Review

Maturity Framework for Operationalizing Machine Learning Applications in Health Care: Scoping Review

Yutong Li[1], BSc; Julie Tian[1], BSc; Ariana Xu[1]; Russell Greiner[2], PhD; Jake Hayward[3], MD; Andrew James Greenshaw[1], PhD; Bo Cao[1], PhD

[1]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada
[2]Department of Computer Science, University of Alberta, Edmonton, AB, Canada
[3]Department of Emergency Medicine, University of Alberta, Edmonton, AB, Canada

Corresponding Author:

Bo Cao, PhD
Department of Psychiatry
University of Alberta
4-142 KATZ
Edmonton, AB T6G 2R3
Canada
Phone: 1 780-407-6504
Email: cloudbocao@gmail.com

Abstract

Background: The exponential growth of publications regarding the application of machine learning (ML) tools in medicine highlights the significant potential for ML to revolutionize the field. Despite the multitude of literature surrounding this topic, there are limited publications addressing the implementation and feasibility of ML models in clinical practice. Currently, Machine Learning Operations (MLOps), a set of practices designed to deploy and maintain ML models in production, is used in various information technology and industrial settings. However, the MLOps pipeline is not well researched in medical settings, where multiple barriers exist to implementing ML pipelines into practice.

Objective: This study aims to detail how MLOps is implemented in health care and propose a maturity framework for the health care implementations.

Methods: A scoping review search was conducted according to the Joanna Briggs Institute Manual for Evidence Synthesis. Results were synthesized using the 3-stage basic qualitative content analysis. We searched 4 databases (eg, MEDLINE, Embase, Web of Science, and Scopus) to include any studies that involved proof of concept or real-world implementation of MLOps in health care. Studies not reported in English were excluded.

Results: A total of 19 studies were included in this scoping review. The MLOps workflow identified within the studies included (1) data extraction (19/19 studies), (2) data preparation and engineering (18/19 studies), (3) model training (19/19 studies), (4) measured ML metrics and model evaluation (17/19 studies), (5) model validation and test in production (14/19 studies), (6) model serving and deployment (15/19 studies), (7) continuous monitoring (14/19 studies), and (8) continual learning (13/19 studies). We proposed a 3-stage MLOps maturity framework for health care based on existing studies in the field, that is, low (5/19 studies), partial (1/19 studies), and full maturity (13/19 studies). There were 8/19 studies that discussed ethical, legislative, and stakeholder considerations for MLOps implementations in health care settings.

Conclusions: We investigated the implementation of MLOps in health care with a corresponding maturity framework. It is evident that only a limited number of studies reported the implementation of ML in health care contexts. Hence, it is imperative that we shift our focus toward creating an environment that supports the development of ML health care applications, such as improving existing data infrastructure, and engaging partners to support the development of MLOps applications. Specifically, we can include patients, policymakers, and health care professionals in the creation and implementation of ML applications. One of the main limitations includes the varying quality of each extracted study in terms of how the MLOps implementation was presented. Hence, it was difficult to verify the presence and discuss in depth all steps of the MLOps workflow for each study. Furthermore, due to the inherent nature of a scoping review protocol, there may be a compromise on an in-depth discussion of each step within the MLOps workflow.

J Med Internet Res 2025;27:e66559; doi: 10.2196/66559

Keywords: maturity framework; health care applications; scoping review; machine learning; ML; machine learning operations; MLOps; data; database; clinical practice; health care; health care implementations

Introduction

Over the last decade, there has been exponential growth in the number of machine learning (ML) publications in the medical field, where ML is defined as a group of algorithms and statistical methods that enable computers to learn from data and create predictions [1-3]. Despite the significant number of such studies, fewer than 10% have been implemented in clinical settings [4]. ML model implementations in the clinical setting have significant potential to improve patient care, as ML models have been shown to be effective in medical use cases, such as classifying images, supporting medical diagnoses, and triaging patients [5-10]. To best operationalize the ML models, a general standardized workflow can help enable the universal implementation of ML systems across multiple health care disciplines and sites to ensure that ML models are being implemented in an ethical and practical manner. The ethical and practical implementation of Machine Learning Operations (MLOps) allows for the continued implementation of the ML tool in clinical practice while improving patient care [11,12]. Because the MLOps principles are a pre-existing framework that is adopted successfully in many industrial fields, such as technology, finance, and transportation, it may be a promising avenue to establish an MLOps workflow that can be standardized to ensure universal quality of ML tools across multiple medical domains and sites [11]. Here, we investigate how MLOps is adopted in the health care field, including considerations for implementing MLOps and propose a maturity framework within the context of health care applications.

MLOps is an extension of development operations, a software engineering practice that involves continuous integration (CI) and continuous delivery of software to ensure speed and high quality of production [11,13]. MLOps incorporates frameworks, such as continual learning (CL), where models are retrained and deployed based on new data when there is a decay in model performance; and continuous monitoring (CM) where the performance of the ML model is monitored [11,14]. Currently, the standards for MLOps include variations of the following steps: data preparation (data extraction and data engineering); model development (model training and measuring model performance); and model operationalization (model validation and testing in production, model serving and deployment, CM and CL) [11,12,14,15]. The first stage involves defining the use case and outcomes to inform the data collection and handling process. Data and feature engineering follows the data collection process, where important features or predictors are created. Furthermore, the data will be cleaned and normalized according to the requirements of the use case and model [12]. Following this stage, the best ML algorithm and hyperparameters will be selected using ML performance metrics. When the ML model is assessed to be the best model, the model will be presented to end-users on a user-friendly platform [16]. MLOps is the presence of CM, where the ML model performance is continuously monitored, and when the above steps are automated, the ML model is updated on new data when the ML model decays to a predefined metric (CL), allowing for the preservation of model performance [14].

