Burying the Lead: Adjusting Goals to Manage Functional Limitations of AI Tools in Healthcare

Jacqueline Kernahan[1], Richard Bartels[2], Mark de Reuver[1], Daniel Oberski[3], Roel Dobbe[1]

[1]Delft University of Technology, Delft, The Netherlands
[2]University Medical Centre Utrecht, Utrecht, The Netherlands
[3]Utrecht University, Utrecht, The Netherlands
j.a.kernahan@tudelft.nl, r.t.bartels-6@umcutrecht.nl, g.a.dereuver@tudelft.nl, d.l.oberski@uu.nl, r.i.j.dobbe@tudelft.nl

Abstract

Artificial intelligence based tools are being developed for decision support in healthcare, however, they are frequently found to lack the required functionality to achieve the clinical goals for which they were built. This results in wasted time, money and resources for hospitals attempting to implement and operate such tools. To determine how functionality issues can be resolved prior to tool implementation, it is necessary to understand why such tools are being designed and then built. Our research focuses on clinical decision support tools with functionality issues arising from target variable invalidity. In this paper, we analyze published articles which present clinical decision support tool designs related to clinical goals. These tools use machine learning models trained on electronic health record data. We find that design decisions driven by data availability can introduce construct invalidity in clinical decision support tool designs, leading to an inability of the tool to address the clinical goal. We observe that alternative goals to the main clinical goal are used to justify continued development. We show that functional limitations of the tool related to the clinical goal can be obscured by imprecise terminology in the model's stated functionality. Finally, we highlight the need for reconsidered approaches to dataset creation, defining success criteria, and the reporting and transparency of research outcomes as they relate to clinical goals.

Introduction

In recent years, the interest in, and experimentation with, artificial intelligence (AI) applications has expanded and intensified across a variety of industry sectors. Increasingly, there are expectations that AI tools will help to improve decision making and provide substantial efficiency gains, particularly in sectors which face issues of labor shortages and productivity deficits. But despite these aspirations, there is growing evidence that the actual benefits of AI are falling short of these expectations, while its future impacts are highly uncertain (Thais 2024), both in terms of offering promised functionality (Raji et al. 2022) as well productivity gains (Acemoglu 2025). This uncertainty in functionality is mostly evidenced by a lack of consensus across AI experts of what "capabilities" we now have and may expect from AI models, as laid out in the first International AI Safety report (Bengio et al. 2025). Additionally, safety and other core values cannot be ascribed to the AI model as a "capability", and require intimate and integrative engagement with the social and institutional contexts in which these are necessary (Dobbe 2025). Therefore, although some clear "general" capabilities of certain AI technologies may stabilize over time, an ongoing challenge will be understanding how these can be used for domain-specific decision support to deliver on context-specific requirements while preventing undesirable outcomes.

The potential for negative impacts to stakeholders caused by AI-based decision automation has resulted in sustained discourse in both academia and industry on whether AI tools *ought to* be used for automating certain decision tasks. In the meantime, we are observing that existing AI decision support tools do not function such that they can support the aims of the decision task. Thus the more pertinent question arises: are AI tools, as they are currently being developed and built even *able to* be used for certain decision tasks? Failure to challenge the claimed functional capabilities of AI tools in decision support has led to a 'fallacy of AI functionality' (Raji et al. 2022). Recent work has shown that misplaced assumptions of AI model functionality have led to the deployment of tools that do not work as expected, resulting in significant harms to stakeholders including the denial of appropriate medical care (Buonora et al. 2024), withholding of essential financial support (Egan 2019; Dancu 2021), and disruption to education (Klee 2023).

In an effort to work towards a more rigorous understanding of what AI functionalities are possible and how we understand and attain these, this research seeks to examine why tool capabilities are falling short of what is required, leading either to implementation failure (Staub et al. 2023) or harms to stakeholders from deployed products (Raji et al. 2022). Previous research in this area has considered the shortfall between machine learning (ML) model performance claims and reality (Kapoor et al. 2024; Saxena et al. 2025), and how this impacts claims relating to social outcomes achieved (Birhane et al. 2022; Kou 2024). However, many instances of functionality failure have been shown to occur, not because of performance shortfall, but because of construct validity issues inherent in the design of the ML model (Buonora et al. 2024; Obermeyer et al. 2019; Raji et al. 2022; Larson et al. 2016). These issues make it impos-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sible for it to possess the required functionality to achieve the defined social goal (Raji et al. 2022; Buonora et al. 2024; Obermeyer et al. 2019; Larson et al. 2016). It remains unclear why ML tools with these issues are being designed and built. Thus, in this paper, we focus on gaining more insight into the conditions that lead to non-functional tools being designed, built, and implemented.

Healthcare is one of the sectors which has seen significant and growing expectations around the potential benefits of machine learning tools. Globally, the healthcare sector is experiencing challenges related to a lack of health personnel, organizational and procedural inefficiencies, inequity in access to quality healthcare, and growth in demand due to aging populations (Lekadir et al. 2022). Artificial intelligence tools have been proposed as a solution to help relieve pressure on healthcare providers, by reducing the burden of administrative tasks and providing clinical decision support (WHO Guidance 2021). Our study specifically focuses on the functionality of ML decision support systems in clinical healthcare.

Machine learning based clinical decision support (CDS) tools can synthesize large amounts of patient data to make improved predictions related to patient diagnosis, prognosis, treatment selection and triage. Despite a demand for ML-based CDS tools across multiple clinical applications, in many cases, these tools are not working as required. In some instances, tools were deployed before their functional limitations were discovered (Buonora et al. 2024; Wong et al. 2021; Obermeyer et al. 2019), creating significant safety risks to users and patients.

However, in many cases, functional issues with tools are identified by users during the implementation process, leading to implementation failure. That is, implementation of the tool is stopped and the desired outcomes and benefits are not realized. Identification at this stage prevents harm to patients but still results in wasted time, money, and resources for hospitals. Studies analyzing implementation failures of clinical decision support tools have found that they were often rejected by users as their functional capabilities did not meet the needs of the users within the clinical operational context (Abell et al. 2023; Westerbeek et al. 2021).

To reduce losses due to implementation failure, it is necessary to identify functional issues prior to implementation. Therefore, we seek to understand why there are clinical decision support tools being designed and then built, whose functionality does not meet the needs of users. This despite strong demand, interest, and use cases for such tools. As improved predictive capabilities are a core justification for the use of these tools, we in particular examine tools where issues in their design mean they do not predict the outcome they need to predict.

