

[89] Enid Mumford and Maria Weir. 1979. *Computer systems in work design.the ETHICS method: eective technical and human implementation of computer systems*. New York: Wiley.

[90] Daniel B. Neill. 2013. Using artificial intelligence to improve hospital inpatient care. *IEEE Intell. Syst.* 28, 2 (2013), 92–95. DOI: https://doi.org/10.1109/MIS.2013.51

[91] Kristen Nygaard and Olav Terje Bergo. 1975. The trade unions—New users of research. *Personnel Review* 4, 2 (2 1975), 5–10. DOI: https://doi.org/10.1108/eb055278

[92] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan–A web and mobile app for systematic reviews. *System. Rev.* 5, 1 (12 2016), 1–10. DOI: https://doi.org/10.1186/S13643-016-0384-4

[93] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *System. Rev.* 10, 1 (12 2021), 1–11. DOI: https://doi.org/10.1186/S13643-021-01626-4/FIGURES/1

[94] Brijesh Patel and Partho Sengupta. 2020. Machine learning for predicting cardiac events: What does the future hold? *Expert Rev. Card. Therapy* 18, 2 (2020), 77–84. DOI: https://doi.org/10.1080/14779072.2020.1732208

[95] Vimla L. Patel, Edward H. Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R. Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. 2009. The coming of age of artificial intelligence in medicine. *Artif. Intell. Med.* 46, 1 (5 2009), 5–17. DOI: https://doi.org/10.1016/j.artmed.2008.07.017

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

[97] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How AI developers overcome communication challenges in a multidisciplinary team. *Proc. ACM Hum.-comput. Interact.* 5, CSCW1 (2021), 1–25. DOI: https://doi.org/10.1145/3449205

[98] Fernanda Polubriaginof, Nicholas P. Tatonetti, and David K. Vawdrey. 2015. An assessment of family history information captured in an electronic health record. *Annual Symposium proceedings. AMIA Symposium* 2015 (2015), 2035–2042. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765557//pmc/articles/PMC4765557//pmc/articles/PMC4765557/?report=abstract.

[99] Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras. 2017. *Principles of Data Wrangling*. O'Reilly Media, Inc. 94 pages.

[100] Yvonne Rogers. 2011. Interaction design gone wild: Striving for wild theory. *Interactions* 18 (2011), 58–62.

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

[102] Santiago Romero-Brufau, Kirk D. Wyatt, Patricia Boyum, Mindy Mickelson, Matthew Moore, and Cheristi Cognetta-Rieke. 2020. Implementation of artificial intelligence-based clinical decision support to reduce hospital readmissions at a regional hospital. *Appl. Clin. Inform.* 11, 4 (8 2020), 570–577. DOI: https://doi.org/10.1055/s-0040-1715827

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

[104] Ahmed Seffah, Michel C. Desmarais, and Eduard Metzker. 2005. HCI, usability and software engineering integration: Present and future. In *Human-Computer Interaction Series*. Vol. 8. Springer, 37–57. DOI: https://doi.org/10.1007/1-4020-4113-6_3

[105] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 59–68. DOI: https://doi.org/10.1145/3287560.3287598

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

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

[108] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.–Comput. Inter.* 36, 6 (2020), 1–10. DOI: https://doi.org/10.1080/10447318.2020.1741118

[109] Jesper Simonsen and Toni Robertson. 2012. *Routledge International Handbook of Participatory Design*. 1–300 pages. DOI: https://doi.org/10.4324/9780203108543

[110] Anders Søgaard. 2021. Explainable natural language processing. In *Synthesis Lectures on Human Language Technologies*. Vol. 14. Morgan & Claypool, 1–123. DOI: https://doi.org/10.2200/S01118ED1V01Y202107HLT051

[111] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'20)*. Association for Computing Machinery, Inc., 56–67. DOI: https://doi.org/10.1145/3351095.3372870

[112] James Stewart and Robin Williams. 2005. The wrong trousers? Beyond the design fallacy: Social learning and the user.

[113] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards a process model for co-creating AI experiences. In *Proceedings of the Designing Interactive Systems Conference*. ACM, New York, NY, 1529–1543. DOI: https://doi.org/10.1145/3461778.3462012

[114] Lucy Suchman, Randall Trigg, and Jeanette Blomberg. 2002. Working artefacts: Ethnomethods of the prototype. *Br. J. Sociol.* 53, 2 (06 2002), 163–179. DOI: https://doi.org/10.1080/00071310220133287

[115] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput.-Hum. Interact.* 27, 5 (8 2020), 1–53. DOI: https://doi.org/10.1145/3398069

[116] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–18. https://doi.org/10.1145/3411764.3445432

[117] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–15. DOI: https://doi.org/10.1145/3290605.3300831

[118] 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. 2019. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* 25, 9 (9 2019), 1337–1340. DOI: https://doi.org/10.1038/s41591-019-0548-6.

[119] Terry Winograd. 2006. Shifting viewpoints: Artificial intelligence and human-computer interaction. *Artificial Intelligence* 170, 18 (12 2006), 1256–1258. DOI: https://doi.org/10.1016/j.artint.2006.10.011

[120] Christine T. Wolf. 2019. Explainability scenarios: Towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Vol. Part F147615. ACM, New York, NY, 252–257. DOI: https://doi.org/10.1145/3301275.3302317

[121] Yan Xu, Kai Hong, Junichi Tsujii, and Eric I-Chao Chang. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *J. Amer. Medic. Inform. Assoc.* 19, 5 (2012), 824–832. DOI: https://doi.org/10.1136/amiajnl-2011-000776

[122] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A case study of exploring the right things to design with language intelligence. In *Proceedings of the Conference on Human Factors in Computing Systems*. Vol. 12, ACM, New York, NY. 1–12. DOI: https://doi.org/10.1145/3290605.3300415

[123] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS'18)*. 585–596. DOI: https://doi.org/10.1145/3196709.3196730

[124] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the Conference on Human Factors in Computing Systems*. DOI: https://doi.org/10.1145/3313831.3376301

[125] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'19)*. Association for Computing Machinery, New York, NY, 1–11. DOI: https://doi.org/10.1145/3290605.3300468

[126] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'16)*. Association for Computing Machinery, New York, NY, 4477–4488. DOI: https://doi.org/10.1145/2858036.2858373

[127] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop. In *Proceedings of the ACM Designing Interactive Systems Conference*. ACM, New York, NY, 1245–1257. DOI: https://doi.org/10.1145/3357236.3395528

Received 13 November 2021; revised 25 November 2022; accepted 28 November 2022