Despite the emphasis on CM and CL within MLOps, not all organizations or use cases have the capability to reach full MLOps maturity [11,17]. There are various descriptions for the stages of MLOps maturity, with varying degrees of the implementation of CM and CL in the MLOps framework [11,17-20]. For example, Google has outlined the three stages of maturity for presented MLOps processes: (1) Manual process, (2) ML pipeline automation, and (3) CI and continuous delivery automation [19]. For the first stage, companies perform every step of the MLOps workflow pipeline manually. A problem with this method is that the ML framework is not frequently updated or monitored. The second stage is ML pipeline automation, where the data ingestion and data engineering steps are automated, allowing for the automation of model retraining and testing to ensure continuous delivery of model predictions. However, the deployment of the ML pipeline is completed manually. The final stage of ML pipeline automation is where the data ingestion and engineering steps, model training, testing, validation, serving, monitoring, and deployment of the ML pipeline are all automated. Microsoft has outlined a 5-level model, where MLOps practices are absent at level 0 as the ML model training and deployment are manually completed. Level 4, or full MLOps automated operations, is defined as the automated training, monitoring, and deployment of an ML system. Similar to the Microsoft maturity model, Garg et al [18] have outlined a 3-stage model where level 0 is the absence of automation of the ML system and level 2 is the full maturity of the system. John et al [11] described an MLOps maturity model from a different perspective, where the first stage of the maturity model is when an organization adopts automated data collection, and the final stage is when the organization adopts automated data collection, model deployment, and model monitoring.

In the context of health care, there are distinct considerations for the application of MLOps, as health care presents unique challenges to implementing MLOps. For instance, the clinical environment is unique in that we have additional stakeholder considerations (eg, clinician, patient, and community considerations), regulatory considerations, and considerations surrounding health care data [21]. Because of the evolving nature of clinical practice, regulatory environment, and population dynamics, there is a benefit to implementing a fully mature MLOps pipeline workflow with the inclusion of human oversight [22-25]. This is because the ML model's performance may decay as time progresses due to shifts in the patient population in terms of the patient demographics and the operation of the clinical environment (eg, changes in medical technologies, new procedures, and emergence of new disease outbreaks), which highlights the potential added benefit of an adaptive ML prediction system.

In this paper, we will investigate how the MLOps framework has been applied to health care, considerations for implementing MLOps, and the maturity framework within the context of health care applications. For our purposes, we have chosen to adopt a hybrid of all the MLOps maturity models for health care implementations, which can be applied to the literature described in the scoping review. In the hybrid description of the maturity model, we have identified 3 stages of maturity within the studies selected for this scoping review. These 3 stages include low, partial, and full maturity. Low maturity is the complete absence of CM and CL. Partial maturity is the presence of CM, with a lack of CL, as the retraining of the model was manually triggered by the engineering team. Full maturity is the presence of CM and CL.

Methods

Search Strategy

The Joanna Briggs Institute framework for scoping reviews was used to guide the execution of this study, as this framework provides a structured approach for searching, charting, and analyzing the data [26-31]. Details regarding the scoping review framework are presented in the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews), a checklist of essential items for scoping review reporting (Checklist 1) [32]. We conducted a 3-step search strategy proposed by Aromataris and Riitano [33]. The first 2 steps involved 2 rounds of searching, which were conducted to ensure that the search process was comprehensive. The first round consisted of a search of 4 databases (MEDLINE, Web of Science, Scopus, and Embase) and subsequent analysis of relevant keywords in the title and abstract of extracted literature (Multimedia Appendix 1). We have chosen these 4 databases because they capture the medical, engineering, and computer science fields. To specify, studies were screened for title, abstract, and keywords by reviewers YL and JT, with AX resolving conflicts. We collected keywords relating to MLOps and health care that were not included in the first round of searching. These keywords were incorporated into the final search terms described in Multimedia Appendix 1, and a second search where a total of 2712 studies were identified. The second search expanded the initial search terms to include a broader range of studies. The third step involved searching through the reference list of selected studies to identify additional papers, which were added to the papers that were screened. Five studies were identified and manually uploaded following hand-searching the reference lists of included studies. The most recent search was conducted on 29 January 2025.

Search Terms

Search terms relating to ML, development operations, and all health care fields were included. To ensure that search terms were comprehensive, 2 rounds of searching were performed. The first round involved collecting and brainstorming search terms relating to ML, MLOps, and health care. The second search involved collecting search terms from the first round of abstract, title, and keyword screening for the studies collected. For example, the term MLHOps was not present in the search terms for the first round of searching. However, it was identified as a keyword in one study within the first round of the search; hence, it was incorporated as a search term for the second round of searching. Finalized search terms from the Web of Science, Embase, Ovid MEDLINE, and Scopus were presented in Multimedia Appendix 1. Finally, we hand-searched the reference lists of the included studies for other potential MLOps studies that were not identified.

Selection Criteria

We used the PCC framework ("Population," "Concept," and "Context") to select the studies that will be included in the scoping review (Table 1; [31]). Population was defined as any population, as MLOps can be applied to any individual in any country in a health care setting. The main concept of interest is MLOps. The context pertains to all health care practices and fields. We only included studies where MLOps was implemented in a clinical application. This included studies that were a proof of concept of a clinical application, or studies implemented in a real-world clinical setting. Research papers consisting of peer-reviewed and preprint articles that discuss the application of MLOps in health care settings were included in this review. Studies were excluded if they were not written in English, did not discuss an instance of MLOps implementation in a medical setting, or did not focus on health care applications. Perspective pieces describing MLOps concepts were also excluded. There were no set limitations on the publication date.

Table 1. Summary of the "population," "concept," and "context" framework used in this scoping review, exclusion criteria, and other additional inclusion criteria.

Category	Description
Population	Any population with a medical condition.
Concept	Any applications or proof of concept involving MLOps[a].
Context	All clinical practices and fields.
Inclusion criteria	Research articles consisting of peer-reviewed and preprint articles that discuss the application of MLOps in health care settings.

Category	Description
Exclusion criteria	Studies were excluded if they were not written in English, did not discuss an instance of MLOps implementation in a medical setting, or did not focus on health care applications. Perspective pieces describing MLOps concepts were also excluded.

[a]MLOps: Machine Learning Operations.

Study Selection

Following the search for the studies, identified studies were first exported into Zotero (Corporation for Digital Scholarship) and then imported into Covidence (Veritas Health Innovation Ltd), an online platform for managing reviews. Duplicates were removed before title and abstract screening. Title and abstract screening and full-text screening were completed separately by YL, JT, and AX, and each study was reviewed by 2 reviewers according to the PCC criteria mentioned in the "search criteria" section. Conflicts that occurred in either stage were resolved by consensus among the reviewers.

