



Integrating AI into healthcare systems: A multivocal literature review

Giulio Mallardi^{ID*}, Fabio Calefato^{ID}, Luigi Quaranta^{ID}, Filippo Lanubile^{ID}

Department of Computer Science, University of Bari, Bari, Italy

ARTICLE INFO

Keywords:

Model deployment
ML pipeline
ML-enabled system
Health informatics

ABSTRACT

The integration of artificial intelligence (AI) into healthcare systems promises to improve patient care, enhance operational efficiency, and facilitate personalized medicine. The goal of this paper is to provide a comprehensive review of the current challenges that hinder the seamless adoption of AI in healthcare. Additionally, the paper aims to delineate the best practices for achieving optimal integration of AI within the medical domain. To achieve these objectives, we employ a Multivocal Literature Review (MLR), a systematic literature review methodology that incorporates both peer-reviewed publications and non-peer-reviewed sources, including technical blog posts and white papers. Substantial evidence in the literature points to challenges related to data quality, model bias, interoperability, patient privacy, and the susceptibility of AI systems to adversarial attacks. Additionally, there is growing awareness of challenges such as the distributional shift between training and production data, as well as the critical need for continuous monitoring and retraining of AI models within dynamic clinical settings. Based on our review, we advocate for the adoption of best practices aimed at mitigating the identified challenges, including rigorous model evaluation, standardization of data practices, and promotion of interdisciplinary collaboration. Furthermore, we emphasize the need for responsible AI that aligns with principles of fairness, transparency, security, and reliability, underscoring the importance of multi-stakeholder engagement.

1. Introduction

The healthcare domain is undergoing a profound transformation driven by the integration of artificial intelligence (AI) in several key areas, including patient care, diagnosis, and operational efficiency. However, this integration is complex and introduces a growing number of challenges. As AI becomes more pervasive in healthcare, there is a critical requirement for a structured approach to manage the end-to-end lifecycle of AI models, encompassing development, deployment, and ongoing maintenance.

This complexity has motivated a growing body of research aimed at identifying the barriers and enablers of AI integration across the healthcare domain. While several surveys have explored specific technical or clinical dimensions — such as diagnostic tools, model interpretability, or data privacy — many focus on isolated stages of the AI pipeline. Moreover, most existing reviews are limited to academic sources, overlooking insights from real-world implementation and practice.

This paper aims to provide a comprehensive analysis of the key challenges that hinder seamless AI adoption in healthcare and then outline the best practices for integrating AI models into healthcare systems. In particular, the focus is on the integration of models based on data-driven AI approaches, including machine learning (ML) and deep learning (DL).

* Corresponding author.

E-mail addresses: giulio.mallardi@uniba.it (G. Mallardi), fabio.calefato@uniba.it (F. Calefato), luigi.quaranta@uniba.it (L. Quaranta), filippo.lanubile@uniba.it (F. Lanubile).

<https://doi.org/10.1016/j.smhl.2025.100631>

Received 6 February 2025; Received in revised form 30 July 2025; Accepted 13 December 2025

Available online 16 December 2025

2352-6483/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Table 1
Comparison of selected related reviews on AI in healthcare.

Study	Year	Scope	Source types
Zhang, Xing, Zou, and Wu (2022)	2022	Data-centric view of ML in healthcare	Academic literature
An, Rahman, Zhou, and Kang (2023)	2023	ML algorithms in healthcare	Academic literature
Waring, Lindvall, and Umeton (2020b)	2020	AutoML in healthcare	Academic literature
Qayyum, Qadir, Bilal, and Al-Fuqaha (2021)	2021	Security and privacy of ML in healthcare	Academic literature
Khattak et al. (2023)	2023	MLOps in healthcare	Academic literature
This work	2024	AI integration in healthcare	Academic + Grey literature

While there is an extensive body of research on AI applications in healthcare, existing reviews tend to concentrate on specific technologies or narrowly defined clinical use cases — such as diagnostic tools, predictive modeling, or security threats — often without addressing the broader organizational and lifecycle challenges of real-world AI integration. Moreover, the majority of these surveys are based solely on academic literature, which may not fully capture evolving practices and implementation barriers encountered in operational settings. This creates a gap in understanding how AI models are actually deployed, maintained, and governed within healthcare ecosystems. Our study addresses this gap by adopting a multivocal approach that integrates both academic and grey literature, offering a more practice-informed and comprehensive synthesis.

To achieve our goal, we conducted a Multivocal Literature Review (MLR), a systematic review that combined insights from peer-reviewed academic publications and non-peer-reviewed sources like technical blog posts and white papers. With this approach, we synthesize perspectives from a diverse range of stakeholders — including healthcare professionals, data scientists, IT specialists, and policymakers — to provide a comprehensive understanding of the current state of AI integration in healthcare.

The central themes that emerged from our MLR encompass model deployment, monitoring, governance, ethics, and regulatory considerations, collectively offering a holistic overview of the inherent intricacies of AI adoption in healthcare. For each theme, we delineate a set of corresponding challenges and best practices. Among the challenges frequently discussed in the literature are technical issues such as distributional shifts and adversarial attacks. A distributional shift occurs when the statistical properties of input data in a production environment differ from those seen during model training, potentially leading to degraded performance. Adversarial attacks involve subtle, intentionally crafted perturbations to input data that deceive AI models into making incorrect predictions, posing serious risks to robustness and security—particularly in clinical contexts.

This paper is organized as follows. Section 2 examines the integration of AI models into healthcare systems, providing context for the study. Section 3 delineates the methodology used for our MLR. Section 4 presents the key themes that emerged from our analysis, along with associated challenges and best practices. Section 5 synthesizes and discusses the findings, exploring their implications for the field. Finally, Section 6 concludes the paper, summarizing key insights and suggesting directions for future research.

2. Related work

In this section, we position our study within the broader landscape of existing research on AI integration in healthcare. Specifically, we focus on related surveys and literature reviews that we have identified through systematic search. By examining these works, we aim to clarify how our study builds upon and extends current knowledge—particularly through its more integrated, and cross-cutting perspective. Table 1 provides a comparative overview of key characteristics of the related studies, including publication year, thematic scope, and the types of sources analyzed (academic and/or grey literature).

The literature review by Zhang et al. (2022) examines the critical transition of machine learning in healthcare from the development phase to actual deployment, emphasizing the central role of data throughout the process. The authors highlight several innovations aimed at enhancing ML model performance, such as deep generative models and federated learning for dataset augmentation. In addition, the review explores the use of transformer models to handle larger datasets and improve clinical text modeling.

The review conducted by An et al. (2023) investigates machine learning applications in healthcare. The authors focus on supervised and unsupervised techniques for improving the accuracy and efficiency of healthcare data processing, particularly concerning heart rate data transmission. The review discusses the integration of ML with mobile health technologies and notes the need for more research to demonstrate significant improvements in data accuracy and network efficiency.

The literature review presented by Waring et al. (2020b) examines the state-of-the-art for automated machine learning (AutoML) in the context of healthcare. Covering challenges and opportunities, they analyze 101 papers, highlighting AutoML’s ability to match or surpass human performance in various tasks, often more quickly. The review addresses AutoML’s scalability issues beyond small datasets and discusses aspects like automated feature engineering and neural architecture search, illustrating AutoML’s potential to enhance clinical research and health outcomes.

A comprehensive survey by Qayyum et al. (2021) examines the security and robustness of ML/DL models in healthcare applications. The survey addresses critical concerns, including adversarial attacks, data privacy, and model reliability. It categorizes healthcare applications into four areas (prognosis, diagnosis, treatment, and clinical workflow optimization), illustrating the diverse potential of these technologies. The survey also identifies vulnerabilities at different stages of the ML pipeline and proposes a taxonomy of solutions to mitigate security and robustness challenges. This study underscores the need for secure and robust ML frameworks to enable safe deployment in life-critical healthcare settings while also highlighting ongoing research needs in this rapidly evolving field.

[Khattak et al. \(2023\)](#) explore the integration of Machine Learning Operations (MLOps) into healthcare settings to ensure the reliable, efficient, and ethical deployment and maintenance of ML models. This work provides a comprehensive survey of current work and offers guidelines for practitioners on implementing Machine Learning pipelines for healthcare, addressing foundational concepts, setup, long-term monitoring, updating, and ethical considerations.

[Paleyes, Urma, and Lawrence \(2022\)](#) review numerous case studies across various industries, revealing that issues can arise at every stage of the ML workflow, from data collection and preprocessing to model training and deployment. They identified key challenges, including data dispersion, lack of high-quality labels, integration difficulties, and maintaining model performance over time. Their survey also emphasizes the importance of ethical considerations, user trust, and legal compliance throughout all development stages. By systematically mapping these challenges to specific stages of the ML pipeline, the authors provide a structured framework for understanding and addressing the practical difficulties faced by practitioners. This work confirms the intricacies of achieving regulatory compliance with ML solutions deployed in healthcare.

[Chan and Wong \(2023\)](#) identify and examine common barriers to the deployment of clinical prediction models in healthcare settings. They highlight issues such as poor reproducibility and replicability, unfairness in model predictions, and improper development and evaluation of models, emphasizing the need for comprehensive reporting and addressing biases and heterogeneities in clinical data to enhance model reliability and deployment success.

Finally, [Jackson, Yaqub, and Li \(2019\)](#) address the challenges of integrating ML models into healthcare environments using agile methodologies. The authors note that traditional actuarial models, often developed in isolation from client-facing software, lead to difficulties in maintenance and adaptation. To counter these issues, they propose a hybrid agile approach that combines data science and software engineering. This method involves iterative development and continuous experimentation, allowing for dynamic adjustments and thorough debugging at various stages of model deployment. [Jackson et al.](#) emphasize the importance of cross-functional collaboration, with data scientists and engineers working closely throughout the development cycle to ensure seamless integration and high-quality output. The proposed approach aims to facilitate the efficient deployment of predictive analytics models, specifically for healthcare cost estimation, while providing a framework that can be adjusted to similar applications.

In summary, these reviews highlight several common challenges associated with transitioning ML applications from development to real-world deployment in healthcare settings. Notably, they emphasize data quality, security, and integration issues. Moreover, they underscore the persistent demand for research aimed at enhancing the accuracy, efficiency, and reliability of ML-enabled systems for the medical domain.

These contributions address technical, organizational, or ethical concerns in isolation or within narrowly defined scopes. In contrast, our work offers a cross-cutting synthesis of challenges and best practices across the full AI model lifecycle—from development and deployment to monitoring and governance. By integrating findings from both academic and grey literature, we provide a more comprehensive and implementation-oriented perspective that complements and extends previous surveys.

3. Methodology

Selecting an appropriate research methodology for our study required careful consideration of the research objectives and contextual factors, thereby ensuring the validity and reliability of the conclusions drawn.

The Multivocal Literature Review (MLR) is a systematic literature review approach that encompasses both peer-reviewed academic publications, referred to as ‘white’ literature, and non-peer-reviewed technical documents — such as white papers and blog posts — known as ‘grey’ literature.

We opted to conduct an MLR, as this type of review is particularly well-suited for examining the diverse perspectives within the existing literature, including insights from practitioners. Unlike traditional systematic literature reviews (SLRs), which primarily rely on peer-reviewed academic sources to synthesize empirical evidence, MLRs intentionally include grey literature to capture practical experiences and emerging trends. This broader evidence base is particularly valuable in dynamic fields like AI in healthcare, where state-of-the-art developments are often documented outside of academic channels. The inclusion of both white and grey literature enables a richer, more timely, and practice-informed synthesis of the challenges and solutions in this domain ([Garousi, Felderer, & Mäntylä, 2019](#)). Including grey literature helps to reduce publication bias, offers a more balanced perspective of the available evidence, and ensures that the study findings reflect a broader spectrum of research, including studies with null or negative results that might otherwise go unreported. By incorporating grey literature, we aim to present a more complete and nuanced understanding of the challenges and best practices in implementing AI in healthcare systems. The following paragraphs describe the process adopted to conduct the MLR, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure consistency and rigor throughout the review ([Page et al., 2021](#)).

3.1. Research questions

Research questions play a pivotal role in delineating the focus necessary to identify relevant studies and extract pertinent data ([Wohlin, 2014](#)). We employed the PICO mnemonic ([Petersen, Vakkalanka, & Kuzniarz, 2015](#)) (Population, Intervention, Comparison, and Outcome) to deconstruct the review objectives into searchable keywords, facilitating the formulation of precise research questions. Widely used in the medical and social sciences, PICO encourages researchers to systematically consider the key components of their investigations ([Petticrew & Roberts, 2006](#)). [Kitchenham and Charters \(2007\)](#) provide a comprehensive guide for Systematic Literature Reviews (SLR), establishing an alignment of PICO elements to a technical context. Refer to [Table 2](#) for a detailed description of these elements and their application in the current review.

Two research questions were formulated by carefully analyzing each component of the PICO criteria, with particular emphasis on the desired outcome:

Table 2
PICO criteria.