We conducted a literature review to identify recently published articles where ML-based clinical decision support tools were designed and built. From these, we identified examples of designs in the literature where functionality issues arise due to construct validity issues in the predictive variable (also known as the target variable). We performed a detailed, qualitative examination of how these issues arise, how they are addressed, and how they are communicated.

Our core findings are as follows:

(1) Design decisions regarding the choice of the invalid model target variable were driven by the datasets accessible to the developers

(2) The pursuit of alternative goals was used to justify continued development of models that could not support the main clinical end goal

(3) Resultant model limitations were not clearly communicated creating a risk of them remaining unaddressed in future iterations.

In the following section, we examine the existing literature to lay out core definitions on AI functionality and validity with which we ground our study. The methods section outlines the methods and data used for our analysis. In the results section we present the selected case studies and our key findings from the qualitative analysis. Our discussion contextualizes our findings and lists our main reflections and discussion points, and the final section provides our conclusions.

Background

In this section we provide background and definitions for the concepts which will underpin our research methods and analysis: functionality, target variable construct validity and functional failures in clinical decisions support tools.

What is functionality?

The functionality of a system, product, or tool refers to the set of capabilities it has or tasks it can perform. Put simply, what it can do. A function refers to one of those capabilities. Therefore, a system, product, or tool may have multiple functions.

We can more formally define a function of a tool as "the intended and deliberately caused ability to bring about a transformation of a part of the environment of the [tool]" (Roozenburg and Eekels 1995). This introduces the concept of intention. From this, we describe a tool as being 'functional' if its capabilities enable it to bring about an intended outcome. To determine the intended functions, it is necessary to start with a problem to be solved. This problem gives rise to a goal which informs the functions to be realized by the tool (Roozenburg and Eekels 1995). If a tool cannot achieve all the intended outcomes due to limitations in its capabilities, we consider it to have insufficient functionality. For ML-based clinical decision support tools, a clinical problem informs the clinical goal. The subsequent functionality of the tool should be able to achieve intended outcomes that support the realization of the clinical goal.

Development of ML-based CDS tools whose function cannot support the desired outcome can result in significant harms to patients. These include the denial of appropriate medical care (Buonora et al. 2024), unfair allocation of treatment (Obermeyer et al. 2019), and misdiagnosis (Wong et al. 2021). However, functionality is rarely prioritized in the evaluation of ML tools; instead, emphasis is placed on reliability, fairness, and explainability (Ojewale et al. 2025). When the evaluation of functionality does occur, it is often

during implementation or operation of the tools, for example through functionality audits (Mökander et al. 2024). This does not enable the prevention of the development of tools that do not have appropriate functionality.

Target variable construct validity

To evaluate a ML model prior to implementation or deployment, it is necessary to consider its validity (Coston et al. 2023). Functionality issues can arise due to problems with external validity of performance claims (Kapoor et al. 2024; Saxena et al. 2025), or the construct validity of the benchmarks against which the model is assessed (Alaa et al. 2025). Additionally, research has highlighted the common occurrence of unsubstantiated links made between model performance and social outcome claims, often resulting from a lack of engagement with the model application context (Birhane et al. 2022; Kou 2024). A lack of contextual awareness or engagement can give rise to construct validity issues in models. These issues are introduced into the design prior to the measurement of model performance.

Construct validity considers how well a concept is operationalized for use in reality (Drost 2011). In relation to machine learning models, it can be defined as the "validity of inferences regarding the extent to which a model reflects the construct it is aimed at predicting" (Anglin 2024). The variable a model is trained to predict is termed the target variable.

Construct invalidity in target variables can arise in various ways, including by inadequate interpretation of the construct, errors in target variable labels, and biases in proxy labels (Anglin 2024; Guerdan et al. 2023a,b). An example of target variable construct invalidity causing functionality issues in a CDS tool is observed in a high profile case study from the United States, where a clinical decision support tool was developed to identify patients who require high needs care (Obermeyer et al. 2019). The tool was trained on historical records of insurance claims for treatment. This proxy target variable did not accurately reflect the construct that the developers aimed to predict. Thus, the model did not predict which patients required high needs care. It was not functional.

Understanding functional failures of CDS tools

Functionality issues caused by construct validity of target variables have the potential to be a contributing factor to implementation failure of clinical ML models. However, research is yet to be conducted on this.

The purpose of our research is not to confirm a causal relationship between construct validity of target variables and implementation failure of clinical ML tools, but to take the first step towards bridging this research gap. We seek to understand how construct invalidity of target variables occurs in the model design, the impact it has on functionality, and how such products might end up being implemented.

It is relevant to apply this domain-focused lens as clinical ML research differs from more general ML research. Clinical ML research is heavily influenced by other scientific fields such as medical informatics and much of the clinical AI literature is published in venues with a specific bioinformatics or medical focus. Additionally, the models developed in clinical ML are inherently grounded within the clinical context due to the specificity of problems and datasets; therefore, there may be greater contextual awareness than in general ML research (Birhane et al. 2022). Thus, if construct invalidity does occur, it may emerge differently than in other applications or fields.

Methods

The nature of our research question requires a deep and thorough analysis of individual papers within the clinical AI literature. Over the last few years, there has been a significant increase in the volume of published papers presenting the development of AI models for use in clinical contexts such as hospitals and medical practices. Thus, it is not possible to perform a systematic review of every paper which presents an AI model developed for use in a clinical context. We therefore focus on papers in a specific area of clinical AI research, namely papers which present AI models that utilize natural language processing on free-text data from electronic health records (EHRs) for decision support applications. These applications have clear demand from clinicians to help relieve administrative burden and support the synthesis of large volumes of patient data and records (Bongurala et al. 2024). Therefore, it is critical to understand how to ensure such tools are built with the functionality necessary to address user needs.

Data collection

Due to the rapidly evolving nature of the clinical ML field, we seek to capture a contemporary snapshot of the literature, restricting our search to papers published from January 2023 to June 2024. A systematic literature search was conducted on 7 August 2024 from databases Scopus and Pubmed. We extracted relevant articles and screened them for inclusion in the analysis. Databases Scopus and PubMed were used to conduct the literature search. The following string was used to perform the search: (A) AND (B) AND (C) AND (D) AND (E), combining the keywords as specified in Table 1.