Data Charting and Synthesis

The data extraction template was developed by YL using the 3-stage basic qualitative content analysis discussed by Pollock et al [31], and Elo and Kyngäs [34]. For the organization stage, the deductive approach was used, where an extraction framework was developed using pre-existing MLOps frameworks in literature [11,12,14,15,17,19,20]. For the MLOps maturity model, we adopted an inductive approach to account for the various interpretations of the MLOps maturity framework found in the literature. YL, JT, and AX analyzed the 19 chosen studies to find a common MLOps maturity framework that accommodates the diverse perspectives presented in these works [34-52]. Extraction was performed by YL and JT, with AX resolving any conflicts within the Covidence platform. The extraction form included the author and year, aim of the study, population and disease characteristics, location of study, MLOps maturity stage, MLOps workflow (data extraction, data preparation and data engineering, model training, measured ML metrics and evaluation, model validation and test in production, model serving and deployment, and CM and CL), and other considerations for MLOps in health care applications. The data from the extraction template were organized into Table 2 (author and year, aim of the study, population and disease characteristics, and location of study), Table 3 (MLOps workflow), and Table 4 (other considerations for MLOps in health care applications) (Multimedia Appendix 2).

Table 2. The objectives of each study included in the scoping review, along with the year of publication, author(s), study aim, study location, and the population or disease characteristics of individuals in the dataset used for Machine Learning Operations (if applicable).

Author and Year of publication	Population and Disease characteristics	Location	Aim of study
Bahaa et al [35], 2023	Patients with heart disease	United States	This study introduces a proof-of-concept model, aiming to turn unprocessed data into a useful product that provides utility through a rapid, scalable, and repeatable process. The proof-of-concept model was demonstrated through the UCI (University of California Irvine) heart dataset.
Bai et al [36], 2022	Emergency Department Visits	United States	This study aimed to develop a prototype Machine Learning Operations (MLOps) platform for machine learning (ML)–based clinical tools.
Granlund et al [37], 2021	Patients undergoing joint replacement surgery	Finland	This study presents a case study of Oravizio (Solita), a ML-based medical device used for predicting the risk of joint replacement surgery, in addition to discussing policy considerations for ML-based medical devices.
Kanbar et al [38], 2022	Patients diagnosed with epilepsy	United States	This study seeks to develop clinical decision support tools that integrate electronic health records and patient notes to make predictions on surgical candidacy for patients with seizure and to screen emergency department patients for eligibility to enroll in clinical trials.
Karácsony et al [39], 2021	Patients diagnosed with epilepsy	Germany	In this study, an MLOps framework was tailored to the analysis of 2D and 3D seizure videos to classify different types of seizures.
Kleftakis et al [40], 2022	N/A[c]	Greece	This study presents an MLOps framework that monitors the health of the entire body and the patient's overall health status.
Krishnan et al [41], 2022	Internal medicine, critical care, and patients with chest X-rays	Canada	This study details the MLOps framework applied to several medical use cases, such as mortality prediction and imaging applications.

Author and Year of publication	Population and Disease characteristics	Location	Aim of study
Kundu and Bilgaiyan [42], 2023	Patients diagnosed with COVID-19	India	In this study, a different and efficient approach to automating the hyper-parameter tuning process is proposed with the use of DevOps[a] tools for automating the workflow.
Meel and Bodepudi [43], 2021	Patients with skin cancer	United States	This study presents Melatect, a ML prediction tool embedded in an IOS app that predicts the malignancy of a skin lesion.
Mirza et al [44], 2023	N/A	United States	This study details the development of an MLOps platform to predict the workload at clinical trial sites.
Tougui et al [45], 2022	Patients diagnosed with Parkinson	Morocco	This study presents a ML application that uses voice recordings to predict the presence of Parkinson disease. The prediction is presented in a web application.
Tseng et al [46], 2022	Patients undergoing cardiac arrest	Taiwan	This study defined a new implementation guide for running a user-friendly ML system for predicting In-hospital cardiac arrest (IHCA). This system uses the Fast Healthcare Interoperability Resources (FHIR) to manage health care data and the ML application.
Ghosh and Chaki [47], 2025	Patients with kidney cancer	India	This study aims to automate the detection of kidney tumors from computed tomography scans via the integration of MLOps principles for generating predictions.
Imrie et al [48], 2023	Patients with Diabetes	United States	This study seeks to automate the training and deployment of ML models in clinical environments to support physician decision-making, with a demonstration in diabetes risk prediction.
Lombardo et al [49], 2024	Hospital staff and patients	Italy	This study provides a proof-of-concept model for the location-based service tracking of hospital staff to make predictions on the location-based trajectories of hospital staff. This has potential use cases in the prediction of the trajectories of the hospital staff to determine potential wait times in the hospital or to identify potential security concerns.
Lutnick et al [50], 2023	N/A	United States	This study aims to develop an MLOps-based system that enables nontechnical users, such as clinicians, to leverage pretrained ML models to develop predictions for histology images.
Markowitz et al [51], 2024	Patients with central nervous system tumors	United States	This study aims to develop an MLOps pipeline for classifying central nervous system tumors using data from clinical diagnostic tools, including the MethylationEPIC v2.0 BeadChip (Illumina) and other DNA/RNA biomarker-based cancer screening technologies.
Mathew and Joseph [52], 2023	Patients with brain cancer	India	This study develops an MLOps pipeline for the prediction of brain tumors using MRI[b] images.
Moskalenko and Kharchenko [53], 2024	N/A	Ukraine	This study aims to develop a resilient MLOps system for health care applications, as systems in health care applications may be impacted from adversarial attacks and distribution drifts.

[a] DevOps: Development Operations

[b] MRI: Magnetic Resonance imaging.

[c] N/A: not applicable.

Table 3. The checklist for Machine Learning Operations pipeline and Machine Learning Operations maturity[a].

	MLOps[b] maturity level	Data preparation		Model development		Model operationalization		
		Data extraction	Data engineering	Model training	Measured ML[c] Metrics and Evaluation	Model validation and test in production	Model serving and deployment	Continuous (CM[d], CL[e])
Granlund et al [37], 2021	Low maturity	✓[f]	✓	✓	✓	✓	✓	
Karacsony et al [39], 2021	Low maturity	✓	✓	✓			✓	