Component	Description	Examples	Synonyms
Population	Can be a specific role, an application area, or an industry domain.	Healthcare	Health care, Health, Medical, Medicine, Telemedicine
Intervention	The methodology, tool, or technology that addresses a specific issue.	Development and deployment techniques for machine learning models	MLOps, DevOps, ML pipeline, AI pipeline, Data pipeline
Comparison	The methodology, tool, or technology in which the Intervention is being compared (if appropriate).	N/A	N/A
Outcome	Factors of importance to practitioners and/or the results that Intervention could produce.	Insights on best practices and challenges	N/A

R1. *What challenges exist in the integration of AI models in healthcare, particularly those affecting the dependable and efficient deployment of these models in real-world settings?*

R2. *What are the current best practices for developing and maintaining machine learning models in healthcare production environments to ensure reliability and efficiency?*

In the following, we describe how we identified a suitable set of search terms, selected relevant search results from two distinct search engines, complemented our search using the snowballing technique, and finally analyzed the collated corpus to derive a catalog of challenges and best practices.

3.2. Definition of search query

In the initial phase of our MLR, we carefully curated a comprehensive set of relevant search terms. These terms were strategically selected to focus our exploration on two primary domains: the development of AI models and the healthcare sector. To maximize the recall of relevant sources, we included synonymous or closely related expressions, as detailed in the PICO table (Table 2). Specifically, we included MLOps, DevOps, ML Pipeline, AI Pipeline, and Data Pipeline as synonyms of the Intervention because they are strongly related to cutting-edge practices and tools for developing and deploying AI models, encompassing the entire lifecycle from data preparation to model maintenance, and emphasizing automation, collaboration, and continuous improvement in AI-enabled systems. The search query formulated based on the PICO criteria is as follows:

(“MLOps” OR “DevOps” OR “machine learning” OR “ML” OR
 “machine learning pipeline” OR “AI pipeline” OR “data pipeline” OR “model lifecycle” OR “model deployment”)
 AND
 (“healthcare” OR “health care” OR “medicine” OR “medical” OR “telemedicine” OR “health”)

We selected these search terms because integrating machine learning models into healthcare systems requires careful consideration of various software engineering aspects. These include operational practices, the use of appropriate tools, the management of model pipelines, and the model lifecycle. Moreover, by emphasizing terms such as ‘machine learning’ and ‘AI,’ we aimed to identify studies that explore how authors have integrated AI models within healthcare, with the specific objective of uncovering best practices and challenges associated with such integration. Note that we did not explicitly include the terms ‘best practices’ and ‘challenges’ since we intended to infer these concepts from the search results rather than imposing their explicit use as an a priori inclusion criterion. This approach allows for a more comprehensive exploration of the literature landscape, enabling us to identify both explicit and implicit discussions on best practices and challenges.

3.3. Results sampling

We performed a systematic search on two distinct search engines. To retrieve white literature (i.e., scientific papers), we searched using Google Scholar, a broad-coverage abstract database. The retrieval process considered papers in which the specified keywords appeared in critical sections such as title, abstract, and metadata. To retrieve grey literature, we performed a generic web search using Google.

Both Google and Google Scholar generated a very high number of search hits. To balance comprehensive coverage with review efficiency, we established specific stopping criteria for the evaluation of search results. In the case of Google, we decided to stop the review process after excluding 25 consecutive results that did not align with the inclusion criteria outlined in the next section. For Google Scholar, the stopping criterion was triggered after examining three consecutive pages without obtaining results that met our inclusion criteria. This approach ensured a pragmatic balance between thoroughness and practicality, streamlining the review process while maintaining the focus on our key objectives.

Table 3
Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
• Documents, papers, social media posts, blog posts. • Content that discusses issues in deploying AI models in healthcare • Content that outlines best practices in deploying AI models in healthcare • Focus on the software engineering perspective	• Results not in English • Results that are courses, workshops, advertisements, or conferences • Results that are not accessible • Results that are out-of-scope • Results that are secondary studies (e.g., surveys, literature reviews)

3.4. Selection criteria

In the pursuit of a thorough investigation into the deployment of AI models in healthcare from a software engineering perspective, it is crucial to establish clear criteria for the selection of sources. These criteria ensure the relevance, coherence, and reliability of the information gathered during the review process.

As shown in Table 3, the inclusion criteria are carefully designed to embrace a diverse range of sources that contribute meaningfully to the topic of AI model deployment in healthcare, also considering the viewpoint of software engineering. The documents considered for inclusion encompass various formats, such as academic and white papers, blog posts, and other relevant materials. The primary emphasis was placed on sources that investigate both the challenges and best practices associated with the integration of AI models into the healthcare environment. The goal was to compile a comprehensive collection of insights that not only address the intricacies of AI deployment but also provide a nuanced understanding of how software engineering principles play a pivotal role in this context. By incorporating documents that explicitly address challenges and best practices or from which these concepts can be deduced, the inclusion criteria are designed to provide a comprehensive viewpoint on the subject matter. To mitigate potential limitations associated with grey literature, we adopted basic credibility checks commonly used in multivocal reviews. These included evaluating the clarity of the argument, traceability of the source (e.g., institutional or organizational backing), and consistency with findings from peer-reviewed studies. Sources that lacked transparency or appeared purely anecdotal in nature were not included.

Equally important are the exclusion criteria, which serve as a filter to ensure the quality and relevance of the selected sources. Exclusion is warranted if a source does not meet certain predetermined conditions. For example, results not in the English language are excluded to maintain uniformity and accessibility. Additionally, courses, workshops, and advertisements are filtered out as these may not provide the depth of insight required for a focused software engineering analysis. Inaccessible results — those behind paywalls or unavailable during data collection (e.g., 404 errors) — are excluded to ensure all included literature can be thoroughly examined and contribute meaningfully to the synthesis.

We also exclude results deemed out of scope. Specifically, we omit sources lacking a technical perspective or those concerning only the viewpoint of AI models without acknowledging the broader system context. Additionally, we exclude papers from conferences outside the computer science domain. Finally, we exclude secondary studies such as surveys and existing literature reviews to prioritize original research and avoid relying on pre-existing syntheses. However, all retrieved secondary studies are treated as related work and discussed in Section 2.

Overall, the inclusion and exclusion criteria are designed to refine the pool of selected documents, ensuring a more focused and pertinent set of insights. The study selection process, adhering to the PRISMA guidelines, is depicted in Fig. 1, which illustrates the flow of records through the selection phases—identification, screening, eligibility assessment, and inclusion. The process is graphically distinguished into two stages: an initial phase dedicated to peer-reviewed literature, and a subsequent phase focused on grey literature retrieved through structured web searches. In the final step, the two streams converge, forming the unified evidence base analyzed in this review.

3.5. Snowballing

The Google search yielded 145 results before reaching the stopping criterion, with only eight results meeting the specified inclusion/exclusion criteria. In parallel, the Google Scholar search produced 119 results, of which 12 were deemed acceptable according to the same criteria. Consequently, the whole search process yielded a total of 20 acceptable results.

Given the limited number of articles identified through these searches, we decided to employ the snowballing technique to broaden the scope of academic articles for potential inclusion. This involved performing both backward and forward snowballing (Wohlin, 2014), which entails examining references within and citations of the initially selected scholarly literature, respectively.

Snowballing led to the identification of 30 additional academic papers.

3.6. Data extraction & Analysis

Following the augmentation of our data collection process via the snowballing technique, we compiled a comprehensive dataset comprising 50 articles for analysis, as shown in Table 4. This dataset comprises eight results from Google Search, 12

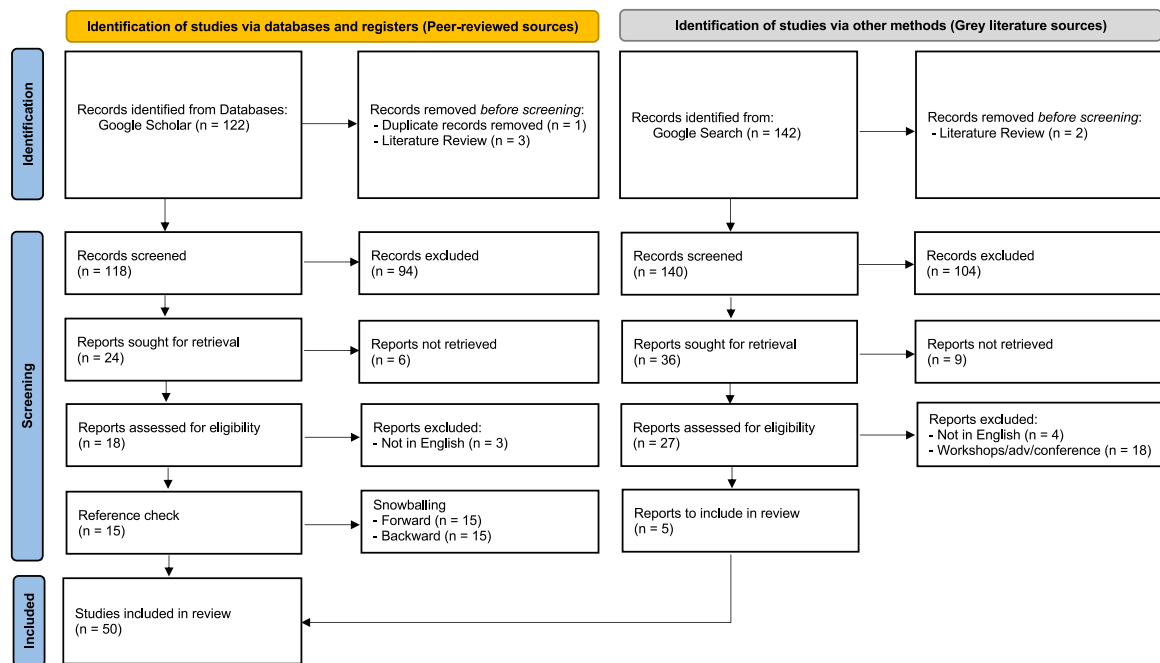


Fig. 1. PRISMA flow diagram for study selection.

Table 4

List of references for the search results.

Source	References
Google results	Eng, Scoville, Gupta, and Aleman (2023), Iyer (2021, 2023), Lavanya (2021) and Reinsch (2024)
Google scholar results	Harris et al. (2022), Huang and Liu (2019), Kaissis, Makowski, Rückert, and Braren (2020), Kompa, Snoek, and Beam (2021), Mena and Gonzalez (2006), Newaz, Haque, Sikder, Rahman, and Uluagac (2020), Ngiam and Khor (2019), Sendak et al. (2020), Tonekaboni, Joshi, McCradden, and Goldenberg (2019), Vellido (2019), Waring, Lindvall, and Umeton (2020a), Stefanie Warnat-Herresthal et al. (2021), Wiens et al. (2019), Willemink et al. (2020) and Yoon, Torrance, and Scheinerman (2021)
Snowballing backward Google Scholar	Antoniou and Mamdani (2021), Chan and Wong (2022), Conor Corbin et al. (2023), Dubovitskaya, Xu, Ryu, Schumacher, and Wang (2017), Finlayson et al. (2021), Mincu and Roy (2022), Nestor et al. (2019), Oner, Cheng, Lee, and Sung (2020), Pfohl, Forciarz, and Shah (2020), Rajkomar, Dean, and Kohane (2019), Salem, Taheri, and Yuan (2018), Sendak, Sirdeshmukh, Ochoa, Premo, Tang, Niederhoffer, Reed, Deshpande, Sterrett, Bauer, Snyder, Shariff, Whellan, Riggio, Gaieski, Corey, Richards, Gao, Nichols, Heintze, Knechtle, Ratliff, and Balu (2022), Tonekaboni et al. (2022), Ulmer, Meijerink, and Cinà (2020) and Verma et al. (2021)
Snowballing forward Google Scholar	Arman Ghavidel and Atashi (2023), Bozkurt et al. (2020), Doyen and Dadario (2022), Finlayson et al. (2019), Jia, Mcdermid, Lawton, and Habli (2021), Jones et al. (2023), Kolyshkina and Simoff (2021), Marulli, Marrone, and Verde (2022), McCoy, Banja, Ghassemi, and Celi (2020), Mccradden et al. (2023), Pandl, Feiland, Thiebes, and Sunyaev (2021), Rosenblatt, Dadashkarimi, and Scheinost (2023), Seneviratne, Shah, and Chu (2019), Štiglic, Kocbek, Fijačko, Zitnik, Verbert, and Cilar (2020) and Varoquaux and Cheplygina (2022)
Total	50 results

from Google Scholar, 15 obtained via backward snowballing, and 15 through forward snowballing. All selected articles meet the inclusion/exclusion criteria outlined in the previous section.

Three team members conducted an independent and detailed review of the selected articles, extracting relevant excerpts focused on recommendations and challenges related to AI use in healthcare. Subsequently, the team gathered for a plenary session to collectively analyze the compiled excerpts. During this session, they collaboratively evaluated problematic excerpts, retaining relevant ones and openly discussing those that prompted disagreement. This process resulted in the creation of a consolidated catalog of excerpts whose relevance was agreed upon by the whole team.

We then conducted a thematic analysis on this catalog following the six-phase approach outlined by [Braun and Clarke \(2012\)](#). We started by familiarizing ourselves with the collected data through careful and repeated reading of each excerpt while making preliminary notes on potential patterns and meanings. In the second phase, we examined the entire catalog to develop an initial set of codes, identifying and labeling relevant features. In the third phase, we organized our codes into potential themes, focusing

on two predetermined analytical domains: *challenges* and *best practices*; this high-level categorization informed our identification of themes pertinent to our research questions. The fourth phase involved reviewing potential themes against both the coded extracts and the entire dataset to ensure coherence and distinctiveness. In the fifth phase, we established clear definitions and names that encapsulated the core content of each theme. The final phase involved crafting the report found in the following section, selecting the most illustrative data extracts, and constructing an analytic narrative that addressed our research questions.