Groups	Keywords
(A)	"clinical decision support" OR "clinical decision system" OR "clinical decision making"
(B)	"machine learning" OR "artificial intelligence" OR "data driven"
(C)	"design" OR "product" OR "solution" OR model OR "system''
(D)	"health records" OR "medical records" OR "patient records" OR "ehr" OR "emr"
(E)	"NLP" OR "natural language processing"

Table 1: Keywords used in database search with the AND operator applied between each group.

The process for identifying core articles is shown in Figure 1. Following the extraction of 72 articles from the databases, duplicate records were removed, resulting in 54 records. Screening was conducted based on two criteria.

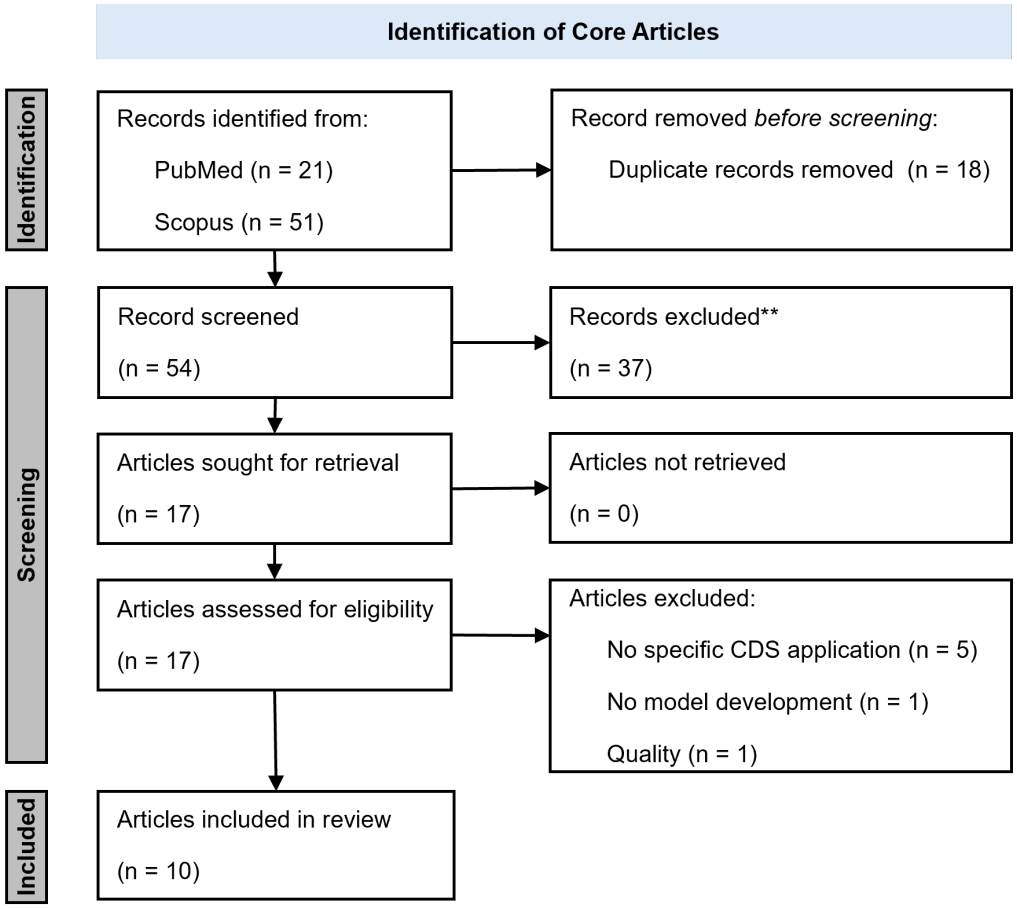

Figure 1: Flowchart of key article selection (adapted from PRISMA (Page et al. 2021))

First, the article needed to present an AI model which uses EHR text data for a clinical decision support application. Second, the article needed to describe the development phase of the ML model. After excluding ineligible records, 17 papers were sought for retrieval and assessed for eligibility by applying the same criteria. This resulted in the identification of 10 core articles for analysis.

Case-study identification

A systematic analysis of each of the 10 core articles was conducted to identify occurrences of designs where the functionality of the model did not support the clinical goal. These were used as case studies on which we performed a detailed qualitative analysis to answer the following questions: (1) How did construct validity issues arise and what was their impact on the CDS tool functionality? (2) Why did the development of the CDS tool continue despite known functionality issues? and (3) How were the known functional limitations of the CDS tool communicated?

To identify CDS tool designs with functionality issues, we considered the construct validity of the target variables in their underpinning ML models. In their work on threats to validity in supervised machine learning, Anglin (2024) presents four threats to construct validity arising in supervised machine learning. Two of these relate to the target variable: "inadequate explication of constructs" and "errors in human labels [of data]" (Anglin 2024).

Presence of construct invalidity due to the first threat can be assessed by comprehensively specifying the construct of interest, that is, what is to be predicted. The second can be assessed by specifying the level of knowledge or types of processes that create the data labels to be used. For CDS tool predictions, this would require the following types of considerations: should the labels be generated by doctors, by pathology results, or by some other means? Should labels come from current or historical clinical settings? Should labels be assigned by specialists or general practitioners? For each article, we extracted these specifications based on the main clinical goal and used them to define the "ideal target variable".

Extraction of specifications for our analysis was done in four steps:

- *Identify the main clinical goal*: Distinct clinical challenges or needs, which were used as justification for the development of the model, were identified and extracted from the introduction sections of each article.
- *Identify the intended CDS tool function*: The proposed function of the CDS tool's machine learning model, as

intended to address the clinical goal, was again extracted from the introduction section of each article.

- *Determine specification 1 (construct of interest)*: If it was stated explicitly in the article, the construct was extracted from the text. If it was not stated explicitly, the construct was inferred from the intended CDS tool function.
- *Determine specification 2 (labeling requirements)*: If it was stated explicitly, the construct was extracted from the text. If not stated explicitly, the construct was inferred from the intended CDS tool function.

We then compared the actual target variable used in the model to these specifications to determine if there was misalignment to the specifications, indicating issues of construct validity.

A limitation of this approach is that we did not have full access to the datasets used. We needed to interpret from the authors' disclosure and justification whether the selected dataset and target variable used did not meet the specifications. Where there is a clear discrepancy between the specifications and the used target variable, we could identify the article as having construct invalidity. However, we cannot claim that the designs in any of the articles had complete construct validity with regards to the clinical goal. There may be discrepancies or target variable bias (Guerdan et al. 2023b) that are not apparent in the information provided in the articles, or that the authors of this paper did not have the expertise to identify. Therefore, we classified each article as "target variable construct invalidity observed" or "no target variable construct invalidity observed."