	MLOps[b] maturity level	Data preparation		Model development		Model operationalization		
		Data extraction	Data engineering	Model training	Measured ML[c] Metrics and Evaluation	Model validation and test in production	Model serving and deployment	Continuous (CM[d], CL[e])
Tougui et al [45], 2022	Low maturity	✓	✓	✓	✓	✓	✓	
Imrie et al [48], 2023	Low maturity	✓	✓	✓	✓	✓	✓	
Lutnick et al [50], 2023	Low maturity	✓	✓	✓		✓	✓	
Bahaa et al [35], 2023	Partial maturity	✓	✓	✓	✓		✓	CM only
Tseng et al [46], 2022	Full maturity	✓	✓	✓	✓		✓	✓
Kanbar et al [38], 2022	Full maturity	✓	✓	✓	✓	✓	✓	✓
Kleftakis et al [40], 2022	Full maturity	✓	✓	✓	✓	✓	✓	✓
Krishnan et al [41], 2022	Full maturity	✓	✓	✓	✓			✓
Kundu and Bilgaiyan [42], 2023	Full maturity	✓	✓	✓	✓			✓
Meel and Bodepudi [43], 2021	Full maturity	✓	✓	✓	✓	✓	✓	✓
Mirza et al [44], 2023	Full maturity	✓	✓	✓	✓	✓	✓	✓
Bai et al [36], 2022	Full maturity	✓	✓	✓	✓	✓	✓	✓
Ghosh and Chaki [47], 2025	Full maturity	✓	✓	✓	✓	✓	✓	✓
Lombardo et al [49], 2024	Full maturity	✓	✓	✓	✓	✓	✓	✓
Markowitz et al [51], 2024	Full maturity	✓	✓	✓	✓	✓		✓
Mathew and Joseph [52], 2023	Full maturity	✓	✓	✓	✓	✓	✓	✓
Moskalenko and Kharchenko [53], 2024	Full maturity	✓		✓	✓	✓		✓

[a] Blank cells denote that a MLOps step was not discussed or found within the paper.
[b] MLOps: Machine Learning Operations.
[c] ML: Machine Learning.
[d] CM: Continuous Monitoring.
[e] CL: Continual learning.
[f] Presence of the MLOps framework step.

Table 4. Other health care–specific considerations for the Machine Learning Operations framework.

Author and year of publication	Other considerations
Bahaa et al [35], 2023	• Team considerations for building the DataOps[a] pipeline were discussed as team members relating to domain knowledge and technical skill sets were considered as crucial aspects of the pipeline. • Other considerations include the archiving or deleting of unused data to ensure the pipeline is cost-effective.
Bai et al [36], 2022	• The authors outlined the core tenets of an MLOps[b] system in health care, which are that the platform must be ethical, auditable, adaptable, automated, and accessible. This ensures that patient privacy is protected, the machine learning (ML) system is transparent, and it maintains a safe level of performance. • Team considerations were key stakeholders like a health informatics professional, health care researcher, and chief information officer that work together to define the key clinical use case and govern the development and safety of the ML model.
Granlund et al [37], 2021	• This paper also focused on the regulatory considerations for ML-based medical device systems. • Four important considerations for deploying ML-based medical device systems in clinical practice involve the medical device's ability to provide clinical benefit to the target patient group, transparency of the device, the performance or safety of the medical device, and risk management of the ML system.
Kanbar et al [38], 2022	• Considerations for patient privacy were discussed as the servers used for prediction ran on a secured Health Insurance Portability and Accountability Act-compliant server. • Team considerations for building an MLOps pipeline. Need to consider the team structure. Clinicians, researchers, support health care staff, and research coordinators worked with the bioinformaticians to create this health care framework.
Karácsony et al [39], 2021	• Considerations for data storage take into account the health data privacy laws in the European Union. Due to data privacy concerns, the epilepsy data is stored in a centralized private server in the cloud where the data can only be accessed via the encrypted Virtual Private Network connection. Patient identifiers were also removed to maintain privacy. • There is also consideration of computational speed and resources in the creation of the MLOps pipeline highlighted in this paper as a central server that hosts the patient data and ML models to minimize the transfer of large file sizes.
Imrie et al [48], 2023	• This study also details the importance of defining the advantages of using ML over traditional statistical methods for clinical prediction and demonstrating the importance of feature importance for debugging ML models within the production pipeline.
Lutnick et al [50], 2024	• Considerations for data access were discussed and integrated into the MLOps platform for histology image analysis.
Moskalenko and Kharchenko [53], 2024	• In medical applications, there is more need for MLOps systems to be more resilient as it is a high-stakes environment. Hence, MLOps systems need to be developed for improved safety and adaptability within the medical system. • Considerations for data protection and model development were discussed to be avenues for adapting a MLOps system for medical applications.

[a] DataOps: Data Operations.
[b] MLOps: Machine Learning Operations.

Results

Data Analysis and Presentation

The search protocol was described as a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart in Figure 1. In total, 2712 studies were identified based on the search terms in Multimedia Appendix 1 from Scopus, Web of Science, Embase, Ovid MEDLINE, and the reference list of included papers. In the screening phase, 1268 studies were screened following the removal of 1444 duplicate studies. There were 1206 excluded studies based on the title, abstract, and keyword screening. Papers without MLOps and health care–related terms or concepts in the title, abstract, and keyword screening were excluded. A full-text screening was completed for the 62 studies to assess eligibility for inclusion. The final 19 extracted papers were presented in 3 tables. Table 2 describes the aims of each selected study, year of publication, author, location where the study was conducted, and the population and disease characteristics. Table 3 describes elements of the MLOps workflow pipeline used by each of the studies, in addition to a summary of the MLOps maturity model for each study. Table 4 outlines additional considerations for health care applications not present within the MLOps workflow pipeline. Figure 2 outlines the MLOps workflow pipeline as a graphical representation of the number of studies per stage.

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for screening Machine Learning Operations health care studies. MLOps: Machine Learning Operations.