Throughout this process, we adopted a collaborative and iterative approach, continually refining our codes and themes to achieve consensus among team members.

4. Results

Our thematic analysis reveals two distinct categories in the integration of AI within healthcare systems: domain-specific considerations and general AI challenges that manifest in medical contexts. We first examine challenges and corresponding best practices unique to healthcare AI integration (Sections 4.1 and 4.2). We then analyze general AI challenges and practices as they apply to healthcare settings (Sections 4.3 and 4.4). Table 5 lists all the challenges and best practices extracted through the multivocal literature review.

4.1. Healthcare-specific challenges

Integrating AI in healthcare systems presents unique challenges stemming from three domain's distinct characteristics: clinical data and patient treatment considerations, healthcare safety and compliance requirements, and domain-specific security and standardization needs.

4.1.1. Clinical data & Patient treatment

Lack of healthcare data. One of the primary impediments to the advancement and practical application of AI algorithms lies in the scarcity of adequately large, curated, and representative training datasets, particularly those with expert labeling (Mincu & Roy, 2022; Willemink et al., 2020). This challenge is particularly pronounced in the post-acute care domain (Iyer, 2023). Obtaining substantial amounts of standardized clinical data proves to be a formidable task, further complicated by the fact that such data often originates from a limited number of institutions. In this scenario, the data may disproportionately represent common diseases or healthy populations, making it challenging to capture patterns associated with rarer conditions. The reluctance of institutions to share clinical data exacerbates this problem, stemming from concerns about patient privacy and data security. This hesitancy poses a significant barrier to the development of robust deep learning models, which excel in handling large datasets and discerning intricate patterns among variables (Ngiam & Khor, 2019). Consequently, addressing the issue of limited and unevenly distributed healthcare data is pivotal for the successful realization of powerful machine learning models in the medical domain.

Data quality issues. Incorporating AI into healthcare systems faces significant challenges, primarily due to data quality issues. In clinical settings, where precision is paramount, individual errors pose significant challenges, even in models that perform well on average (Tonekaboni et al., 2019). The integrity of data is often compromised by factors such as noise, missing entries, and inherent uncertainties, which in turn erode trust in AI outputs (Tonekaboni et al., 2019). Furthermore, the labor-intensive and costly process of preparing data results in algorithms that generally exhibit poor utility and limited generalization (Willemink et al., 2020). This problem is compounded by confounding variables that may not be consistent across different datasets, leading to unreliable correlations that models learn during training (Chan & Wong, 2022). Additionally, the task of transforming unstructured clinical data into a structured format suitable for machine learning models poses a major barrier to the effective deployment of these technologies. Even with larger datasets, there is not necessarily an improvement in diagnostic accuracy, especially in complex medical conditions such as prodromal Alzheimer's disease (Varoquaux & Cheplygina, 2022). Compounding these issues is the prevalence of mislabeling within datasets, even those curated by professional radiologists, highlighting the need for mechanisms to identify and correct such errors (Pandl et al., 2021). Lastly, systemic biases related to underrepresented groups, exacerbated by gaps in data concerning race and skin tone, underscore the urgent need for more inclusive data collection practices (Doyen & Dadario, 2022).

Issues in patient treatment using AI. Integrating patient treatment data may present challenges (Chan & Wong, 2022; Varoquaux & Cheplygina, 2022). The timing and intensity of treatments can vary; treatments initiated before a prediction are classified as a baseline, whereas those commenced afterward are known as treatment drop-ins. Adjustments in treatment intensity during a patient's care add further complexity (Chan & Wong, 2022). A critical issue arises when treatment data are not included in model training, leading to models that may underestimate risks for untreated patients (Chan & Wong, 2022). Similarly, developing predictive models based only on untreated patients or those with lower outcome probabilities could result in underestimating risks when such models are applied to a broader patient population with varied risk levels (Chan & Wong, 2022). Additionally, models may inadvertently incorporate treatment evidence, such as medical interventions visible in data, which can bias diagnoses. For example, chest X-ray datasets used for diagnosing pneumothorax might include images showing a chest drain, a treatment not present at the time of initial diagnosis, thereby potentially skewing the AI's learning process (Varoquaux & Cheplygina, 2022). Thus, the accurate integration of treatment data is paramount to developing effective and unbiased AI-driven diagnostic tools in healthcare.

Table 5

Challenges (CH) and Best Practices (BP) for integrating AI into healthcare systems, ranked by the number of supporting references.

Code	Source references	Support
CH02 Bias and unfairness in training data	Bozkurt et al. (2020), Doyen and Dadario (2022), Jones et al. (2023), McCoy et al. (2020), Mccradden et al. (2023), Mena and Gonzalez (2006), Mincu and Roy (2022), Oner et al. (2020), Rajkomar et al. (2019), Sendak et al. (2020), Varoquaux and Cheplygina (2022), Wiens et al. (2019) and Willemink et al. (2020)	13
CH05 Discrepancies between production data and training data distribution	Doyen and Dadario (2022), Kompa et al. (2021), Mccradden et al. (2023), Mincu and Roy (2022), Nestor et al. (2019), Varoquaux and Cheplygina (2022) and Yoon et al. (2021)	7
CH06 Issues in model evaluation	Chan and Wong (2022), Doyen and Dadario (2022), Kompa et al. (2021), Tonekaboni et al. (2019), Varoquaux and Cheplygina (2022), Wiens et al. (2019) and Yoon et al. (2021)	7
CH09 Attacks against the AI-enabled system	Doyen and Dadario (2022), Finlayson et al. (2019), Kaissis et al. (2020), Marulli et al. (2022), Newaz et al. (2020), Rosenblatt et al. (2023) and Salem et al. (2018)	7
CH04 Data quality issues	Doyen and Dadario (2022), Harris et al. (2022), Pandl et al. (2021), Tonekaboni et al. (2019), Varoquaux and Cheplygina (2022) and Willemink et al. (2020)	6
CH07 Challenges in interpretability and explainability	Doyen and Dadario (2022), Jia et al. (2021), Kolyshkina and Simoff (2021), Pandl et al. (2021), Vellido (2019) and Yoon et al. (2021)	6
CH03 Lack of healthcare data	Iyer (2023), Mincu and Roy (2022), Ngiam and Khor (2019), Willemink et al. (2020) and Yoon et al. (2021)	5
CH01 Lack of standards	Bozkurt et al. (2020), Finlayson et al. (2021) and Yoon et al. (2021)	3
CH10 Patient privacy violation	Dubovitskaya et al. (2017), Kaissis et al. (2020) and Willemink et al. (2020)	3
CH11 Unintended effects of data anonymization or de-identification	Mincu and Roy (2022), Nestor et al. (2019) and Willemink et al. (2020)	3
CH08 Issues in patient treatment using AI	Chan and Wong (2022) and Varoquaux and Cheplygina (2022)	2
CH12 Technical debt in AI implementation	Harris et al. (2022) and Iyer (2023)	2
BP07 Mitigate bias to ensure equitable outcomes	Antoniou and Mamdani (2021), Conor Corbin et al. (2023), Doyen and Dadario (2022), Finlayson et al. (2021), Jones et al. (2023), McCoy et al. (2020), Mccradden et al. (2023), Mena and Gonzalez (2006), Oner et al. (2020), Pandl et al. (2021), Varoquaux and Cheplygina (2022), Wiens et al. (2019) and Yoon et al. (2021)	13
BP02 Rely on standards	Bozkurt et al. (2020), Doyen and Dadario (2022), Huang and Liu (2019), Iyer (2023), Kaissis et al. (2020), Kolyshkina and Simoff (2021), Nestor et al. (2019), Rajkomar et al. (2019), Vellido (2019) and Verma et al. (2021)	10
BP03 Ensure security and anonymity of patient data	Dubovitskaya et al. (2017), Eng et al. (2023), Kaissis et al. (2020), McCoy et al. (2020), Nestor et al. (2019), Pandl et al. (2021), Reinsch (2024), Salem et al. (2018) and Willemink et al. (2020)	9
BP12 Ensure continuous model monitoring and retraining	Antoniou and Mamdani (2021), Conor Corbin et al. (2023), Finlayson et al. (2021), Iyer (2021), Mincu and Roy (2022), Ngiam and Khor (2019), Reinsch (2024) and Seneviratne et al. (2019)	8
BP01 Comply with healthcare regulations	Harris et al. (2022), Jia et al. (2021), Mincu and Roy (2022), Rajkomar et al. (2019), Sendak et al. (2020), Wiens et al. (2019) and Willemink et al. (2020)	7
BP06 Train models by leveraging data from multiple sites	Dubovitskaya et al. (2017), Huang and Liu (2019), Kaissis et al. (2020), Mincu and Roy (2022), Stefanie Wernat-Herresthal et al. (2021) and Willemink et al. (2020)	6
BP09 Ensure interpretability and explainability	Doyen and Dadario (2022), Jia et al. (2021), Kolyshkina and Simoff (2021), McCoy et al. (2020), Ngiam and Khor (2019), Pandl et al. (2021), Reinsch (2024), Tonekaboni et al. (2019, 2022), Ulmer et al. (2020) and Vellido (2019)	11
BP11 Implement techniques to evaluate the reliability of the model	Antoniou and Mamdani (2021), Eng et al. (2023), Finlayson et al. (2021), Jia et al. (2021), Kompa et al. (2021), Marulli et al. (2022), Mincu and Roy (2022), Pfohl et al. (2020), Varoquaux and Cheplygina (2022), Verma et al. (2021) and Willemink et al. (2020)	11
BP15 Deploy the model silently	Antoniou and Mamdani (2021), Conor Corbin et al. (2023), Eng et al. (2023), Jia et al. (2021), Tonekaboni et al. (2022), Varoquaux and Cheplygina (2022), Verma et al. (2021) and Wiens et al. (2019)	8
BP04 Implement techniques to assure data quality	Eng et al. (2023), Pandl et al. (2021), Sendak et al. (2022) and Willemink et al. (2020)	4
BP05 Implement techniques to mitigate lack of data	McCoy et al. (2020), Mena and Gonzalez (2006), Pandl et al. (2021) and Willemink et al. (2020)	4

(continued on next page)

Table 5 (continued).

Code	Source references	Support
BP14 Prevent enhancement and adversarial attacks	Doyen and Dadario (2022), Finlayson et al. (2019), Kaissis et al. (2020) and Rosenblatt et al. (2023)	4
BP10 Document datasets and models	Bozkurt et al. (2020), Doyen and Dadario (2022), Mincu and Roy (2022), Varoquaux and Cheplygina (2022) and Vellido (2019)	5
BP13 Ensure reproducibility and replicability of ML pipeline	Bozkurt et al. (2020), Doyen and Dadario (2022) and Reinsch (2024)	3
BP08 Include patient treatment information to enhance accuracy of model prediction	Chan and Wong (2022) and Finlayson et al. (2021)	2

4.1.2. Healthcare safety & Compliance

Bias and unfairness. As observed by Harris et al. (2022), “Algorithms can only learn from a digital representation of the world[;] that representation, in turn, cannot encode moral or ethical standards. Unfair outcomes, discrimination against sub-populations, and bias are all reported shortcomings” (Harris et al., 2022). In the healthcare domain, the issues of bias and unfairness are critical, as biased or unfair predictions can pose a substantial risk of patient harm. When deploying ML models in healthcare, researchers must not only focus on technical efficacy but also consider the ethical implications to ensure fair treatment across diverse patient populations (Jones et al., 2023; McCoy et al., 2020; Wiens et al., 2019).

Bias and unfairness can arise when the dataset used to train a model exhibits disparities, such as imbalanced age distributions, unequal proportions of racial and gender groups, differences in the tools or technologies employed, varying prevalence of diseases, or a lack of accurate representation of the target population for which a model is deployed (Arman Ghavidel & Atashi, 2023; Bozkurt et al., 2020; Doyen & Dadario, 2022; Mincu & Roy, 2022; Sendak et al., 2020; Willemink et al., 2020). This challenge, referred to as “Spectrum Bias” by Varoquaux and Cheplygina (2022), underscores the critical need to address discrepancies in data to ensure a model’s generalizability.

Jones et al. (2023) outline three main causes of bias: prevalence disparities, presentation disparities, and annotation disparities. Prevalence disparities occur when there is a causal connection between sensitive attributes (e.g., race, gender, or socioeconomic status) and the prevalence of a disease. Presentation disparities arise when there is a causal link between sensitive attributes and the disease-related physiology or manifestations. Annotation disparities occur when there is a causal connection between sensitive attributes and the criteria used for diagnosis, labeling, or annotation.

A model trained on historical data may inherently learn the prescribing habits of physicians, thus potentially overlooking optimal practices. While machine learning techniques excel at identifying patterns within historical data, they can also shed light on systemic disparities, such as the provision of inferior care to vulnerable groups. Acknowledging and addressing these issues is essential for the ethical deployment of ML models in healthcare settings (Rajkomar et al., 2019). In addition, Mccradden et al. (2023) emphasize that common ways to mitigate bias, such as Randomized Control Trials (RCTs), may fall short in effectively including marginalized groups, highlighting the necessity for a more inclusive approach.