Case-study analysis

For each of the articles classified as "target variable construct invalidity observed", we performed a qualitative analysis to answer the following questions.

(1) How did construct validity issues arise and what is their impact on the CDS tool functionality? In the first phase of analysis, we examined the methods and the discussion of research limitations provided by the authors to understand the selection process of the target variable. From this, we determined how the construct invalidity arose, and how it affected the functionality of the CDS tool.

(2) Why did development of the CDS tool continue? In the second phase of our analysis, we sought to understand the justification given for pursuing the development of a CDS tool where target variable construct invalidity meant that the functionality could not support the achievement of the clinical end goal, and where this limitation was known to the authors. For each case-study article, we performed a detailed reading of the text to identify all the goals presented in the text.

The articles were reviewed sentence by sentence and references to goals were coded. These were identified by signifying words that indicated an intention of the research, such as "goal", "aim", "objective", "sought to", "intended". To identify indirect references to goals, we also assessed and coded each sentence to determine if a problem, outcome or benefit was presented.

Sentences were then grouped with others that referred to the same concept, and goal descriptions were assigned to each concept. For example, the problem, "patients are being underdiagnosed", and the goal, "we aim to improve diagnosis rates", would be grouped together, and the goal description "improve diagnosis rates" assigned. For each goal, we indicated its dependency on the construct validity of the target variable.

(3) How were functional limitations of the CDS tools communicated? In the third phase of analysis, we examined the stated functionality in each article to determine if the functional limitations of the developed model, and its inability to address the clinical end goal, were clearly communicated. We defined the stated functionality as a short summary statement outlining what the developed AI model does. For example *"the developed model identified patients with type 2 diabetes using clinical notes in EHRs"*. These were extracted from the conclusion sections of the abstract. Where there was no stated functionality in the abstract, we extracted them from the discussion section in the main body of the paper.

The stated functionalities, by nature of being summary statements providing an overview of the tools' capabilities, are abstractions of the detailed descriptions of the tools. Ideally, such abstractions should convey the minimum amount of information necessary for a reader to understand the functionality of the tool (Leveson 2017). However, it is also possible that these statements may be misabstractions (de Troya et al. 2025) of the tool's functionality. A misabstraction is defined by de Troya et al. (2025) as:

> a representation of an entity, phenomenon, or procedure that omits critical contextual information and renders that representation problematic when it is reintegrated into the context of the sociotechnical system for which it has been made.

While the stated functionality may be a factually accurate representation of the tool's functionality, there could be critical information missing that prevents an understanding of its limitations. In this case we would consider the stated functionality to not be clearly communicated. We therefore assessed the stated functionality of each tool in two phases. First, we determined if the stated functionality was a factually accurate description of the ML model functionality. If it was *not* factually accurate we concluded that the actual functionality was not clearly communicated. If the stated functionality *was* factually accurate, we then assessed whether it was a misabstraction of the tool's capabilities. If the statement was a misabstraction, we concluded that the functionality was not clearly communicated. This decision process is shown in Figure 2.

Results

In the following four sections, we present the results from each phase of our analysis. In the first section, we review each of the 10 articles and identify three where the model designs had construct validity issues related to the main clinical goal. In the second section, we find that construct validity

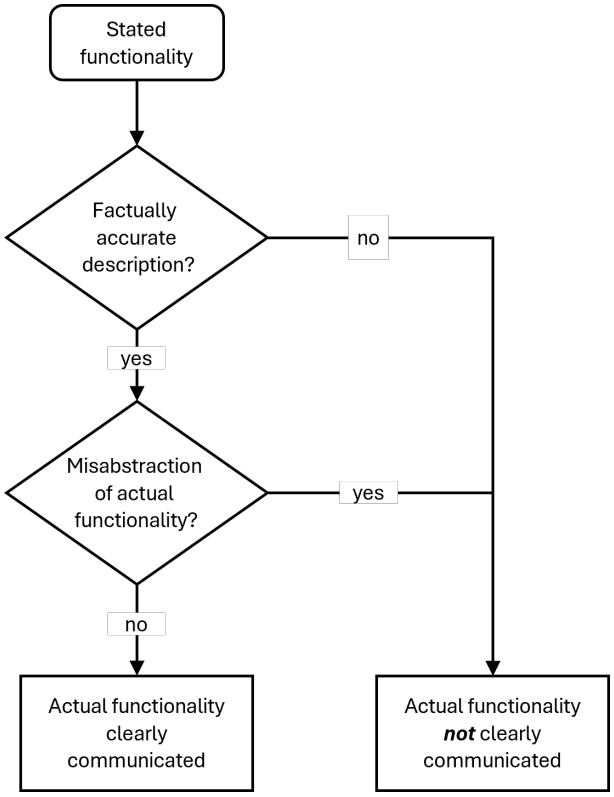

Figure 2: Decision tree to determine if the stated functionality clearly communicated the actual functionality of the CDS tool.

issues arose due to the use of pre-determined datasets which had target variables that did not align to the functional requirements of the clinical goal. We discuss the impact of the use of these datasets on the CDS tool functionality. In the third section, we examine why development of the models continued despite known functionality issues and categorize the types of goals that were used for this justification. In the fourth section, we find that the stated functionality of each of the developed CDS tools does not clearly communicate their limitations in relation to the main clinical goal.

Construct invalidity was identified in three cases

The clinical goals specified in the core papers span a wide range of medical specializations and uses. In our dataset, the application type was distributed approximately equally between prognosis (4 cases), diagnosis (3 cases) and triage (3 cases). Seven of the articles sought to use the CDS to improve the quality of predictions for their use case, and four of the articles aimed to enable earlier predictions with the CDS (one article aimed to do both). Our dataset indicates the variability and range of potential uses for machine learning and natural language processing based clinical decision support systems in medical applications.

Out of the 10 articles, three were observed to have construct validity issues relating to the target variable used in the model, indicating that there would be subsequent functionality issues in their CDS tools. In one case, a target variable matching the definition of the required construct was not available in the dataset (Fudickar et al. 2024). In the other two, the labeling of the target variable does not sufficiently represent the required construct due to limitations in the labeling process (Abdel-Hafez et al. 2023; de Hond et al. 2024). These issues were acknowledged and discussed by the authors of each article.

In the following subsections, we elaborate on three main issues observed in the articles with construct invalidity. First, adhering to a predetermined dataset that is misaligned to clinical goals; second, shifting primary goals justifying continued development of models with construct invalidity; and third, an overstated functionality of the models.