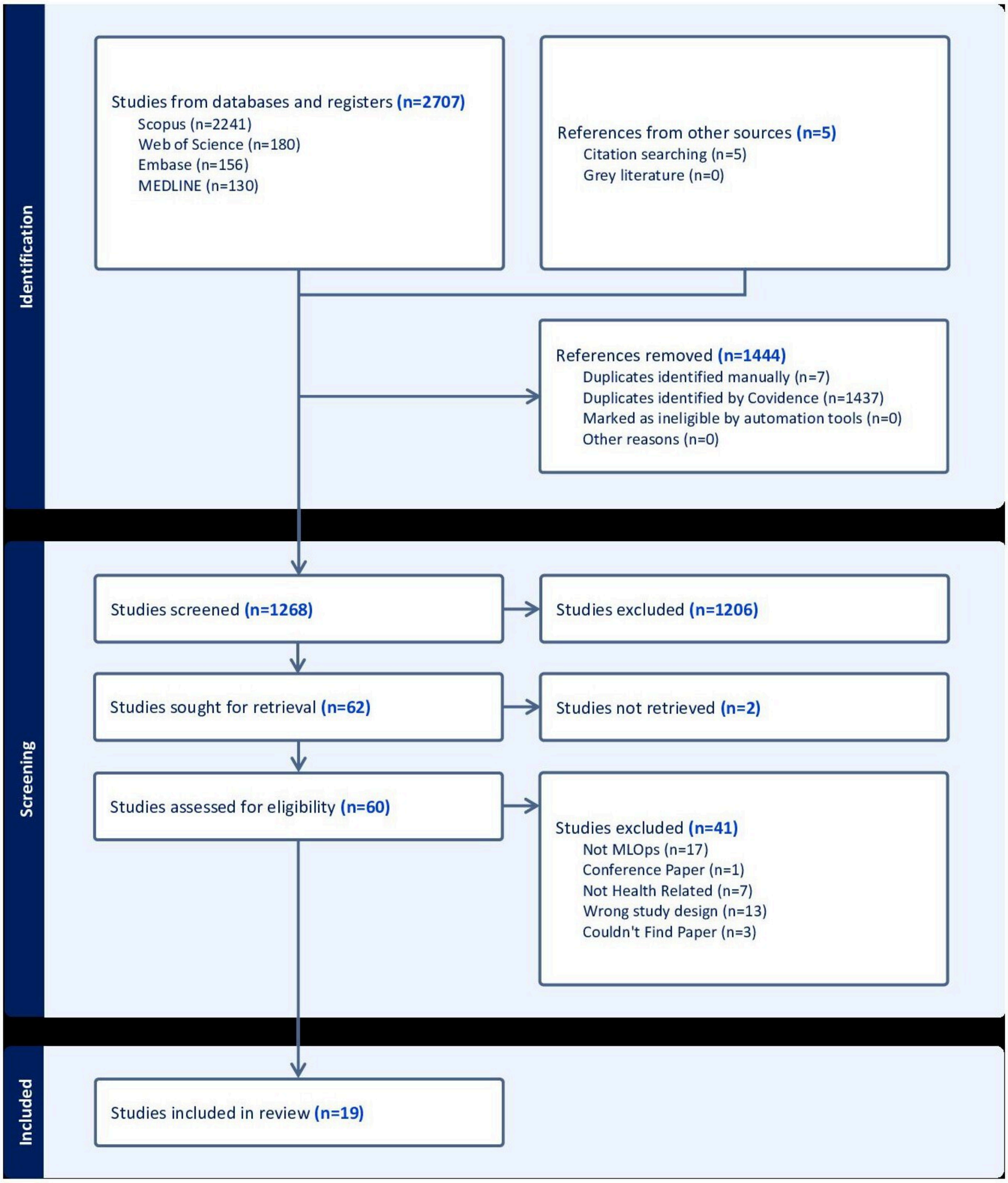
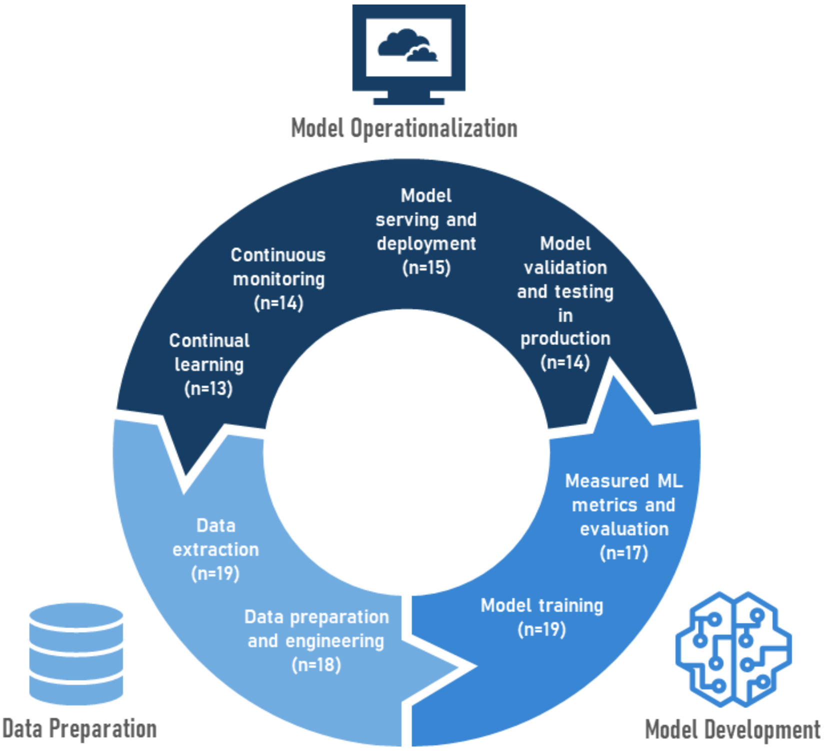

Identification
Studies from databases and registers (n=2707)
Scopus (n=2241)
Web of Science (n=180)
Embase (n=156)
MEDLINE (n=130)
References from other sources (n=5)
Citation searching (n=5)
Grey literature (n=0)
References removed (n=1444)
Duplicates identified manually (n=7)
Duplicates identified by Covidence (n=1437)
Marked as ineligible by automation tools (n=0)
Other reasons (n=0)
Screening
Studies screened (n=1268)
Studies excluded (n=1206)
Studies sought for retrieval (n=62)
Studies not retrieved (n=2)
Studies assessed for eligibility (n=60)
Studies excluded (n=41)
Not MLOps (n=17)
Conference Paper (n=1)
Not Health Related (n=7)
Wrong study design (n=13)
Couldn't Find Paper (n=3)
Included
Studies included in review (n=19)

Figure 2. Graphical representation of Machine Learning Operations workflow pipeline and maturity stage. ML: machine learning.


Model Operationalization
Model serving and deployment (n=15)
Continuous monitoring (n=14)
Model validation and testing in production (n=14)
Continual learning (n=13)
Measured ML metrics and evaluation (n=17)
Data extraction (n=19)
Model training (n=19)
Data preparation and engineering (n=18)
Data Preparation
Model Development

Data Extraction

The MLOps workflow pipeline from all 19 studies discussed the implementation of a data extraction step, where the data used for developing the ML model were extracted from a pre-existing database [35,41-45,47,48,50,52,53], or the collection of data from various electronic health records or the collection of patient physiological measurements from clinical settings [36-40,46,49]. Markowitz et al [51] obtained data from a combination of pre-existing databases and in-hospital measurements.