Challenges in interpretability and explainability. The integration of machine learning into healthcare raises critical concerns about accountability, interpretability, and explainability. As ML models, particularly those based on deep learning, increase in complexity, they often become less interpretable, making it difficult for users to understand how decisions are made, thus potentially creating an accountability gap where it is unclear whether responsibility should rest with human operators or the algorithms themselves (Yoon et al., 2021). This lack of transparency is problematic in healthcare, where explainability is crucial for trust and effective decision-making between patients and practitioners. The quality and provenance of data are equally important, as these factors directly influence the reliability of ML outputs and their explainability (Jia et al., 2021). Moreover, while approximation methods like the Shapley value are used to explain model decisions, their computational complexity often necessitates simplifications that may not adequately reflect the model’s reasoning, further complicating trust in such systems (Pandl et al., 2021). Current interpretability methods, including SHAP and LIME, have been found to produce inconsistent results or are not robust to minor input variations, which is particularly concerning in sensitive medical applications (Kolyshkina & Simoff, 2021). These challenges underscore the need for advanced interpretability techniques and stringent data management practices to ensure that ML tools in healthcare are both understandable and reliable (Doyen & Dadario, 2022; Vellido, 2019).

Patient privacy violation risks. When interacting with patient data, the risks of privacy violation are behind the corner. Ensuring that patient data are accessible for AI development involves multiple complex steps, with a necessary focus on minimizing the risk of reidentification through potential cross-linking of different data sources (Willemink et al., 2020). In addition, re-identified patient records are highly sought after by health insurance companies, which may use them to discriminate against individuals based on their health conditions, thus posing a serious risk to patient confidentiality (Kaissis et al., 2020). Together, these issues underscore the critical need for robust data protection measures to safeguard patient privacy as AI technologies become more prevalent in the healthcare sector.

4.1.3. Healthcare security & Standards

Attacks against AI-enabled systems. Integrating AI into healthcare systems presents several challenges related to data security and model robustness against adversarial threats. Adversarial examples, designed to mislead algorithms through minor but intentional data modifications, pose significant risks, particularly when such perturbations remain imperceptible yet effective in deceiving machine learning models (Finlayson et al., 2019). Moreover, algorithmic attacks in machine learning include adversarial attacks, which manipulate inputs to corrupt outputs (e.g., model poisoning), and model-inversion attacks, which derive information about the training data by analyzing the algorithm's behavior, often using generative algorithms to recreate parts of the dataset (Kaissis et al., 2020). Re-identification attacks also highlight vulnerabilities in anonymized datasets, revealing individual identities by linking anonymized data with other identifiable datasets, thereby necessitating rigorous data handling measures such as defacing or skull stripping in medical imaging, which can inadvertently decrease the accuracy and robustness of resulting models (Kaissis et al., 2020). Furthermore, biometric recognition systems face threats from spoofing attacks, where fake biometric data such as fingerprints or iris scans are used to deceive the system (Salem et al., 2018). These challenges underscore the critical need for robust AI systems in healthcare that can withstand various forms of adversarial and spoofing attacks while maintaining data privacy and integrity.

Unintended effects of data anonymization or de-identification. Data anonymization or de-identification are valuable techniques for privacy concerns but can affect model quality. For example, the removal or conversion of DICOM metadata, common during the transition to formats such as NIFTI, results in the loss of crucial information that could enhance AI algorithm development because metadata plays a significant role in understanding and processing data (Willemink et al., 2020). Moreover, de-identification practices often involve altering the timestamps associated with patient records to obscure the actual dates of care, as evidenced in the MIMIC-III dataset where dates are shifted to fictitious years far in the future (Nestor et al., 2019). This modification impedes the ability to train models that can accurately generalize over time since temporal analysis requires consistent and realistic time frames. Furthermore, even with rigorous de-identification, studies have shown that anonymized imaging data can still potentially be used to reconstruct patient faces, highlighting residual risks of re-identification and raising concerns about the true anonymity of such datasets (Mincu & Roy, 2022). These challenges underscore the complexity of ensuring privacy while retaining the informative value of healthcare data in AI applications.

Lack of standards in medical context. As highlighted by Sendak et al. (2020), “Reporting of machine learning models often fails to follow established best practices, and model performance measures are not standardized across publications.” This lack of standardization is especially pronounced in AI-enabled healthcare systems, affecting various phases of development and documentation, as well as practices and nomenclature. For instance, concerning this last aspect, varied naming conventions are used across different sites, undermining interoperability. An illustrative example is the adoption of the International Classification of Diseases (ICD)-9, which is inconsistent with ICD-10, as noted by Finlayson et al. (2021). The lack of standards hampers the consistency of reports and code sharing among research groups, as highlighted by Bozkurt et al. (2020). Even with the use of Electronic Health Records (EHR), the evident absence of standards hinders full interoperability, resulting in challenges when scaling AI-enabled systems. This issue is similarly recognized in the lack of standardized metrics for model or dataset benchmarking, creating obstacles in comparing different products and settings, as emphasized by Sendak et al. (2020).

4.2. Healthcare-specific best practices

Our analysis identifies essential practices for successfully deploying AI in healthcare settings. We organize these practices into corresponding categories that mirror our challenge classification in Section 4.1.

4.2.1. Clinical data & Patient treatment

Implement techniques to assure data quality. Ensuring data quality is paramount in transitioning AI into healthcare systems (Pandl et al., 2021), particularly in the realm of medical imaging (Willemink et al., 2020). Structuring data into homogenized and machine-readable formats is essential, followed by linking images to accurate ground-truth information such as labels or segmentations. Accurate ground truth definition is crucial for building reliable AI models, especially in radiology, where large datasets are required. The ML-DQA (Data Quality Assurance) framework suggested by Sendak et al. (2022) offers a systematic approach to address data quality concerns, particularly in structured EHR data. This framework encompasses three phases: data element pre-processing, ML-DQA checks, and adjudication. In the pre-processing phase, steps like entity resolution and rules-based data transformations ensure consistency and reliability in the data. ML-DQA checks in the second phase assess the completeness, conformance, and plausibility of the data, drawing upon established quality frameworks. Adjudication by clinicians in the final phase confirms data fitness for downstream analyses and model development, reinforcing the importance of interdisciplinary collaboration in ensuring data quality. By implementing these techniques, healthcare systems can enhance the reliability and effectiveness of AI-driven solutions, ultimately improving patient care and outcomes.

Implement techniques to mitigate lack of data. Addressing the challenge of limited data availability in healthcare is crucial for the effective deployment of machine learning models. While collecting data from multiple sources (Pandl et al., 2021) is essential for generating large datasets necessary for training, it can be both expensive and time-consuming (Mena & Gonzalez, 2006). Therefore, there is a growing need for classifiers capable of reliable diagnosis with small amounts of patient data. One potential solution to increase data sample size is the utilization of generative adversarial networks (GANs), which have shown promise in synthesizing realistic images for augmenting training datasets. GANs operate by modeling and sampling from the implicit density function of input

data, comprising a generator network that produces synthetic data and a discriminator network that distinguishes between real and synthetic data. In healthcare, GANs have been employed to augment datasets, alleviate privacy restrictions, and address dataset imbalances. Particularly in clinical datasets, GANs have been instrumental in augmenting or balancing datasets while preserving patient privacy. By leveraging techniques like GANs, healthcare systems can mitigate the challenges posed by the lack of data, facilitating the development of robust machine learning models for improved patient care and outcomes (Willemink et al., 2020).

Include patient treatment information to enhance accuracy of model prediction. The consideration of patient treatment data holds significant importance for enhancing model performance. Various methods are employed to handle treatments effectively within predictive models, ensuring accurate and reliable outcomes (Chan & Wong, 2022; Finlayson et al., 2021). One common approach is to account for treatment data by integrating it directly into the analysis. This can involve several techniques, such as ignoring treatment altogether, focusing analysis solely on untreated patients, or combining treatment with the adverse outcome as a composite measure. By doing so, we obtain a more comprehensive understanding of the patient's condition and the effectiveness of the treatment administered. Another method involves explicitly modeling treatment within the predictive framework. This means incorporating baseline treatments received by patients as predictive variables in the model. By accounting for these factors, the model can better discern the impact of treatment on outcomes, leading to more precise predictions (Chan & Wong, 2022). Additionally, the hypothetical prediction method offers a unique perspective by conducting counterfactual analyses. This involves predicting the risk of outcomes in a hypothetical scenario where treated patients are not given treatment. By simulating these alternate realities, the model can provide insights into the potential outcomes of different treatment strategies, aiding in decision-making processes (Chan & Wong, 2022). Furthermore, estimands play a crucial role in scenarios where the treatment strategy varies across the deployment population. It is imperative to account for these differences to avoid inaccurate predictions. By defining clear estimands, we can ensure that the model accurately captures the effects of treatment within diverse patient populations, thus improving its overall performance and reliability (Chan & Wong, 2022).

Train models by leveraging data from multiple sites. The utilization of data from various sources across multiple sites is necessary for training robust and effective models. Different techniques and approaches have emerged to enable the aggregation and utilization of decentralized data while addressing concerns regarding privacy, security, and data sovereignty. Federated learning has emerged as a prominent method for privacy preservation in next-generation machine learning applications, both within industries and particularly in medical AI contexts. By allowing models to be trained collaboratively across distributed datasets without centralized data aggregation, federated learning ensures that sensitive patient information remains localized while still contributing to model improvements (Kaissis et al., 2020; Mincu & Roy, 2022). Initiatives like the Federated Tumour Segmentation (FeTS) Initiative exemplify ongoing efforts to leverage federated learning for collaborative research across international healthcare institutions. By addressing challenges related to model aggregation and standardizing training data, these initiatives pave the way for more effective collaboration and knowledge sharing in medical research. Cross-silo federated learning expands the scope of collaborative model training by leveraging data from disparate sources, such as EHRs from multiple institutions. Despite potential challenges like imbalanced datasets or limited samples per institution, federated learning models have demonstrated comparable performance to centralized training methods, affirming their viability for real-world implementation in healthcare settings. Community-based federated learning (CBFL) represents a tailored approach to address challenges posed by non-identically independently distributed (non-IID) data, such as EHRs. By clustering EMR data into communities and training separate models per community, CBFL enhances the efficiency and effectiveness of the learning process compared to traditional federated learning methods (Huang & Liu, 2019). Emerging approaches like Swarm Learning (SL) offer decentralized alternatives to traditional federated learning, leveraging distributed machine learning and blockchain technology to enable collaborative model training without the need for a central coordinator. By sharing model parameters via a permissioned blockchain network, SL ensures data sovereignty, security, and confidentiality while facilitating collaborative learning among individual nodes (Stefanie Warnat-Herresthal et al., 2021). Blockchain technology further enhances data management and sharing in healthcare scenarios, offering solutions for primary patient care, data aggregation for research purposes, and connecting different healthcare players for improved patient care. By leveraging blockchain's transparency, traceability, and privacy-preserving features, these scenarios facilitate secure and efficient data exchange among trusted entities while ensuring patient privacy and data authenticity (Dubovitskaya et al., 2017).

4.2.2. Healthcare safety & Compliance

Comply with healthcare regulations. Transitioning AI into healthcare systems necessitates a stringent adherence to healthcare regulations to ensure patient safety and data privacy. Model reliability is paramount, requiring ongoing research to bolster the robustness of machine learning products used in clinical care. This includes active fields of research focused on model updating and validation, ensuring the validity, safety, and efficacy of these products. Incorporating EHR data into machine learning models demands meticulous attention to data quality and transformation processes to ensure initial credibility (face validity) and robust performance. However, challenges persist in adopting established frameworks for assessing EHR data quality, and without interoperability between EHR systems, scaling, and diffusion of machine learning products remain hindered (Rajkomar et al., 2019; Sendak et al., 2020). Regulatory bodies like the FDA and initiatives such as Software as a Medical Device (SaMD) and Assurance of Machine Learning for Autonomous Systems (AMLAS) provide frameworks for validating and regulating AI-driven healthcare technologies, emphasizing the need for adherence to government-required regulatory steps (Jia et al., 2021). Moreover, safeguarding protected health information, as outlined by regulations like the Health Insurance Portability and Accountability Act (HIPAA) (Willemink et al., 2020) or GDPR, is imperative to maintain patient privacy and trust. For this reason, the International Medical Devices Regulator Forum (IMDRF) defined that verification and validation should provide clear guidance for ensuring the

efficacy and safety of AI-driven healthcare technologies, emphasizing the importance of objective evidence to confirm compliance with specified requirements and intended use (Jia et al., 2021). Granlund, Stirbu, and Mikkonen suggest using “locked” state models since regulatory bodies usually approve them and do not change their behavior in a post-market phase, preferring deterministic models that are unable to evolve continuously. With emerging standards and regulatory frameworks reflecting the dynamic nature of AI in healthcare, a comprehensive approach that integrates compliance with healthcare regulations and best practices is essential for the safe and effective integration of AI technologies into clinical care settings.

Mitigate bias to ensure equitable outcomes. In the transition of AI into healthcare systems, mitigating bias is essential to ensure equitable outcomes in clinical settings. Bias often arises due to “dataset shifts”, where differences between training data and real-world application data affect performance. Ensuring representative medical datasets and monitoring demographic and phenotypic breakdowns are vital steps to mitigate biases that may arise from training models on predominantly homogeneous populations (Finlayson et al., 2021). Randomized controlled trials (RCTs) (Antoniou & Mamdani, 2021) remain the gold standard for establishing intervention efficacy, while sensitivity analysis methods help identify potential sources of errors in effect estimates.