Pre-determined datasets were misaligned to the clinical goal resulting in construct invalidity

Abdel-Hafez et al. (2023) aim to reduce the time required for clinicians to triage to ear, nose, throat (ENT) specialists and ensure triage is done in line with the hospital's clinical prioritization criteria (CPC), enabling equitable access to these specialists for patients. This requires that the target variable 'triage categories' is used, and that the labeling of this target variable is in line with the CPC. However, the authors note that they trained the model on historical data where the labels were not always adhering to the CPC, creating a construct validity issue. The authors explained this was done as there were not enough records that did adhere to the CPC and the time and effort that would be required to manually re-label the records was prohibitive for this project. There-

fore, they went ahead with the historical data for training despite the construct invalidity.

A similar challenge is faced by de Hond et al. (2024). The defined clinical goal is to improve early detection of depression in patients receiving cancer treatment, as they are currently underdiagnosed. This presents a challenge in obtaining the necessary target variable as, if patients are currently underdiagnosed, the ground truth data is not available for every patient with depression in early cancer treatment. This is acknowledged by the authors who state, "depressive symptoms may have existed and not been recorded or ignored by the oncology-focused clinicians, or that the patient did not express their depressive symptoms to their oncology-focused clinician." In each of these cases, there would have been no diagnosis. Consequently, in the dataset used to train the model, these patients would have a 'no depression' label. A number of proxy target variables were tested to attempt to identify depression cases that were not officially diagnosed; however, none of these were found to be good enough for the final model. In their conclusion, the authors include a call to refine the target variable labeling in future studies.

The CDS tool built by Fudickar et al. (2024) seeks to allocate patients to the optimal treatment for their lower back pain. However, the necessary target variable "optimal treatment for lower back pain" is not available in the dataset used. The ML model instead predicts which treatments patients were referred to for their lower back pain based on a combination of patient questionnaires and EHR data. There is no follow-up data to confirm that these are the most effective or optimal treatments, creating a functional limitation of the model in achieving the clinical goal. This is noted by the authors who could "not conclude whether the triaging to the treatment was the correct triaging in terms of the treatment with the highest benefit to the patient and whether the patient indeed was successfully treated". The authors indicate that the reason they did not use a dataset with the necessary target variable is that it was not available to them. They note that future longitudinal studies which include treatment results would be required to create such a dataset. This implies that it is not currently possible to achieve the clinical goal with a machine learning based CDS tool due to the lack of required data.

In each of the cases above, the authors had access to a specific dataset. However, upon analysis, this dataset was determined to be insufficient to build a model whose function can address the clinical goal. Although the authors describe how datasets with appropriate target variables could be created, it is determined in each case that the time and resource commitment required to do this is prohibitive. Thus, the authors were unable to obtain the necessary data to build a model that achieved the function required to support the clinical goal. This did not result in abandonment of development in any of the three cases. Instead, the use of the available dataset was framed as a design constraint on the model, and development continued.

Alternative goals were used to justify continued development of CDS tools

For Fudickar et al. (2024), de Hond et al. (2024), and Abdel-Hafez et al. (2023) misalignment of the available dataset with the main clinical goal meant that the model functionality necessary to achieve the clinical goal was not possible. However, in each case the main clinical goal was not the only goal being pursued by the authors. Achievement of one or more of the other goals defined in the articles was therefore used as justification for the continued development of the CDS tool. In the following section, we examine the types of goals that were used to justify continued development despite a lack of construct validity to the main clinical goal.

Partial achievement of the main clinical goal The clinical goals defined in the reviewed papers generally describe the improvement of an existing clinical process. Achievement of this improvement often requires that multiple clinical sub-goals are achieved simultaneously. Abdel-Hafez et al. (2023) and de Hond et al. (2024) both specify two sub-goals of the main clinical goal, one relating to the increase in quality of the clinical process through improved decision making, the other relating to the increase in efficiency of the process through automation. In both cases, although the available EHR datasets were not sufficient to achieve the sub-goals relating to process quality, they could still be used to achieve the efficiency related clinical sub-goals. Therefore, the authors proceeded with the development of the CDS tool. The potential interdependency of these sub-goals is not reflected on in the re-prioritization of goals.

Achievement of technical sub-goals With the development and use of AI-based clinical decision support tools for hospital settings still in early stages, the successful design of a CDS tool for a particular purpose or setting often requires achievement of one of more technical sub-goals, which themselves require design and testing. When considered in isolation, these technical goals often do not depend on the construct validity of the target variable to the clinical goal, and so are able to be pursued with the available dataset. For example, Fudickar et al. (2024) aim to optimise the predictive performance of their CDS tool. Therefore, they define a technical goal to determine the improvement in the predictive performance of a model obtained by using an additional source of data. To achieve this goal, the target variable used does not need to have construct validity with the clinical goal. It just needs to be kept consistent between models, so that the difference can be measured. The implicit assumption is that this performance increase will remain when a dataset with an appropriate target variable becomes available. Therefore, using the available dataset is justified by the pursual of this technical goal instead of the original clinical end goal.

Achievement of developers' goals The development of a technical solution such as a CDS tool requires the negotiation of the goals and priorities of stakeholders including clinicians, patients and hospital administrators. These form the basis for clinical goals and are the focus of considerations related to the success or failure of the design functionality.

However, the developers of CDS tools are themselves stakeholders and they have their own goals and priorities which can be distinct from the clinical goal. For example, where the developers are authors of academic publications, making a scientific contribution by addressing a gap in the literature is a goal of the development that is distinct from that of a user of the tool. Examples of scientific contributions made by the papers include determining the predictive value of a data source (Fudickar et al. 2024), assessment of process automation feasibility (Abdel-Hafez et al. 2023), and applying a CDS tool to a novel application (de Hond et al. 2024). None of these contributions required that the clinical goal was able to be achieved by the CDS tool functionality. Thus it was not necessary for the authors to prioritize ensuring tool functionality. In fact, it is important for continued scientific progress to report on attempts to develop CDS tools where functionality was not achieved, so that challenges and limitations found can be resolved by other researchers in future iterations. In such cases it is necessary that the limitations of the developed tool are clearly communicated, which we discuss next.