Data Preparation and Data Engineering

Data preparation and engineering involve loading data into a database, preprocessing raw data to create relevant predictive variables or features, and cleaning and formatting data so it is suitable for training and testing. There were 18 studies [35-52] that described data preparation and engineering steps for deploying the ML model. Because medical applications require the protection of patient privacy, studies such as Bai et al [36] described the removal of 18 protected health information identifiers. Kanbar et al [38], and Lutnick et al [50], also discussed the importance of protecting sensitive patient data during the MLOps workflow. To select the top features for model training, algorithms such as least absolute shrinkage and selection operator, or ElasticNet were used on the training dataset [37,45]. The assessment of data quality was also discussed in Kleftakis et al [40]. Moskalenko and Kharchenko [53] did not describe the feature engineering step involved in the creation of their MLOps workflow.

Model Training

The model training step was described in all 19 studies [35-53] for the development of the models. Studies also emphasized the importance of transparency and of the ML model within the context of health care applications [37].

ML Model Evaluation

There were 17 studies [35-38,40-49,51-53] out of 19 studies that described the model metrics used to evaluate the ML models used for the studies. Examples of metrics used to evaluate model performance include accuracy, F1 (harmonic mean of precision and recall), the area under the receiver operating curve, sensitivity, and specificity. Two studies [39,50] did not describe the ML model evaluation technique used.

Model Validation and Testing in Production

Model validation and testing in production involve elements relating to internal and external validation. Internal validation involves assessing the performance of the ML model on the dataset used to develop the model [54]. External validation involves the assessment of the ML model performance on a dataset not used for the development of the ML model. Furthermore, testing in production refers to the process of testing the model in a real-world clinical scenario. Examples of testing in production may refer to the comparison of the model prediction to a clinician's prediction. Fourteen [34,36-38,41,42,45-52] out of 19 studies discussed how the ML models were validated and tested in production. A total of 11 studies [37,38,40,44,45,47-50,52,53] performed internal validation only and 3 studies [36,43,51] performed both external and internal validation.

There were 2 studies that emphasized the importance of testing the model in production to ensure the model has clinical benefits and is ethical for use in patient populations.

Kanbar et al [38] performed shadow deployment, which is the comparison of ML model accuracy to the clinician's accuracy in terms of decision-making. Specifically, the prediction from the model was compared against a prediction from an epileptologist for 1 year. In addition to model performance against a baseline, the bias of the model was also considered. Kleftakis et al [40] emphasized the importance of testing whether a model surpassed the accuracy of a previous model before it could be placed into production.

Model Serving and Deployment

Model serving and deployment is the process of packaging the ML model and releasing it using an application programming interface on a user-friendly interactive platform. Fifteen studies [35-40,43-50,52] out of 19 studies described the process in which the ML model was released into a user-friendly platform.

Maturity Stage and Continuous Monitoring and Continuous Learning

Here we propose 3 MLOps maturity stages: low maturity, partial maturity, and full maturity. Low maturity was defined as the absence of any CM or CL. Hence, the ML pipeline was used to make predictions with no regard for performance degradation; essentially, it is a "locked" model. Five studies [37,39,45,48,50] out of 19 studies have low MLOps maturity. The second stage is partial maturity, where there was the presence of CM by the MLOps system; however, there was a lack of automatic triggering of the MLOps workflow pipeline, as retraining of the ML was manually scheduled by the engineering team [35]. One study involved partial maturity of the MLOps workflow [35]. The third stage involves the full maturity of the model, where the model is continuously monitored and exhibits CL. Here, the model is retrained using new data from either a pre-existing database or newly collected data, and the updating of the MLOps workflow pipeline is determined by the deterioration in the model's performance. Thirteen studies [36,38,40-44,46,47,49,51-53] out of 19 studies contained full maturity of their MLOps workflow pipelines.

Other Considerations Specific to Health Care Applications

Because of the sensitive nature of health care data, considerations relating to ethical practice, privacy, stakeholder management, and MLOps design principles were discussed in 8 studies. Specifically, 5 studies [36-39,50] discussed the importance of data management to protect patient privacy. Two studies [38,39] emphasized the importance of applying server governance to adhere to the legislation present in the region where the study was used. Granlund et al [37] also discussed the importance of legislation on how an MLOps workflow pipeline could be deployed. Notably, the US Food and Drug Administration and the European Union on Medical Devices have supported the use of "locked" or ”frozen” ML models, models that cannot be improved or modified following the release of the model into production [37,55,56]. Another essential aspect of integrating a successful MLOps workflow pipeline into health care applications was the selection of relevant stakeholders and experts [35,36]. Health care professionals and an effective engineering team were needed to run an MLOps workflow pipeline, as health care professionals provide domain knowledge to define the use case and govern the data usage, and the engineering teams work closely with the health care professionals to build the MLOps application to ensure that it is user-friendly. Finally, 2 studies discuss how the MLOps workflow can be adapted to health care applications. Moskalenko and Kharchenko [53] develop a resilient MLOps pipeline that can better withstand adversarial attacks and model performance decays in consideration of the high-stakes environment of health care applications. Imrie et al [48] discuss how an MLOps workflow needs to justify the use of ML in health care applications through the comparison of ML to traditional statistical methods, such as the Cox Proportional Hazards model. Furthermore, the implementation of ML in health care systems should be interpretable to clinical staff for the improved trust in the ML application.

Discussion

Principal Findings

MLOps is crucial for health care applications as it provides a structured framework to operationalize ML models, ensuring they are implemented ethically and practically. By following a tailored MLOps framework, health care organizations can streamline the deployment, monitoring, and updating of ML models, thus enhancing patient care and operational efficiency [57]. MLOps also describes a maturity model with various stages, from manual processing of ML pipelines to fully mature pipelines with CM and CL, indicating different stages of maturity and sophistication in the implementation [11,14]. The MLOps workflow and maturity model help organizations understand their current capabilities and identify areas for improvement to achieve higher efficiency and reliability in deploying ML models in health care. In this study, we highlighted the steps of an MLOps workflow pipeline, which included data extraction, data preparation and data engineering, model training, model evaluation, model validation and testing in production, model serving, and CM/CL.