Various approaches, including data augmentation, sampling methods, and resampling techniques like SMOTE, help address unbalanced data challenges and promote diverse representation within training datasets. For instance, augmentation strategies may involve the synthetic generation of new data points to enhance minority classes, such as through oversampling or interpolation-based techniques. In medical imaging tasks, generative models — e.g., GANs — can be employed to create realistic representations of underrepresented pathological classes. Similarly, structured clinical data may benefit from methods like SMOTE and its variants, which synthetically generate new samples for sparsely populated subgroups.

Safeguarding against data leakage and utilizing rich, multi-institutional data sets with explicit inclusion criteria further contribute to mitigating biases and promoting equitable outcomes in AI-driven healthcare (Arman Ghavidel & Atashi, 2023). Ultimately, a comprehensive approach that encompasses diverse sampling strategies, rigorous data segregation (Oner et al., 2020), and data augmentation (Chan & Wong, 2022) is essential to foster fairness and inclusivity in healthcare AI applications.

Ensure interpretability and explainability. Ensuring interpretability and explainability in machine learning models for healthcare systems is fundamental to fostering trust, transparency, and confidence in their deployment. The Medical Device Regulation (MDR) mandates manufacturers to design software ensuring repeatability, dependability, and performance, aligning with the imperative to guarantee the safety of medical devices (Reinsch, 2024). In the context of healthcare, temporal explanations pose a challenging task, necessitating ML models to elucidate predictions based on changes in individual patient states. Clinicians emphasize the importance of transparent design, advocating for models that mirror the established methodology of evidence-based medical decision-making, thus enhancing interpretability (Vellido, 2019). Transparency in data collection, analysis, and usage is imperative for building patient trust and ensuring accountability in healthcare ML applications (McCoy et al., 2020). The complexity of healthcare data underscores the need for interpretable ML models. Higher interpretability facilitates easier comprehension and explanation of predictions, empowering healthcare experts to make informed, data-driven decisions. Interpretability approaches can be classified into model-specific and model-agnostic techniques, each serving distinct purposes in providing understandable explanations for ML predictions (Štiglic et al., 2020). Explainability in ML involves illuminating the underlying workings of models, enabling humans to interpret representations of their decisions. This includes methods such as Shapley values, which quantify individual data instances’ contributions to model training, enhancing trustworthiness in healthcare ML applications (Pandl et al., 2021). Interpretability and explainability are crucial for ensuring trust in ML solutions deployed in healthcare settings. A clear understanding of how and why ML-driven decisions are made is essential for integrating these solutions into routine clinical practice, thereby improving patient outcomes and healthcare delivery (Kolyshkina & Simoff, 2021).

Deploy models silently. A cautious approach is necessary before these models can be deployed into patient care. The concept of “silent deployment” (Antoniou & Mamdani, 2021; Wiens et al., 2019) emerges as a pivotal step, allowing for prospective validation in a controlled environment. Here, predictions are made in real-time but not acted upon, providing an opportunity for clinical experts to scrutinize the system’s performance and identify potential errors. This silent testing (Verma et al., 2021) phase serves several critical purposes. Firstly, it ensures that the model functions optimally in a clinical setting, highlighting any discrepancies between training and deployment environments. Moreover, it tests the system’s stability and reliability, uncovering any weaknesses in the data pipeline or infrastructure. Lastly, it addresses human factors concerns, considering the impact of the AI-enabled system on end-users and incorporating qualitative feedback into metric design. Throughout this validation process, three overarching categories are addressed: model functionality and performance, system stability and reliability, and human factors concern (Tonekaboni et al., 2022). Additionally, emphasis is placed on clinical validation, assessing the system’s ability to generate clinically meaningful outputs in its operational environment. Furthermore, the robustness of the system is crucial. It must demonstrate an ability to accurately distinguish between different conditions or recommended treatments without over-reliance on specific input features. This ensures that the model’s performance remains consistent across diverse scenarios, contributing to its reliability and efficacy in real-world healthcare settings.

4.2.3. Healthcare security & standards

Prevent enhancement and adversarial attacks. Safeguarding against adversarial attacks is critical to maintaining the integrity and reliability of algorithms (Doyen & Dadario, 2022). Researchers have delved into various strategies to fortify models against these threats, with a primary focus on retraining algorithms with exposure to adversarial examples (Newaz et al., 2020). By subjecting models to these adversarial scenarios during training, their resilience to such attacks can be significantly enhanced. One proactive

approach involves implementing data fingerprinting techniques, where a unique hash of the data is extracted and stored at the point of capture (Newaz et al., 2020). This fingerprint serves as a reference point for detecting any tampering or alterations to the data, enabling investigators to assess the integrity of information fed into algorithms. Robust optimization methodologies have emerged as effective countermeasures against adversarial attacks, bolstering a model's resistance to manipulation (Rosenblatt et al., 2023). By retraining models with robust optimization techniques, their ability to withstand adversarial perturbations can be notably strengthened. Additionally, detection methods play a key role in identifying and mitigating attacks (Rosenblatt et al., 2023). Augmentation and input transformations are among the techniques that can be leveraged to defend against enhancement attacks, although their effectiveness may vary compared to adversarial attack defenses (Rosenblatt et al., 2023). Data provenance tracking, exemplified by tools like DataLad, offers another layer of defense by monitoring the origin and lineage of data (Rosenblatt et al., 2023). However, it is crucial to recognize that data manipulation could occur before the initiation of provenance tracking. Therefore, replication of studies in independent datasets serves as a robust alternative solution to verify the integrity of findings (Rosenblatt et al., 2023).

Ensure security and anonymity of patient data. Trustworthy machine learning in healthcare must protect patient privacy while enabling diverse, high-quality data collection. This involves privacy-preserving techniques like excluding sensitive data, summarization methods, and addressing mislabeled instances to maintain security and anonymity (Pandl et al., 2021). Ensuring the security and anonymity of patient data in medical devices with machine learning is crucial for compliance with regulations such as the General Data Protection Regulation (GDPR). Manufacturers must implement robust data protection measures, including data anonymization and encryption, to safeguard patient information. Additionally, conducting thorough risk assessments and establishing comprehensive data governance frameworks are essential to identify potential vulnerabilities and ensure compliance with regulatory requirements (Reinsch, 2024). When dealing with radiological imaging data, which is common in AI-based medical devices, considerations must be made regarding patient and imaging characteristics, as well as image processing techniques (Willemink et al., 2020). Proper de-identification of protected health information is crucial to comply with regulations like HIPAA and GDPR, necessitating the removal of identifying information from DICOM metadata and images (Willemink et al., 2020). Several techniques and technologies are available to enhance the security and privacy of patient data. Secure multi-party computation (SMPC) ensures confidentiality by processing encrypted data shares across multiple parties, preventing any single entity from accessing the complete dataset (Kaissis et al., 2020). Advanced de-identification methods, such as homomorphic encryption, provide a means to perform computations on encrypted data without compromising privacy (Kaissis et al., 2020; Salem et al., 2018). DeepZeroID exemplifies this approach, offering a privacy-preserving biometric verification system that utilizes homomorphic encryption to ensure the security and integrity of biometric data (Salem et al., 2018).

Document datasets and models. Clear communication and documentation are paramount in the collaboration between data analysts and medical experts within healthcare machine learning projects. The data analyst may solicit specific information from the medical expert, including a comprehensive outline of medical requirements regarding interpretability and explainability, realistic insights into the limitations and possibilities of interpretability within analytical models, a detailed description of the actual medical decision-making process in clinical settings, and assurance of result verification (Vellido, 2019).

Efforts to improve the reporting and critical evaluation of prediction model studies in healthcare have led to the development of frameworks like the Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (TRIPOD) statement and the Prediction model Risk Of Bias ASsessment Tool (PROBAST) (Doyen & Dardario, 2022; Mincu & Roy, 2022). These initiatives focus on enhancing transparency and reducing bias in such studies. Recently, work has been underway to extend these frameworks to accommodate the increasing use of machine learning in healthcare research. This includes the creation of AI-specific versions of both TRIPOD (TRIPOD-AI) and PROBAST (PROBAST-AI), which aim to address the unique challenges associated with prediction models that use machine learning techniques.

Documenting datasets is crucial for transparency and reproducibility in healthcare machine learning. Suggestions include providing detailed descriptions of dataset provenance, demographics, and content, as well as conducting exploratory data analysis to assess dataset quality. It is also recommended to release transparency artifacts, such as standardized questionnaire templates, alongside research papers (Mincu & Roy, 2022). Furthermore, to enhance reporting quality, authors are advised to adhere to validated guidelines provided by organizations like CONSORT-AI and SPIRIT-AI. These guidelines offer checklists to ensure comprehensive and standardized reporting practices in healthcare machine learning studies (Mincu & Roy, 2022).

Rely on interoperability. In the transition of AI into healthcare systems, relying on established standards is crucial for seamless integration and effective operation. One such standard is EHRs, which have become ubiquitous in modern healthcare. EHRs not only provide clinicians with networked access to comprehensive patient data, including medical images, signals, and textual information but also enable the secondary use of this data for developing machine learning models. These models, leveraging EHR data, have shown promise in various clinical applications, such as mortality risk prediction and sepsis treatment optimization (Huang & Liu, 2019; Iyer, 2023; Reinsch, 2024). Moreover, the adoption of standards like DICOM (Digital Imaging and Communications in Medicine) further enhances electronic data management in medical imaging, ensuring interoperability and facilitating the sharing of medical images across different healthcare systems (Kaissis et al., 2020). Additionally, methodologies like CRISP (Cross-Industry Standard Process for Data Mining) and its machine learning extension, CRISP-ML (Kolyshkina & Simoff, 2021), provide frameworks for algorithm development and testing, ensuring consistency and reliability in machine learning solutions. By adhering to these standards and methodologies, healthcare organizations can optimize the utilization of AI technologies while maintaining data integrity, interpretability, and interoperability, ultimately enhancing patient care and outcomes.

4.3. General AI challenges in healthcare context

Beyond healthcare-specific challenges, implementing AI in medical settings encounters common machine learning difficulties that, while not unique to healthcare, require careful consideration given the critical nature of medical applications.

Discrepancies between production and training data distributions. In transitioning AI into healthcare systems, a significant challenge is ensuring model robustness amid inevitable data shifts between training and deployment environments. Kompa et al. (2021) and Mincu and Roy (2022) highlight how these data shifts, including changes in patient demographics, clinical practices, or even temporal changes in healthcare practices, can lead to a deterioration in model performance. The consequences of these shifts are not trivial; models that perform well in controlled test settings often fail in real-world scenarios due to mismatched data characteristics, as demonstrated by different researches (Doyen & Dadario, 2022; Nestor et al., 2019; Varoquaux & Cheplygina, 2022). Furthermore, factors like the unbalanced class distribution commonly found in healthcare data exacerbate these issues, impacting the reliability of mortality predictions and other clinical outcomes (Arman Ghavidel & Atashi, 2023). As a result, without ongoing monitoring and adaptation to these changes, deployed models may not only become less accurate over time but could also potentially lead to harmful outcomes (Doyen & Dadario, 2022). Thus, addressing these challenges is crucial for the successful integration of AI in healthcare settings.

Issues in model evaluation. In the transition of AI into healthcare systems, a multitude of challenges affecting model evaluation and trust in machine learning applications have been identified. The reliability of ML predictions is compromised by various sources of uncertainty. The uncertainty inherent in data measurements (known as Aleatoric uncertainty) cannot be mitigated by increasing data collection (Kompa et al., 2021), while model uncertainty arises from the variability in model parameters and their interpretation of observed data (Tonekaboni et al., 2019). Additionally, experimental setups often produce results that vary with different random seeds, hyper-parameters, or data samples, emphasizing the need for robust statistical practices to distinguish noise from reproducible findings (Varoquaux & Cheplygina, 2022; Wiens et al., 2019). Furthermore, the context-specific accuracy of models can be skewed by unaccounted differences in patient treatment, exemplified by models erroneously associating asthma with a reduced risk of death from pneumonia due to treatment biases (Chan & Wong, 2022).

Technical debt in AI implementation. The transition of AI into healthcare systems often involves significant technical debt, making the process costly and hazardous. A large proportion of this debt stems from the extensive additional code required beyond the ML core model, which only represents about 5% of the total codebase (Harris et al., 2022). This auxiliary code, which includes elements such as “glue-code”, “pipeline jungles”, and “dead experimental codepaths”, creates complexity and potential errors during deployment (Harris et al., 2022). Furthermore, the design of EHR systems highlights the importance of integrating functionalities that enhance data consistency, accuracy, and error prevention through AI-driven auto-calculation processes. These features not only maintain consistency and reliability in clinical data management but also help mitigate technical debt by reducing reliance on manual processes and minimally impacting retraining operations (Iyer, 2023).

4.4. General AI best practices in healthcare context

To address general AI challenges in healthcare deployment, established machine learning best practices must be adapted to meet medical domain requirements. While these practices are common across AI applications, their implementation in healthcare demands heightened attention.