Stated functionality of the developed models obscured the existence of their limitations

In their article, Fudickar et al. (2024) discuss the limitation of their design in relation to the clinical goal. The authors note that the ground truth data on optimal treatments for lower back pain was not available for the development of their model. They note that in the future, longitudinal studies could be conducted to develop a dataset with an appropriate target variable, and that this could be used to train future versions of the model. Despite their awareness of the limitations, the stated function of the developed model, as it was written in the abstract, did not convey those limitations. The abstract states in its conclusion that the study indicates that extracting additional data from EHR text increases the accuracy of ML models in "suggesting optimal treatments for individual patients with [lower back pain]." This inaccurately conveys that the model's functionality supports the clinical goal by predicting optimal treatments, despite the fact that this is not the case.

Abdel-Hafez et al. (2023) and de Hond et al. (2024) both discuss the functional limitations caused by the lack of available ground truth data, and their attempts to address these, in detail. Both indicate the importance of considering data limitations in future iterations of the models. In each of these articles, the stated functionality of the developed models do provide factually accurate descriptions of what the models do. However, the language obscures the existence of the model limitations, through the omission of critical contextual information, resulting in a misabstraction. Abdel-Hafez et al. (2023) state that a model was developed to "automate the process of categorizing medical referrals based on clinical prioritization criteria guidelines." However, the categorization does not necessarily adhere to these guidelines, as the model was optimized using historical data with categories assigned prior to the development of the guidelines, and containing errors and biases as a result. de Hond et al. (2024) state that a model was developed to "identify patients with cancer at risk for depression within one month of chemo- or radiotherapy treatment". However, due to a lack of ground-truth data, the model is not trained to identify all patients with depression, only those formally diagnosed.

This demonstrates how refocusing the design purpose away from the main clinical goal not only sees it unaddressed by the tool functionality, but leads to situations where clear and precise reporting of the relationship between the main clinical goal and the tool's capabilities and limitations is not prioritized in the communication of the research outputs.

Discussion

In this study, we investigated why machine learning tools without the necessary functionality to achieve clinical goals are being built for clinical decision support applications. Our research focused on CDS tools where functionality issues occurred due to construct invalidity of the ML model target variable.

In our case studies, we found that design decisions involving the selection of the invalid model target variable were dictated by the 'real-world' dataset available to the developers. Opportunities to develop construct-target aligned datasets based on functionality requirements were prevented by time and resource constraints. We observed that each article had defined multiple other goals, some related to, and some distinct from the main clinical goal. Goals that could be achieved using the available dataset were used to justify the continued development of the ML-based CDS tool, despite known functionality issues. Finally, we found that imprecise terminology in the descriptions of tool functionality obscured the functional limitations. A detailed reading of the article was required to understand that the developed CDS tool would not address the main clinical goal, increasing the interpretive labor for users of these studies (Kou 2024).

Losing sight of the clinical goal

When the focus of the development of a solution or tool shifts away from the original goal, the problem underpinning that goal can become 'orphaned' from the solution. That is, the solution is no longer addressing that problem (Siffels and Sharon 2024). Our results illustrate that this is a risk in current clinical ML research practices. Problems whose solutions require upfront investment or effort in obtaining necessary data are orphaned, while problems and goals that can be achieved with existing data are prioritised. In the examples we analyzed we observed this leading to trade-offs between quality and efficiency.

In design best practices, it is common to adjust the design goal during problem refinement (Roozenburg and Eekels 1995). Designers iteratively review their goals to ensure they address a real underlying problem. As designers engage more deeply with the context through research, prototyping, or stakeholder engagement, they frequently encounter new insights that challenge their original problem understanding. Sometimes, the initial problem was misunderstood, requiring the problem definition to be refined and the goals updated. However, in the cases we observed, the legitimacy of

the problem underpinning the main clinical goal was never challenged or deemed necessary to refine. So when the goals were updated due to limitations of the available dataset, the initial problems still remained, existing but unresolved. If development proceeds and these problems continue to be deprioritized, they may become orphaned from the design process and fail to be addressed by the eventual implemented solution (Siffels and Sharon 2024).

In the case studies examined, the shift in focus from the main clinical goal to more achievable sub-goals occurred without consideration of potential dependencies between each sub-goal and their underlying problems, or the relative importance of addressing these problems for the clinical stakeholders. In each of the three cases we analyzed, the original clinical goal set out to improve the quality of clinical decisions, by improving on the current state of disease diagnosis or treatment allocation, as well as their speed. However, the available dataset did not allow for quality improvement to be achieved, and so instead, efficiency through process automation became the focus. The authors of the articles did not discuss whether the quality improvement was necessary to achieve meaningful efficiency gains (e.g. to reduce time to review CDS outputs), or whether the efficiency gains brought about by automation were desirable if the quality improvements could not be achieved. This illustrates a risk in the digitization of processes, where goals of quality improvement are abandoned because the costs or effort to achieve them are seen as prohibitive. Meanwhile, cheaper and easier goals remain, such as those related to efficiency improvements through automation.

When a tool is developed that only addresses efficiency, not quality, the justification is often that it is better than nothing. However, although it may fulfill short-term needs, failing to address the broader long-term goals will result in services that cease to improve over time. Focusing on efficiency while neglecting to invest in quality can lead to further deterioration of processes, as historical standards become calcified and new improvements become difficult to implement within the constraints of outdated systems (Wetter 2007; Van den Hooff and Hafkamp 2018).

A lack of clarity in the communication of the actual functionality of CDS tools, creates a risk that developers and users may lose track of functional limitations, or of the original clinical goal, in future design iterations. If knowledge of the limitations is lost during the design and development process, they could then resurface at a later phase, either during implementation or operation. This creates risks of implementation failure, or in harms to stakeholders during operation of the tool. If the original clinical goals are forgotten, then attempts to realize these goals, either by using ML CDS technology or by other means, may cease. This would result in a clinical problem that remains unaddressed in the long term.

Recommendations for CDS design research

To avoid potential negative impacts to healthcare services, there should be a focus on developing goal-led designs, instead of data or technology-driven designs. Based on our findings, we present three recommendations for clinical ML research to mitigate the risk of developing ML-based CDS tools with target variable invalidity.

Recommendation 1: Invest in dataset design and creation Ideally, the prioritization and specification of goals should drive the creation of datasets, but there is often limited focus on developing appropriate datasets. Although a specific dataset may be needed to address the clinical problem, obtaining it can conflict with other goals or constraints. This leads to adjustments in goals to match what is achievable with available data. When the necessary dataset is unavailable, the clinical problem becomes deprioritized. There is often an implicit assumption by developers that functionality issues will be addressed in future iterations as data quality improves, despite the lack of clear roadmaps for obtaining and integrating these improved datasets. Building models with poor-quality data while waiting for the right dataset not only wastes time, but also risks implementing ineffective tools. Hence, the clinical ML research field should have an increased focus on enabling the design, creation, and use of suitable datasets for achieving clinical goals. Data specifications should be defined in relation to the end goal, not pre-defined by available datasets to inform goal selection.