Among all the selected studies, the data extraction phase was present. However, there were differences in the origin of the data among all studies. One major difference was the origin of the data, as 12 studies [35,41-45,47,48,50-53] out of 19 studies used a pre-existing database to test their proof-of-concept MLOps implementation. The fact that the majority of studies use a pre-existing database may suggest that there are barriers to data access, a well-known challenge in digital health tools [58,59]. Especially when these studies originate from diverse locations (Table 2), they suggest that MLOps is still an emerging area of research globally, and it may not be feasible to design large-scale implementation with real-time data in medical applications.

All 19 studies differed in model validation and testing in production, model serving and deployment, and CM and CL [36-39,43-46,48,51,52]. For example, studies differed in how they performed validation. This can be problematic, as the method of validation is a critical step in determining whether an MLOps implementation can be generalized to novel datasets or clinical sites. Internal validation refers to assessing the model's performance on the dataset used for its development, ensuring it performs well within the known parameters [54]. External validation, however, involves testing the model on a novel dataset, which was not used during the development phase. This step is crucial as it demonstrates the model's ability to generalize to unseen data, making it the gold standard for validation, as MLOps implementations may need to be used for multiple different scenarios within a clinical domain [60,61]. In addition to validation, models must be tested in the clinical setting to ensure that they can improve patient experience and are able to perform as well or better compared to clinical predictions. Kanbar et al [38] demonstrated an instance of testing in production where shadow deployment of the model against the clinical prediction occurred to determine the performance of the model against the clinician's prediction. Shadow deployment is where the model generates unused predictions in the background, which tests the effectiveness of the model in a real-life clinical environment without impacting patient experience, an essential step in determining the efficacy of the model in a real-life setting [62]. This is essential for determining whether a MLOps implementation will be safe and useful for operational decision-making in a clinical setting.

Five studies did not incorporate CM and CL due to resource constraints or the elementary stage of their MLOps implementation [34–38]. For example, Granlund et al [37] did not have CM and CL since their model was deployed in a "locked" or "frozen" state as governed by the European Union legislation for medical devices. "Frozen" models, which are not updated postdeployment, are supported by some regulatory bodies to maintain the consistency and safety of the ML models [63,64]. "Frozen" models are beneficial because they provide a stable and predictable performance, which is critical for maintaining trust among health care professionals and patients [65]. However, the downside is that "frozen" models may become outdated as new data and trends emerge, potentially reducing their accuracy over time. Without proper CM and CL, models can degrade in performance, resulting in unreliable predictions which can create more risk for the ML model use case [66,67]. While full MLOps maturity can be preferable with full human oversight, it is not always feasible due to limitations, such as regulations, expertise, and infrastructure. Furthermore, there are risks to full MLOps maturity as well. Due to adversarial attacks on the newly trained ML model, new errors can be introduced by training on new data (eg, a change in how the data was recorded, leading to inaccurate data), and the ML model may undergo catastrophic forgetting [22,24,25]. Hence, there must be a balance between maintaining model stability with the risks introduced by retraining while ensuring periodic updates to capture new insights and evolving data trends. Finally, another important consideration for CM and CL is the decision of the performance metric that will trigger the retraining of the ML model. The choice of the metric is important as well since different ML model metrics measure different aspects of the model performance; for instance, sensitivity focuses on the ability of an ML model to identify true positive cases, and specificity measures the ability of an ML model to identify true negatives.

Clinical Considerations for MLOps Implementations

Furthermore, an essential consideration for medical MLOps applications discussed by 2 studies [38,43] was the presence of bias in ML models and the incorporation of bias checking within the MLOps workflow pipeline. Bias in ML models can lead to unfair treatment of certain patient groups and hinder the model's ability to make accurate predictions for individuals of certain backgrounds and underestimate their health care needs [68,69]. Bias can stem from various sources, including nonrepresentative training data, model design choices, and the historical and systemic inequities embedded in the health care system that resulted from differential routine care for different populations [70-72]. To make improvements upon this concern, we must ensure diverse and representative training data and conduct fairness audits. These audits involve evaluating the model's performance across different groups to identify and mitigate any biases. External validation can also mitigate bias by testing the model's performance on completely new data, ensuring it can generalize beyond the training environment, which is crucial for reliable clinical use. This process helps identify any hidden biases that were not apparent during internal validation. Furthermore, in cases where external validation may not be feasible, temporal validation can be used where the validation set consists of the data from the same patient population in a future time frame [73]. Finally, testing the ML model in a clinical environment through shadow deployment can help identify potential biases and assess its effectiveness and safety in real-world clinical decision-making.

, because the health care environment itself contains special considerations for patient care, there are notable distinctions between ML deployments in health care and other industries. This is because traditional MLOps principles focus on the technical aspects of operationalizing ML algorithms, such as speed of delivery and management of the data workflow, in addition to the engineering teams working on the MLOps pipeline workflows [12]. In comparison, health care is a conservative environment where there are numerous other considerations surrounding patient ethics, legal implications, clinical workflows, and characteristics of the health care data that will impact how an MLOps workflow will be operated. One key challenge is that current health care data infrastructure is primarily designed for the recording of health service–related events for operational purposes. However, this strategy may not support the technical requirements for MLOps implementation. Thus, health care data infrastructure needs to be adapted for MLOps applications. Furthermore, an essential step of MLOps specific to health care applications would be the inclusion of community stakeholders throughout the entire process of developing the ML tools, especially when engaging community stakeholders enables the development of inclusive ML models that add value to the community and the clinical setting [74,75]. Specifically, it is important for community stakeholders and clinicians to collaborate to develop acceptable standards for features used in the model, model performance, model interpretability, and how to maintain the model long term. Especially when medical systems demand adaptive ML models due to changes in the patient population or clinical environment, as ML model performance will decay over time [76,77]. Achieving full MLOps maturity in health care MLOps systems requires close human oversight. Human oversight must be present for the risk management of the MLOps implementation to ensure that the ML model is still performing to an acceptable standard before the redeployment of the new model [22]. Strategies for ensuring the safety of the newly retrained model could be the shadow deployment of the model or testing models on specific sensitivity population subgroups and consultation with clinicians before a full release. However, there may be instances where full MLOps maturity may not be feasible for clinical environments, as clinical workflows may not have sufficient infrastructure to support the full maturity of the ML pipeline workflow. In these situations, the use case of the clinical models may be important to work with key stakeholders to determine the level of maturity that is acceptable. For example, in clinical use cases where there are minimal shifts in the data distributions, the ML model may be manually updated at lower frequencies.