Implement techniques to evaluate model reliability. Ensuring the reliability and efficacy of machine learning models in healthcare requires a multifaceted approach. Predictive uncertainty, a crucial aspect, can be quantified through various methods tailored to the underlying model. For instance, linear regression often utilizes prediction intervals, while neural networks benefit from techniques like Monte Carlo Dropout and ensembling to estimate uncertainty (Kompa et al., 2021). Incorporating Gaussian processes further enhances predictive uncertainty estimates, offering built-in mechanisms for this purpose. Additionally, allowing models to abstain from making predictions, known as “selective prediction” (Kompa et al., 2021), communicates elevated uncertainty levels, prompting human intervention when necessary. Evaluation of segmentation performance, vital in tasks like medical imaging, can be conducted using metrics like the Dice coefficient or algorithms like STAPLE (Willeminck et al., 2020). Moreover, when deploying models in new clinical contexts, locally validating them through silent testing mitigates risks associated with generalization from academic or specialty settings (Antoniou & Mamdani, 2021). Clinical validation extends beyond predictive accuracy to assess the generation of clinically meaningful outputs in real-world environments. Robustness is also fundamental (Verma et al., 2021), ensuring models can differentiate between various conditions or treatments without relying excessively on specific input features. Furthermore, evaluating fairness and trustworthiness is imperative to avoid harming patients and build confidence in model outcomes. Fairness criteria, such as demographic parity and equalized odds, address biases in model predictions, while transparency and robustness demonstrate model reliability. Effective model validation requires considering sample sizes that offer sufficient statistical power and robust confidence intervals. Performance evaluation encompasses standard machine learning practices, while technical validation ensures the correctness and reliability of the software.

Ensure continuous model monitoring and retraining. Evaluation of ongoing system performance is essential for ensuring the continued reliability and effectiveness of machine learning solutions in healthcare. Regular retraining with updated data, monitoring for changes in data distribution, and consulting with clinical experts are vital steps in evaluating and maintaining system performance (Antoniou & Mamdani, 2021). Unlike conventional medical devices, AI tools have the unique capability of continuous improvement through incremental learning. This iterative process allows AI models to adapt and evolve with new data, ensuring they remain up-to-date and effective in healthcare applications (Finlayson et al., 2021; Ngiam & Khor, 2019). Dataset shift presents a significant challenge in deploying machine learning models in healthcare settings. Changes in data characteristics or clinical practices can, therefore, lead to the underperformance of models (Conor Corbin et al., 2023; Seneviratne et al., 2019).

Ensure reproducibility and replicability of ML pipeline. Recent studies have demonstrated that making model code publicly available can significantly enhance the reproducibility of machine learning experiments. This practice enables researchers to validate and replicate results, contributing to the transparency and trustworthiness of ML applications (*“recent studies showed [...] that public sharing of model code could enhance reproducibility”* (Bozkurt et al., 2020)). In healthcare, identifying and implementing best practices for designing ML studies is essential. Reproducibility and generalizability safeguards are crucial to ensure the equitable implementation of ML technologies across healthcare settings. These safeguards help to establish reliable and consistent results, fostering confidence in the outcomes of ML applications. Ensuring the reproducibility of an ML algorithm in the healthcare setting requires adherence to specific criteria: *“To ensure an ML algorithm applied in the healthcare setting is fully reproducible, some have suggested that a study should produce the same results according to (1) technically identical conditions related to code and dataset release, (2) statistically identical conditions”* (Doyen & Dadario, 2022).

5. Discussion

Our comprehensive review highlights several key findings on integrating artificial intelligence (AI) within healthcare systems. First, ‘responsible AI’ or ‘trustworthy MLOps’ is paramount, as emphasized by Khattak et al. (2023). This involves addressing ethical concerns such as fairness, interpretability, security, and resilience in AI applications. We found that data quality, bias, and security are critical challenges that must be addressed for successful AI deployment in healthcare. A critical aspect that emerges from the literature is the trade-off between model performance and practical deployability in clinical environments. Highly expressive models, such as deep neural networks, often achieve superior predictive accuracy but at the cost of reduced interpretability, which may hinder their adoption in high-stakes domains like healthcare. Furthermore, optimization for performance metrics can lead to overfitting on non-representative training data, thereby exacerbating issues related to fairness, generalizability, and robustness. These considerations highlight the necessity for evaluation frameworks that balance accuracy with explainability, bias mitigation, and system resilience. Second, biases in predictive models, including omitted variables, representation, and measurement biases, are significant issues, as noted by Chan and Wong (2023). Our review identifies practical strategies to mitigate these biases, such as data augmentation and fairness-enhancing algorithms, to ensure equitable healthcare outcomes. Third, our multivocal literature review expands upon previous reviews by incorporating both peer-reviewed and grey literature, offering a more comprehensive view of AI integration in healthcare. While earlier reviews focused primarily on data quality issues, bias, and interoperability challenges, this review highlights new emerging concerns, such as the distributional shift between training and production data, the need for continuous model retraining, and the potential of federated learning. It also places greater emphasis on best practices, including leveraging multiple data sources, implementing robust security measures, and ensuring compliance with healthcare regulations. The inclusion of actionable insights and operational challenges, such as technical debt and deployment strategies, differentiates it from previous reviews, which are more focused on identifying obstacles rather than proposing solutions.

In the remainder of this section, we discuss these results in detail, explore their implications, and consider our study’s limitations.

5.1. Best practices & Challenges of integrating AI in healthcare

Despite significant progress, several best practices have been identified, along with persistent challenges in the deployment of AI in healthcare. In Table 5, we report a summary of the best practices and challenges identified, along with their support, that is, how frequently they are mentioned in the identified sources. To answer the two research questions, based on the literature, we have categorized the challenges and corresponding best practices into three main clusters: (i) compliance & data management, (ii) model performance & fairness, and (iii) security & deployment, as shown in Fig. 2.

5.1.1. Compliance & Data management

Standards, compliance, and data management are the foundation for reliable AI systems in healthcare. Best practices such as relying on standards, complying with healthcare regulations, and documenting datasets and models ensure consistency, legality, and transparency. These practices provide a foundation that enhances the robustness and accuracy of AI models, particularly when implementing techniques to ensure data quality and mitigate data scarcity by leveraging multiple sites. However, challenges such as the lack of universal standards, insufficient healthcare data, and mismatches between the distribution of production and training data pose significant hurdles. By implementing high-quality data practices and thorough documentation, many of these challenges, such as inconsistent data or limited utility, can be mitigated. For example, standardized data collection methods and detailed documentation of datasets help improve consistency across different sources while ensuring data remains useful and adaptable in AI applications. However, there remains an ongoing need for continuous data validation and adjustment, emphasizing the dynamic effort required to maintain data integrity and ensure AI models are applicable in real-world healthcare scenarios.

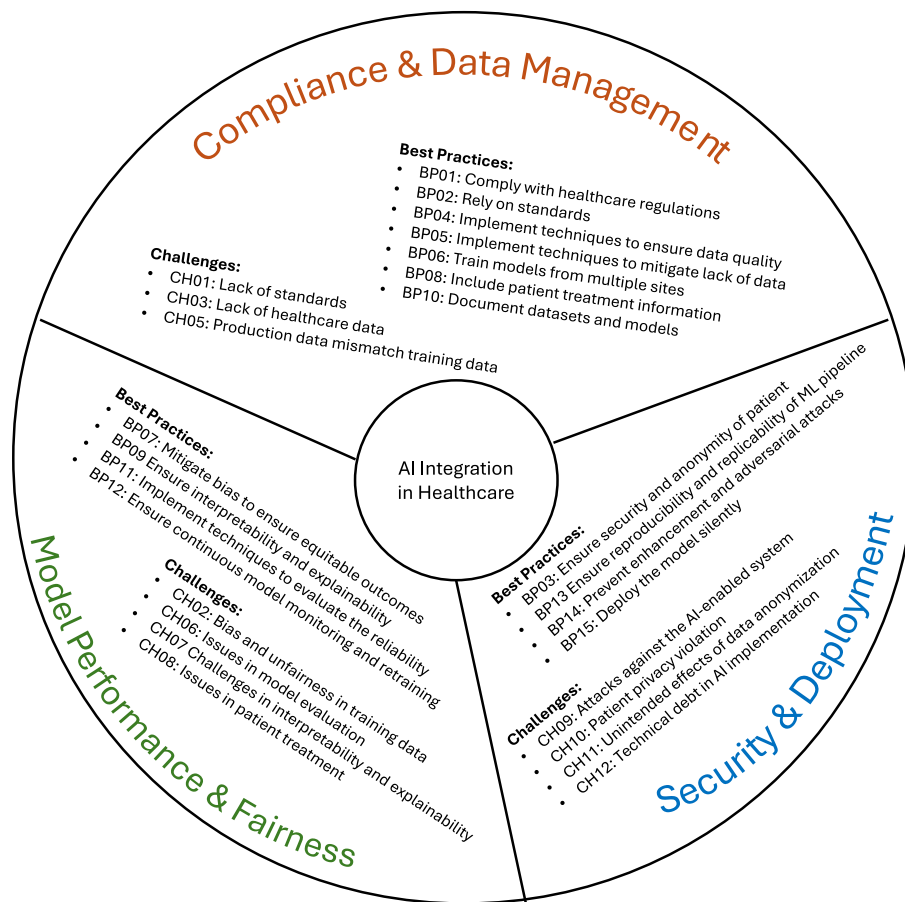


Fig. 2. Clusters of best practices and challenges in healthcare AI.

5.1.2. Model performance & Fairness

Ensuring model performance and fairness is crucial for the ethical deployment of AI in healthcare. Best practices in this cluster focus on mitigating bias, ensuring continuous model monitoring and retraining, evaluating model reliability, and maintaining interpretability and explainability. These practices aim to prevent discriminatory outcomes, keep models relevant, and build trust among healthcare professionals. However, challenges such as inherent biases in training data, problematic AI-driven patient treatment, model evaluation issues, and the difficulty of interpreting complex models persist. Addressing bias and ensuring interpretability are both essential for creating fair and trusted AI systems in healthcare. One of the key challenges is balancing model complexity with transparency; more complex, highly accurate models can be more challenging to interpret, which can reduce trust. This trade-off highlights the need to find a balance between achieving high performance and ensuring the model remains explainable and reliable for healthcare professionals and patients alike. By focusing on mitigating bias, continuously monitoring and retraining models, and maintaining interpretability, these practices help prevent discriminatory outcomes and build trust in AI-driven healthcare.

5.1.3. Security & Deployment

Security, privacy, and effective deployment are paramount for AI systems in healthcare. Best practices include ensuring the security and anonymity of patient data, cautious model deployment, preventing enhancement and adversarial attacks, and ensuring the reproducibility and replicability of the machine learning pipeline. These measures are critical for protecting patient information, maintaining system integrity, and building confidence in AI systems. However, challenges such as potential privacy violations, security threats, federated learning limitations, and the unintended effects of data anonymization create significant obstacles. Securing patient data and preventing attacks are not only essential best practices but also significant challenges, underscoring the need for robust and adaptable security measures. One key challenge is that data anonymization, while critical for protecting privacy, can reduce the utility of the data, requiring a careful balance between maintaining privacy and preserving data effectiveness for AI applications.

5.1.4. Economic feasibility and cost considerations

Another key consideration that emerged from the reviewed literature is the economic feasibility of deploying AI in healthcare settings. Although technical performance and fairness are often prioritized, several studies explicitly highlight cost as a critical factor in real-world implementation. For instance, Sendak et al. (2020) report a deployment cost of over \$220,000 for integrating a predictive model within a hospital system, noting that similar investments would be required for each additional deployment site unless shared infrastructures are adopted. Seneviratne et al. (2019) advocate for conducting cost-utility analyses early in the development process and continuously throughout deployment to assess the clinical and financial trade-offs of AI-based interventions compared to existing practices. Harris et al. (2022) underscore the burden of hidden deployment costs, observing that up to 95% of the total effort — encompassing coding, infrastructure engineering, MLOps, and monitoring — occurs outside of the model itself. Similarly, Wiens et al. (2019) call for the inclusion of economic metrics, such as cost-effectiveness and number needed to benefit, when evaluating the real-world impact of clinical machine learning models. These findings collectively underscore the importance of incorporating financial sustainability considerations into AI deployment strategies, as they are essential for ensuring scalability, institutional adoption, and equitable access across healthcare systems.

5.2. Implications for future research

The integration of AI in healthcare presents several critical avenues for future research. In particular, developing advanced techniques for data interoperability and standardization across healthcare systems is essential, with the aim of creating universally accepted protocols for seamless information sharing. Another crucial direction is enhancing the interpretability and transparency of AI models. This should be done by focusing on methodologies that provide clear, actionable insights for healthcare professionals, such as refined explainable AI (XAI) techniques. Furthermore, addressing security and privacy concerns requires innovative approaches to protect sensitive healthcare data while maintaining regulatory compliance. Future research should also prioritize developing fairness-enhancing algorithms to mitigate biases in AI models, ensuring that AI contributes to reducing rather than exacerbating healthcare disparities. Lastly, exploring the application of federated learning in healthcare offers promising solutions to data privacy concerns while enabling robust AI model development from decentralized data sources.

Research in these areas will be instrumental in overcoming current challenges and advancing the successful integration of AI in healthcare systems, ultimately improving patient outcomes and operational efficiency.