Recommendation 2: Define clear success and failure criteria for solutions Without established success criteria, or specified conditions under which the project should be terminated, the design of a tool is able to continue indefinitely. Continuous redesign according to increasingly onerous constraints is done to achieve sub-goals, even as these changes obstruct or even prohibit achievement of the required functionality for the main clinical goal. This can result in a significant investment of time and resources without ever developing a solution that addresses the main goal. Therefore, prior to commencing the design of ML-based CDS tools, clear definitions of success should be specified. Additionally, indicators of success feasibility, or infeasibility, should be identified so that it is clear when development can no longer continue in a particular direction.

Recommendation 3: Make clear the distinction between value to users and scientific value In many cases, the potential effects or risks of reprioritizing goals due to feasibility issues are not as salient to the research team developing the tool as they are to the users. Often, development teams or organizations have other goals distinct from those of the users, arising from financial or reputational incentives both in industry and academia. For example, in academic literature, achievement of the author specific goals of contributing to the scientific literature is generally independent of the clinical stakeholders' goals. Therefore, there is no inherent incentive to reflect on the relative importance that clinical stakeholders may assign to each of the clinical goals and sub-goals. Nor is there incentive to consider or how users they may be affected if the developed tool is not able to achieve the clinical goal. To create such an incentive, proof of the alignment of research outcomes with clinical end goals needs to be valued more in the scientific literature. Evaluation of the achievement of the clinical goals should be reported on specifically and precisely, making clear the

ways they are different and distinct from the scientific goals.

Conclusion

Our research presents an exploratory case study analysis of how functionality issues in ML-based CDS tools can arise due to construct invalidity of the model target variables. The three case studies in our analysis have shown how the pursuit of alternative goals can be used to justify the continued development CDS tools which have known construct invalidity issues, resulting in tools with functionality issues. While our research surfaces the existence of this phenomena, we are unable to infer its prevalence. This is due to the small sample size of our core articles, the diversity of clinical applications they present. Therefore further research is needed to understand the extent to which this phenomena occurs, and if there are trends which or across different medical applications. The methods presented in this article provide a contribution to this future research, as they outline an analytic framework that can be adapted to, and applied in, systematic reviews of specific CDS tool applications.

Although our research focuses on ML-based clinical decision support tools, the conditions driving the development of non-functional tools - unavailable data, and competing priorities and goals - are not specific to healthcare. Therefore, it is possible that functionality issues observed in sectors such welfare, education and criminal justice may be emerging due to these same conditions. Future work examining functionality failure in different sectors is needed to determine the generalizablity of these findings.

In this paper we have shown how data-driven development of CDS tools can result in target variable invalidity, and that this creates functionality issues that prevent the tool from meeting the needs of users and achieving the clinical goal. We show that developer knowledge of these functionality issues does not necessarily prevent the development of a tool. Instead, alternative goals are used to justify continued development, shifting focus away from the clinical goal and risking the original clinical problem being orphaned from the CDS tool solution.

Acknowledgments

The research was conducted as part of the Gravitation Program Public Values in the Algorithmic Society (AlgoSoc), which is funded by the Dutch Ministry of Education, Culture and Science (OCW) under project number 024.005.017. The authors thank Eva de Winkel, Íñigo de Troya, and Daniel Anadria for providing feedback on earlier versions of this manuscript.

References

Abdel-Hafez, A.; Jones, M.; Ebrahimabadi, M.; Ryan, C.; Graham, S.; Slee, N.; and Whitfield, B. 2023. Artificial intelligence in medical referrals triage based on Clinical Prioritization Criteria. *Frontiers in Digital Health*, 5: 1192975.

Abell, B.; Naicker, S.; Rodwell, D.; Donovan, T.; Tariq, A.; Baysari, M.; Blythe, R.; Parsons, R.; and McPhail, S. M. 2023. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. *Implementation Science*, 18(1): 32.

Acemoglu, D. 2025. The simple macroeconomics of AI. *Economic Policy*, 40(121): 13–58.

Alaa, A.; Hartvigsen, T.; Golchini, N.; Dutta, S.; Dean, F.; Raji, I. D.; and Zack, T. 2025. Medical Large Language Model Benchmarks Should Prioritize Construct Validity. arXiv:2503.10694.

Anglin, K. 2024. Addressing Threats to Validity in Supervised Machine Learning: A Framework and Best Practices for Education Researchers. *AERA Open*, 10.

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; Heidari, H.; Ho, A.; Kapoor, S.; Khalatbari, L.; Longpre, S.; Manning, S.; Mavroudis, V.; Mazeika, M.; Michael, J.; Newman, J.; Ng, K. Y.; Okolo, C. T.; Raji, D.; Sastry, G.; Seger, E.; Skeadas, T.; South, T.; Strubell, E.; Tramèr, F.; Velasco, L.; Wheeler, N.; Acemoglu, D.; Adekanmbi, O.; Dalrymple, D.; Dietterich, T. G.; Felten, E. W.; Fung, P.; Gourinchas, P.-O.; Heintz, F.; Hinton, G.; Jennings, N.; Krause, A.; Leavy, S.; Liang, P.; Ludermir, T.; Marda, V.; Margetts, H.; McDermid, J.; Munga, J.; Narayanan, A.; Nelson, A.; Neppel, C.; Oh, A.; Ramchurn, G.; Russell, S.; Schaake, M.; Schölkopf, B.; Song, D.; Soto, A.; Tiedrich, L.; Varoquaux, G.; Yao, A.; Zhang, Y.-Q.; Albalawi, F.; Alserkal, M.; Ajala, O.; Avrin, G.; Busch, C.; de Carvalho, A. C. P. d. L. F.; Fox, B.; Gill, A. S.; Hatip, A. H.; Heikkilä, J.; Jolly, G.; Katzir, Z.; Kitano, H.; Krüger, A.; Johnson, C.; Khan, S. M.; Lee, K. M.; Ligot, D. V.; Molchanovskyi, O.; Monti, A.; Mwamanzi, N.; Nemer, M.; Oliver, N.; Portillo, J. R. L.; Ravindran, B.; Rivera, R. P.; Riza, H.; Rugege, C.; Seoighe, C.; Sheehan, J.; Sheikh, H.; Wong, D.; and Zeng, Y. 2025. International AI Safety Report. arXiv:2501.17805.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 173–184.