5.3. Limitations of the research

Although our multivocal review of the literature provides comprehensive insights into the challenges and best practices of integrating AI into healthcare systems, several limitations must be acknowledged, along with our efforts to mitigate them.

The rapid pace of innovation in AI and healthcare presents a challenge in covering all the latest advances. To counteract this, we employed a broad search strategy and included recent sources from the gray literature to capture emerging trends that may not yet appear in academic publications. The inclusion of grey literature, while beneficial for capturing a wide range of perspectives, introduces potential subjectivity in source selection and quality assessment. We addressed this by establishing clear inclusion criteria and using a systematic approach to evaluate grey literature sources, ensuring a balance between comprehensiveness and credibility.

The methodological heterogeneity among the reviewed studies posed challenges in drawing uniform conclusions. We addressed this limitation by employing a thematic analysis approach, which allowed us to identify common themes across diverse study designs and methodologies. In addition, to minimize subjective interpretation in our thematic analysis, we employed a team-based approach with multiple researchers independently coding and categorizing the data, followed by consensus discussions to resolve discrepancies. By implementing these measures, we aimed to enhance the robustness and comprehensiveness of our MLR while transparently acknowledging its inherent limitations.

Lastly, our discussion acknowledged the lack of empirical validation for some identified best practices and challenges. We emphasize the need for future research to validate these findings in real-world healthcare settings.

6. Conclusion

Integrating AI into healthcare systems presents significant implementation challenges that, when properly addressed through systematic approaches and best practices, can lead to improved healthcare delivery.

Our review identifies critical barriers and corresponding solutions for successful AI integration in healthcare settings. These include ensuring data quality and representativeness, mitigating bias, maintaining interoperability, safeguarding patient privacy, enhancing AI robustness against adversarial attacks, and ensuring model interpretability and transparency. Addressing these multifaceted issues is crucial to developing responsible and trustworthy AI that is fair, transparent, secure, and reliable in healthcare settings.

Our analysis extends the existing literature by exploring emerging challenges and best practices in AI integration. We identify the importance of balancing innovation with caution, particularly in areas such as data discrepancies, the utility of federated learning for the preservation of privacy, and the need for continuous AI monitoring and retraining. The review emphasizes best practices that include rigorous model evaluation, standardization of data practices, and fostering close collaboration among AI developers, healthcare professionals, policymakers, and patients. We also highlight the potential of AI to advance personalized medicine and

optimize healthcare workflows while noting the need for novel approaches to regulation and validation as AI systems become more autonomous and adaptive.

To successfully integrate AI in healthcare, a multidisciplinary approach is essential. Although the challenges are substantial, they are matched by significant opportunities for innovation. By adhering to the principles of trustworthy AI and addressing the detailed challenges and best practices outlined in our review, we can significantly guide AI to enhance healthcare outcomes and operational efficiency. Moving forward, the evidence suggests that systematic evaluation of AI systems through established frameworks and metrics is needed to ensure their reliable and ethical deployment in healthcare settings. This includes improving dataset benchmarking practices to better assess fairness, robustness, and generalizability, particularly across diverse clinical populations. It also involves developing methodologies to evaluate the long-term effects on patient outcomes and healthcare economics, improving AI literacy among healthcare professionals and the public, and ensuring that AI systems enhance rather than replace the human elements of healthcare.

As we navigate this rapidly evolving landscape, maintaining a balance between embracing AI's transformative potential and upholding fundamental medical principles will be key to realizing a future of healthcare that is more precise, personalized, and accessible to all.

CRediT authorship contribution statement

Giulio Mallardi: Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Fabio Calefato:** Writing – review & editing, Investigation, Formal analysis. **Luigi Quaranta:** Writing – review & editing, Investigation. **Filippo Lanubile:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Andrea Basile for assistance with data collection and preliminary analysis. This study has been co-funded by the Complementary National Plan PNC-I.1 - Research initiatives for innovative technologies and pathways in the health and welfare sector - D.D. 931 of 06/06/2022 (“DARE - Digital lifelong pRevEntion”, code PNC0000002, CUP B53C22006420001), the NRRP Initiative, Mission 4, Component 2, Investment 1.3 - Partnerships extended to universities, research centres, companies and research D.D. MUR n. 341, 15.03.2022 – Next Generation EU (“FAIR - Future Artificial Intelligence Research”, code PE00000013, CUP H97G22000210007), and PNRR “SERICS - SEcurity and RIghts in the CybeRSpace” (PE0000014, CUP: H93C22000620001).

Appendix. List of selected papers

Ref.	Year	Authors	Title	Type
Finlayson et al. (2021)	2021	Finlayson et al.	The Clinician and Dataset Shift in Artificial Intelligence	Article
Bozkurt et al. (2020)	2020	Bozkurt et al.	Reporting of demographic data and representativeness in machine learning models using electronic health records	Article
Harris et al. (2022)	2022	Harris et al.	Clinical deployment environments: Five pillars of translational machine learning for health	Article
Wiens et al. (2019)	2019	Wiens et al.	Do no harm: a roadmap for responsible machine learning for health care	Article
McCoy et al. (2020)	2020	McCoy et al.	Ensuring machine learning for healthcare works for all	Article
Jones et al. (2023)	2023	Jones et al.	No Fair Lunch: A Causal Perspective on Dataset Bias in Machine Learning for Medical Imaging	Article

Willemink et al. (2020)	2020	Willemink et al.	Preparing Medical Imaging Data for Machine Learning	Article
Mincu and Roy (2022)	2022	Mincu and Roy	Developing robust benchmarks for driving forward AI innovation in healthcare	Article
Doyen and Dadario (2022)	2022	Doyen and Dadario	12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context	Article
Varoquaux and Cheplygina (2022)	2022	Varoquaux and Cheplygina	Machine learning for medical imaging: methodological failures and recommendations for the future	Article
Rajkomar et al. (2019)	2019	Rajkomar et al.	Machine Learning in Medicine	Article
Ngiam and Khor (2019)	2019	Ngiam and Khor	Big data and machine learning algorithms for health-care delivery	Article
Chan and Wong (2022)	2022	Chan and Wong	Accounting for treatment during the development or validation of prediction models	Article
Kompa et al. (2021)	2021	Kompa et al.	Second opinion needed: communicating uncertainty in medical machine learning	Article
Nestor et al. (2019)	2019	Nestor et al.	Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks	Article
Stefanie Warnat-Herresthal et al. (2021)	2021	Stefanie Warnat-Herresthal et al.	Swarm Learning for decentralized and confidential clinical machine learning	Article
Kaissis et al. (2020)	2020	Kaissis et al.	Secure, privacy-preserving and federated machine learning in medical imaging	Article
Yoon et al. (2021)	2021	Yoon et al.	Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned?	Article
Jia et al. (2021)	2021	Jia et al.	The Role of Explainability in Assuring Safety of Machine Learning in Healthcare	Article
Kolyshkina and Simoff (2021)	2021	Kolyshkina and Simoff	Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach	Article
Vellido (2019)	2019	Vellido	The importance of interpretability and visualization in machine learning for applications in medicine and health care	Article
Finlayson et al. (2019)	2019	Finlayson et al.	Adversarial attacks on medical machine learning	Article
Salem et al. (2018)	2018	Salem et al.	Utilizing Transfer Learning and Homomorphic Encryption in a Privacy Preserving and Secure Biometric Recognition System	Article
Granlund et al. (2021)	2021	Granlund et al.	Towards Regulatory-Compliant MLOps: Oravizio's Journey from a Machine Learning Experiment to a Deployed Certified Medical Product	Article
Huang and Liu (2019)	2019	Huang and Liu	Patient Clustering Improves Efficiency of Federated Machine Learning to predict mortality and hospital stay time using distributed Electronic Medical Records	Article
Antoniou and Mamdani (2021)	2021	Antoniou and Mamdani	Evaluation of machine learning solutions in medicine	Article
Oner et al. (2020)	2020	Oner et al.	Training machine learning models on patient level data segregation is crucial in practical clinical applications	Article
Štiglic et al. (2020)	2020	Štiglic et al.	Interpretability of machine learning-based prediction models in healthcare	Article
Verma et al. (2021)	2021	Verma et al.	Implementing machine learning in medicine	Article
Conor Corbin et al. (2023)	2023	Conor Corbin et al.	DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record	Article
Seneviratne et al. (2019)	2019	Seneviratne et al.	Bridging the implementation gap of machine learning in healthcare	Article
Rosenblatt et al. (2023)	2023	Rosenblatt et al.	Gradient-based enhancement attacks in biomedical machine learning	Article

Waring et al. (2020a)	2020a	Waring et al.	Automated machine learning: Review of the state-of-the-art and opportunities for healthcare	Article
Ulmer et al. (2020)	2020	Ulmer et al.	Trust Issues: Uncertainty Estimation Does Not Enable Reliable OOD Detection On Medical Tabular Data	Article
Pfohl et al. (2020)	2020	Pfohl et al.	An Empirical Characterization of Fair Machine Learning For Clinical Risk Prediction	Article
Marulli et al. (2022)	2022	Marulli et al.	Sensitivity of Machine Learning Approaches to Fake and Untrusted Data in Healthcare Domain	Article
Mccradden et al. (2023)	2023	Mccradden et al.	What's Fair Is... Fair? Presenting JustEFAB, an Ethical Framework for Operationalizing Medical Ethics and Social Justice in the Integration of Clinical Machine Learning: JustEFAB	Inproceedings
Pandl et al. (2021)	2021	Pandl et al.	Trustworthy machine learning for health care: scalable data valuation with the shapley value	Inproceedings
Sendak et al. (2022)	2022	Sendak et al.	Development and Validation of ML-DQA - a Machine Learning Data Quality Assurance Framework for Healthcare	Inproceedings
Mena and Gonzalez (2006)	2006	Mena and Gonzalez	Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic	Inproceedings
Dubovitskaya et al. (2017)	2017	Dubovitskaya et al.	Secure and Trustable Electronic Medical Records Sharing using Blockchain	Inproceedings
Tonekaboni et al. (2022)	2022	Tonekaboni et al.	How to validate Machine Learning Models Prior to Deployment: Silent trial protocol for evaluation of real-time models at ICU	Inproceedings
Newaz et al. (2020)	2020	Newaz et al.	Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems	Inproceedings
Iyer (2023)	2023	Iyer	Architecting MLOps Solutions for Healthcare	Misc
Tonekaboni et al. (2019)	2019	Tonekaboni et al.	What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use	Misc
Reinsch (2024)	N/A	Reinsch	Regulatory Requirements for Medical Devices with Machine Learning	Misc
Iyer (2021)	N/A	Iyer	Various Challenges for Applying Machine Learning in Healthcare	Misc
Eng et al. (2023)	N/A	Eng et al.	Saving Mothers with ML How CareSource uses MLOps to improve Healthcare in High-Risk Obstetrics	Misc
Lavanya (2021)	N/A	Lavanya	Will MLOps change the future of the healthcare system?	Misc

Data availability

No data was used for the research described in the article.

References

- An, Qi, Rahman, Saifur, Zhou, Jingwen, & Kang, James Jin (2023). A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors*, [ISSN: 1424-8220] 23(9), <http://dx.doi.org/10.3390/s23094178>, URL <https://www.mdpi.com/1424-8220/23/9/4178>.
- Antoniou, T., & Mamdani, M. (2021). Evaluation of machine learning solutions in medicine. *CMAJ : Canadian Medical Association Journal*, 193, E1425–E1429.
- Arman Ghavidel, Rouzbeh Ghousi, & Atashi, Alireza (2023). An ensemble data mining approach to discover medical patterns and provide a system to predict the mortality in the ICU of cardiac surgery based on stacking machine learning method. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4), 1316–1326. <http://dx.doi.org/10.1080/21681163.2022.2063189>.
- Bozkurt, S., Cahan, Eli M., Seneviratne, Martin G., Sun, Ran, Lossio-Ventura, Juan Antonio, Ioannidis, J., et al. (2020). Reporting of demographic data and representativeness in machine learning models using electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 27(12), 1878–1884. <http://dx.doi.org/10.1093/jamia/ocaa164>.
- Braun, Virginia, & Clarke, Victoria (2012). Thematic analysis. In *Handbook of research methods in psychology* (pp. 57–71). American Psychological Association, ISBN: 978-1-4338-1003-9.
- Chan, Wei Xin, & Wong, Limsoon (2022). Accounting for treatment during the development or validation of prediction models. *Journal of Bioinformatics and Computational Biology*, Article 2271001.
- Chan, Wei Xin, & Wong, Limsoon (2023). Obstacles to effective model deployment in healthcare. *Journal of Bioinformatics and Computational Biology*, 21(2), Article 2371001. <http://dx.doi.org/10.1142/S0219720023710014>.
- Corbin, Conor K., Maclay, R., Acharya, Aakash, Mony, Sreedevi, Punnathanam, Soumya, Thapa, Rahul, et al. (2023). DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. *Journal of the American Medical Informatics Association : JAMIA*, 30, 1532–1542. <http://dx.doi.org/10.1093/jamia/ocad114>.