Bongurala, A. R.; Save, D.; Virmani, A.; and Kashyap, R. 2024. Transforming health care with artificial intelligence: redefining medical documentation. *Mayo Clinic Proceedings: Digital Health*, 2(3): 342–347.

Buonora, M. J.; Axson, S. A.; Cohen, S. M.; and Becker, W. C. 2024. Paths forward for clinicians amidst the rise of unregulated clinical decision support software: our perspective on narxcare. *Journal of General Internal Medicine*, 39(5): 858–862.

Coston, A.; Kawakami, A.; Zhu, H.; Holstein, K.; and Heidari, H. 2023. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)*, 690–704.

Dancu, A. 2021. The "Toeslagen Affair:" why did the Dutch government resign last week? https://northerntimes.nl/the-

toeslagen-affair-why-did-the-dutch-government-resign-last-week/. [Accessed 22-05-2025].

de Hond, A.; van Buchem, M.; Fanconi, C.; Roy, M.; Blayney, D.; Kant, I.; Steyerberg, E.; and Hernandez-Boussard, T. 2024. Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study. *JMIR Medical Informatics*, 12(1): e51925.

de Troya, Í.; Kernahan, J.; Doorn, N.; Dignum, V.; and Dobbe, R. 2025. Misabstraction in Sociotechnical Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1829–1842.

Dobbe, R. 2025. AI Safety is Stuck in Technical Terms – A System Safety Response to the International AI Safety Report. arXiv:2503.04743.

Drost, E. A. 2011. Validity and reliability in social science research. *Education Research and Perspectives*, 38(1): 105–123.

Egan, P. 2019. State of Michigan's mistake led to man filing bankruptcy. https://eu.freep.com/story/news/local/michigan/2019/12/22/government-artificial-intelligence-midas-computer-fraud-fiasco/4407901002/. [Accessed 22-05-2025].

Fudickar, S.; Bantel, C.; Spieker, J.; Töpfer, H.; Stegeman, P.; Schiphorst Preuper, H. R.; Reneman, M. F.; Wolff, A. P.; and Soer, R. 2024. Natural Language Processing of Referral Letters for Machine Learning–Based Triaging of Patients With Low Back Pain to the Most Appropriate Intervention: Retrospective Study. *Journal of Medical Internet Research*, 26: e46857.

Guerdan, L.; Coston, A.; Holstein, K.; and Wu, Z. S. 2023a. Counterfactual Prediction Under Outcome Measurement Error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1584–1598.

Guerdan, L.; Coston, A.; Wu, Z. S.; and Holstein, K. 2023b. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 688–704.

Kapoor, S.; Cantrell, E. M.; Peng, K.; Pham, T. H.; Bail, C. A.; Gundersen, O. E.; Hofman, J. M.; Hullman, J.; Lones, M. A.; Malik, M. M.; Nanayakkara, P.; Poldrack, R. A.; Raji, I. D.; Roberts, M.; Salganik, M. J.; Serra-Garcia, M.; Stewart, B. M.; Vandewiele, G.; and Narayanan, A. 2024. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18): eadk3452.

Klee, M. 2023. She Was Falsely Accused of Cheating With AI – And She Won't Be the Last. https://www.rollingstone.com/culture/culture-features/student-accused-ai-cheating-turnitin-1234747351/. [Accessed 22-05-2025].

Kou, T. 2024. From Model Performance to Claim: How a Change of Focus in Machine Learning Replicability Can Help Bridge the Responsibility Gap. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1002–1013.

Larson, J.; Angwin, J.; Kirchner, L.; and Mattu, S. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accessed 22-05-2025].

Lekadir, K.; Quaglio, G.; Tselioudis Garmendia, A.; and Gallin, C. 2022. Artificial Intelligence in Healthcare – Applications, Risks, and Ethical and Societal Impacts. Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

Leveson, N. G. 2017. Rasmussen's legacy: A paradigm change in engineering for safety. *Applied Ergonomics*, 59: 581–591.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2024. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4): 1085–1115.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2025. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25.

Page, M. J.; Moher, D.; Bossuyt, P. M.; Boutron, I.; Hoffmann, T. C.; Mulrow, C. D.; Shamseer, L.; Tetzlaff, J. M.; Akl, E. A.; Brennan, S. E.; et al. 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.

Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 959–972.

Roozenburg, N. F.; and Eekels, J. 1995. *Product design: fundamentals and methods*. Wiley.

Saxena, D.; Jung, J.-Y.; Forlizzi, J.; Holstein, K.; and Zimmerman, J. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25.

Siffels, L. E.; and Sharon, T. 2024. Where technology leads, the problems follow. Technosolutionism and the Dutch contact tracing app. *Philosophy & Technology*, 37(4): 125.

Staub, L.; van Giffen, B.; Hehn, J.; and Sturm, S. 2023. Design Thinking for Artificial Intelligence: How Design Thinking Can Help Organizations to Address Common AI Project Challenges. In Degen, H.; Ntoa, S.; and Moallem, A., eds., *HCI International 2023 – Late Breaking Papers*, 251–267.

Thais, S. 2024. Misrepresented Technological Solutions in Imagined Futures: The Origins and Dangers of AI Hype in the Research Community. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1455–1465.

Van den Hooff, B.; and Hafkamp, L. 2018. Dealing with dissonance: misfits between an EHR system and medical work practices. In *38th International Conference on Information Systems: Transforming Society with Digital Innovation, ICIS 2017*.

Westerbeek, L.; Ploegmakers, K. J.; De Bruijn, G.-J.; Linn, A. J.; van Weert, J. C.; Daams, J. G.; van der Velde, N.; van Weert, H. C.; Abu-Hanna, A.; and Medlock, S. 2021. Barriers and facilitators influencing medication-related CDSS acceptance according to clinicians: a systematic review. *International Journal of Medical Informatics*, 152: 104506.

Wetter, T. 2007. To decay is system: The challenges of keeping a health information system alive. *International Journal of Medical Informatics*, 76: S252–S260.

WHO Guidance. 2021. Ethics and governance of artificial intelligence for health. *World Health Organization*.

Wong, A.; Otles, E.; Donnelly, J. P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penoza, C.; Ghous, M.; and Singh, K. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8): 1065–1070.