- Doyen, S., & Dadario, N. (2022). 12 plagues of AI in healthcare: A practical guide to current issues with using machine learning in a medical context. *Frontiers in Digital Health*, 4, <http://dx.doi.org/10.3389/fdgh.2022.765406>.
- Dubovitskaya, Alevtina, Xu, Zhigang, Ryu, Samuel, Schumacher, M., & Wang, Fusheng (2017). Secure and trustable electronic medical records sharing using blockchain. vol. 2017, In *American medical informatics association annual symposium* (pp. 650–659).
- Eng, Chengyin, Scoville, Russ, Gupta, Arpit, & Aleman, Alvaro (2023). Saving mothers with ML how CareSource uses mlops to improve healthcare in high-risk obstetrics. <https://www.databricks.com/blog/2023/04/03/saving-mothers-ml-how-mlops-improves-healthcare-high-risk-obstetrics.html>.
- Finlayson, S. G., Bowers, John, Ito, Joichi, Zittrain, Jonathan, Beam, Andrew, & Kohane, I. (2019). Adversarial attacks on medical machine learning. *Science*, 363, 1287–1289. <http://dx.doi.org/10.1126/science.aaw4399>.
- Finlayson, S. G., Subbaswamy, Adarsh, Singh, Karandeep, Bowers, John, Kupke, Annabel, Zittrain, Jonathan, et al. (2021). The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, 385(3), 283–286. <http://dx.doi.org/10.1056/NEJMc2104626>.
- Garousi, Vahid, Felderer, Michael, & Mäntylä, Mika V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, [ISSN: 0950-5849] 106, 101–121. <http://dx.doi.org/10.1016/j.infsof.2018.09.006>.
- Granlund, Tuomas, Stirbu, Vlad, & Mikkonen, Tommi (2021). Towards regulatory-compliant MLOps: Oravizio's journey from a machine learning experiment to a deployed certified medical product. *SN Computer Science*, 2(5), 342. <http://dx.doi.org/10.1007/s42979-021-00726-1>, ISSN 2662-995X, 2661-8907.
- Harris, S., Bonnici, Tim, Keen, Thomas, Lilaonitkul, Watjana, White, Mark J., & Swanepoel, Nel (2022). Clinical deployment environments: Five pillars of translational machine learning for health. *Frontiers in Digital Health*, 4, <http://dx.doi.org/10.3389/fdgh.2022.939292>.
- Huang, Li, & Liu, Dianbo (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of Biomedical Informatics*, Article 103291.
- Iyer, Anjali (2021). Various challenges for applying machine learning in healthcare. <https://towardsdatascience.com/various-ways-at-which-machine-learning-could-be-used-in-medical-diagnosis-338c65393509>.
- Iyer, Anjali (2023). Architecting mlops solutions for healthcare. <https://mathco.com/article/architecting-mlops-solutions-for-healthcare/>.
- Jackson, Stuart, Yaqub, Maha, & Li, Cheng-Xi (2019). The agile deployment of machine learning models in healthcare. *Frontiers in Big Data*, [ISSN: 2624-909X] 1, <http://dx.doi.org/10.3389/fdata.2018.00007>.
- Jia, Yan, Mcdermid, J., Lawton, T., & Habli, I. (2021). The role of explainability in assuring safety of machine learning in healthcare. *IEEE Transactions on Emerging Topics in Computing*, 10, 1746–1760. <http://dx.doi.org/10.1109/TETC.2022.3171314>.
- Jones, Charles, de Castro, Daniel Coelho, Ribeiro, Fabio De Sousa, Oktay, O., McCradden, M., & Glocker, B. (2023). No fair lunch: A causal perspective on dataset bias in machine learning for medical imaging. ArXiv, [abs/2307.16526](https://arxiv.org/abs/2307.16526).
- Kaissis, Georgios, Makowski, M., Rückert, D., & Braren, R. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2, 305–311. <http://dx.doi.org/10.1038/s42256-020-0186-1>.
- Khattak, Faiza Khan, Subasri, Vallijah, Krishnan, Amrit, Dolatabadi, Elham, Pandya, Deval, Seyyed-Kalantari, Laleh, et al. (2023). MLHops: Machine learning for healthcare operations. arXiv:2305.02474 URL <https://arxiv.org/abs/2305.02474>.
- Kitchenham, Barbara Ann, & Charters, Stuart (2007). Guidelines for performing systematic literature reviews in software engineering. URL https://www.elsevier.com/_data/promis_misc/525444systematicreviewguide.pdf.
- Kolyshkina, I., & Simoff, S. (2021). Interpretability of machine learning solutions in public healthcare: The CRISP-ML approach. *Frontiers in Big Data*, 4, <http://dx.doi.org/10.3389/fdata.2021.660206>.
- Kompa, Benjamin, Snoek, Jasper, & Beam, A. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4, <http://dx.doi.org/10.1038/s41746-020-00367-3>.
- Lavanya, S. (2021). Will mlops change the future of the healthcare system?. <https://www.analyticsvidhya.com/blog/2021/06/will-mlops-change-the-future-of-the-healthcare-system-major-use-cases-of-mlops-in-health-care/>.
- Marulli, F., Marrone, S., & Verde, Laura (2022). Sensitivity of machine learning approaches to fake and untrusted data in healthcare domain. *Journal of Sensor and Actuator Networks*, 11, 21. <http://dx.doi.org/10.3390/jsan11020021>.
- McCoy, L., Banja, J., Ghassemi, M., & Celi, L. (2020). Ensuring machine learning for healthcare works for all. *BMJ Health & Care Informatics*, 27, Article e100237.
- McCradden, Melissa, Oduki, Oluwadara, Joshi, Shalmali, Akrouf, Ismail, Ndlovu, Kagiso, Glocker, Ben, et al. (2023). What's fair is... Fair? Presenting justefab, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: Justefab. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1505–1519). New York, NY, USA: Association for Computing Machinery, ISBN: 9798400701924, <http://dx.doi.org/10.1145/3593013.3594096>.
- Mena, L., & Gonzalez, Jesus A. (2006). Machine learning for imbalanced datasets: Application in medical diagnostic. In *The florida AI research society* (pp. 574–579).
- Mincu, Diana, & Roy, Subhrajit (2022). Developing robust benchmarks for driving forward AI innovation in healthcare. *Nature Machine Intelligence*, 4, 916–921. <http://dx.doi.org/10.1038/s42256-022-00559-4>.
- Nestor, Bret A., McDermott, Matthew B. A., Boag, Willie, Berner, G., Naumann, Tristan, Hughes, Michael C., et al. (2019). Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. ArXiv, [abs/1908.00690](https://arxiv.org/abs/1908.00690).
- Newaz, Akm Iqtidar, Haque, Nur Intiazul, Sikder, A., Rahman, M., & Uluagac, A. (2020). Adversarial attacks to machine learning-based smart healthcare systems. In *Global communications conference* (pp. 1–6). <http://dx.doi.org/10.1109/GLOBECOM42002.2020.9322472>.
- Ngiam, K., & Khor, Ing Wei (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet. Oncology*, 20(5), e262–e273.
- Oner, Mustafa Umit, Cheng, Yu-Chia, Lee, H. K., & Sung, Wing-Kin (2020). Training machine learning models on patient level data segregation is crucial in practical clinical applications. MedRxiv.
- Page, Matthew J, McKenzie, Joanne E, Bossuyt, Patrick M, Boutron, Isabelle, Hoffmann, Tammy C, Mulrow, Cynthia D, et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal*, 372, <http://dx.doi.org/10.1136/bmj.n71>.
- Paleyev, Andrei, Urma, Raoul-Gabriel, & Lawrence, Neil D. (2022). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, [ISSN: 0360-0300] 55(6), <http://dx.doi.org/10.1145/3533378>.
- Pandl, Konstantin D., Feiland, Fabian, Thiebes, Scott, & Sunyaev, A. (2021). Trustworthy machine learning for health care: scalable data valuation with the shapley value. In *Proceedings of the conference on health, inference, and learning*. <http://dx.doi.org/10.1145/3450439.3451861>.
- Petersen, Kai, Vakkalanka, Sairam, & Kuzniar, Ludwik (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, [ISSN: 0950-5849] 64, 1–18. <http://dx.doi.org/10.1016/j.infsof.2015.03.007>.
- Petticrew, Mark, & Roberts, Helen (2006). *Systematic reviews in the social sciences: A practical guide*. Malden: Blackwell Publishing, <http://dx.doi.org/10.1002/9780470754887>, ISBN 1-4051-2110-6 (Hardcover); 978-1-4051-2110-1 (Hardcover).
- Pfohl, S., Foryciarz, Agata, & Shah, N. (2020). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, Article 103621. <http://dx.doi.org/10.1016/j.jbi.2020.103621>.
- Qayyum, Adnan, Qadir, Junaid, Bilal, Muhammad, & Al-Fuqaha, Ala (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156–180. <http://dx.doi.org/10.1109/RBME.2020.3013489>.
- Rajkomar, A., Dean, Jeffrey, & Kohane, I. (2019). Machine learning in medicine. *The New England Journal of Medicine*, 380, 1347–1358. <http://dx.doi.org/10.1056/NEJMr1814259>.

- Reinsch, Daniel (2024). Regulatory requirements for medical devices with machine learning. <https://www.johner-institute.com/articles/regulatory-affairs/and-more/regulatory-requirements-for-medical-devices-with-machine-learning/>.
- Rosenblatt, M., Dadashkarimi, J., & Scheinost, D. (2023). Gradient-based enhancement attacks in biomedical machine learning. arXiv.
- Salem, M., Taheri, S., & Yuan, Jiann-Shiun (2018). Utilizing transfer learning and homomorphic encryption in a privacy preserving and secure biometric recognition system. *Computer*, 8, 3.
- Sendak, M., D'Arcy, Joshua, Kashyap, Sehj, Gao, M., Nichols, M., Corey, Kristin M., et al. (2020). A path for translation of machine learning products into healthcare delivery. *EMJ Innovations*, <http://dx.doi.org/10.33590/emjinnov/19-00172>.
- Sendak, M., Sirdeshmukh, Gaurav, Ochoa, Timothy N., Premo, H., Tang, Linda, Niederhoffer, Kira L., et al. (2022). Development and validation of ML-DQA - a machine learning data quality assurance framework for healthcare. In *Machine learning in health care* (pp. 741–759). <http://dx.doi.org/10.48550/arXiv.2208.02670>.
- Seneviratne, Martin G., Shah, N., & Chu, Larry (2019). Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, 6, 45–47. <http://dx.doi.org/10.1136/bmjinnov-2019-000359>.
- Štiglic, G., Kocbek, Primož, Fijačko, Nino, Zitnik, M., Verbert, K., & Cilar, Leona (2020). Interpretability of machine learning-based prediction models in healthcare. vol. 10, In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.
- Tonekaboni, Sana, Joshi, Shalmali, McCradden, Melissa D., & Goldenberg, Anna (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use.
- Tonekaboni, S., Morgenshtern, G., Assadi, A., Pokhrel, Aslesha, Huang, Xi, Jayarajan, Anand, et al. (2022). How to validate machine learning models prior to deployment: Silent trial protocol for evaluation of real-time models at ICU. In *ACM conference on health, inference, and learning* (pp. 169–182).
- Ulmer, Dennis, Meijerink, L., & Cinà, G. (2020). Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. ArXiv, [abs/2011.03274](https://arxiv.org/abs/2011.03274).
- Varoquaux, Gaël, & Cheplygina, Veronika (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digital Medicine*, [ISSN: 2398-6352] 5, 48. <http://dx.doi.org/10.1038/s41746-022-00592-y>.
- Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32, 18069–18083.
- Verma, A., Murray, Joshua, Greiner, R., Cohen, Joseph Paul, Shojania, K., Ghassemi, M., et al. (2021). Implementing machine learning in medicine. *CMAJ : Canadian Medical Association Journal*, 193, E1351–E1357. <http://dx.doi.org/10.1503/cmaj.202434>.
- Waring, Jonathan, Lindvall, C., & Umeton, R. (2020a). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, 104, Article 101822. <http://dx.doi.org/10.1016/j.artmed.2020.101822>.
- Waring, Jonathan, Lindvall, Charlotta, & Umeton, Renato (2020b). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, [ISSN: 0933-3657] 104, Article 101822. <http://dx.doi.org/10.1016/j.artmed.2020.101822>.
- Warnat-Herresthal, Stefanie, Schultze, Hartmut, Shastri, Krishnaprasad, Manamohan, Sathyanarayanan, Mukherjee, Saikat, Garg, Vishesh, et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594, 265–270. <http://dx.doi.org/10.1038/s41586-021-03583-3>.
- Wiens, Jenna, Saria, Suchi, Sendak, Mark, Ghassemi, Marzyeh, Liu, Vincent X., Doshi-Velez, Finale, et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340. <http://dx.doi.org/10.1038/s41591-019-0548-6>.
- Willemink, M., Koszek, Wojciech A., Hardell, Cailin, Wu, Jie, Fleischmann, D., Harvey, H., et al. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4–15.
- Wohlin, Claes (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering* (pp. 38:1–38:10). New York, NY, USA: Association for Computing Machinery, ISBN: 9781450324762, <http://dx.doi.org/10.1145/2601248.2601268>.
- Yoon, C., Torrance, R., & Scheinerman, Naomi (2021). Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, 48, 581–585. <http://dx.doi.org/10.1136/medethics-2020-107102>.
- Zhang, Angela, Xing, Lei, Zou, James, & Wu, Joseph C. (2022). Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12), 1330–1345. <http://dx.doi.org/10.1038/s41551-022-00898-y>.