Limitations

Five main limitations were identified within this study. First, the quality of the studies varied; hence, it was a challenge to identify every MLOps workflow pipeline step detailed in Table 3 and Multimedia Appendix 3, as each study had a different way of describing the MLOps workflow pipeline steps. Furthermore, the reported details of each MLOps workflow pipeline differed, as certain studies gave details regarding each step while other studies did not discuss certain steps of their pipeline. Hence, it was assumed that if a step in the MLOps workflow pipeline was not discussed, the authors did not complete the step. In addition, there were different descriptions for CM and CL in different studies, so if a described step matched our definition, we would assume that CM and CL were completed. Second, MLOps is poorly defined in the health care realm; there were limited studies in the health care space identified as MLOps studies, so there may be MLOps studies that may not be identified and screened by the protocol. A potential mitigation strategy for the future may be to include MeSH terms within the search strategy. Third, because we only included studies in English, studies in other languages were left out. Fourth, scoping reviews inherently have limitations [78]. While they aim to summarize a broad range of literature, this breadth can sometimes compromise the depth and quality of information on specific topics. Quality assessments of the included studies were not performed as these evaluations are beyond the methodological scope and purpose of scoping reviews [79]. This absence of quality evaluation can hinder the assessment of the strength and reliability of the evidence. Finally, the extracted studies contain limitations to their research design, as health care–specific essential steps for MLOps were not discussed.

Conclusions

In conclusion, we have examined how MLOps is implemented in health care settings and proposed a 3-stage MLOps maturity framework for health care. Even though certain clinical settings may not have full MLOps maturity, it is important to push for full maturity as ML models can better reflect the dynamic nature of the health care data and the population it serves. Widespread clinical use requires rigorous validation, and CM and CL, and close collaboration with health care providers and regulatory units to ensure these models and the platform that the models are hosted on meet the high standards required for patient care. Current MLOps implementations exist at 3 different maturity stages: low maturity, partial maturity, and full maturity, with thirteen models discussed in the reviewed studies having shown promising results and being on the path to clinical implementation with full MLOps maturity.

Acknowledgments

This research was undertaken, in part, thanks to funding from the Canada Research Chairs program, Alberta Innovates, the Institute for Advancements in Mental Health, Mental Health Foundation, Mental Health Research Canada, MITACS Accelerate program, Simon & Martina Sochatsky Fund for Mental Health, Howard Berger Memorial Schizophrenia Research Fund, the Abraham & Freda Berger Memorial Endowment Fund, the Alberta Synergies in Alzheimer's and Related Disorders (SynAD) program, University Hospital Foundation and University of Alberta.

Authors' Contributions

Conceptualization: YL, BC
Writing – original draft: YL, JT, AX
Writing – review & editing: YL, JT, AX, RG, JH, AJG, BC
Resources (instrumental support): RG, JH, AJG, BC
Supervision: BC

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of search terms used for Ovid MEDLINE, Scopus, Web of Science, and Embase.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Extraction template categories and locations of extracted items.
[DOCX File (Microsoft Word File), 12 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Full descriptions extracted from each study related to Machine Learning Operations pipelines and maturity frameworks.
[DOCX File (Microsoft Word File), 30 KB-Multimedia Appendix 3]

Checklist 1

Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews checklist for Scoping Reviews.
[DOCX File (Microsoft Word File), 15 KB-Checklist 1]

References

1. El Naqa I, Murphy MJ. What is machine learning? In: Machine Learning in Radiation Oncology. Springer; 2015:3-11. [doi: 10.1007/978-3-319-18305-3_1] ISBN: 978-3-319-18305-3
2. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med. 2020;3(1):126. [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
3. Rebala G, Ravi A, Churiwala S. Machine Learning Definition and Basics. An Introduction to Machine Learning Springer; 2019:1-17. [doi: 10.1007/978-3-030-15729-6_1] ISBN: 978-3-030-15729-6
4. Poddar M, Marwaha JS, Yuan W, Romero-Brufau S, Brat GA. An operational guide to translational clinical machine learning in academic medical centers. NPJ Digit Med. May 17, 2024;7(1):129. [doi: 10.1038/s41746-024-01094-9] [Medline: 38760407]
5. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med. Apr 26, 2023;6(1):74. [doi: 10.1038/s41746-023-00811-0] [Medline: 37100953]
6. Li Y, Song Y, Sui J, et al. Prospective prediction of anxiety onset in the Canadian longitudinal study on aging (CLSA): a machine learning study. J Affect Disord. Jul 2024;357:148-155. [doi: 10.1016/j.jad.2024.04.098]
7. Liu YS, Kiyang L, Hayward J, et al. Individualized prospective prediction of opioid use disorder. Can J Psychiatry. Jan 2023;68(1):54-63. [doi: 10.1177/07067437221114094] [Medline: 35892186]
8. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. Feb 22, 2019;23(1):64. [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
9. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun. Aug 11, 2020;11(1):3923. [doi: 10.1038/s41467-020-17419-7] [Medline: 32782264]
10. Song Y, Qian L, Sui J, et al. Prediction of depression onset risk among middle-aged and elderly adults using machine learning and Canadian Longitudinal Study on Aging cohort. J Affect Disord. Oct 15, 2023;339:52-57. [doi: 10.1016/j.jad.2023.06.031] [Medline: 37380110]
11. John MM, Olsson HH, Bosch J. Towards mlops: a framework and maturity model. Presented at: 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA); Sep 1-3, 2021:1-8; Palermo, Italy. [doi: 10.1109/SEAA53835.2021.00050]
12. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access. 2023;11:31866-31879. [doi: 10.1109/ACCESS.2023.3262138]
13. Makinen S, Skogstrom H, Laaksonen E, Mikkonen T. Who needs mlops: what data scientists seek to accomplish and how can mlops help? Presented at: 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN); May 30-31, 2021:109-112; Madrid, Spain. [doi: 10.1109/WAIN52551.2021.00024]
14. Huyen C. Designing Machine Learning Systems. O'Reilly Media, Inc; 2022. ISBN: 978-1-09-810793-2
15. Alla S, Adari SK. What is mlops? In: Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure Berkeley. Apress; 2021:79-124. [doi: 10.1007/978-1-4842-6549-9_3] ISBN: 978-1-4842-6549-9

16. Symeonidis G, Nerantzis E, Kazakis A, Papakostas GA. MLOps - definitions, tools and challenges. Presented at: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC); Jan 26-29, 2022:0453-0460; Las Vegas, NV. [doi: 10.1109/CCWC54503.2022.9720902]
17. Lima A, Monteiro L, Furtado A. MLOps: practices, maturity models, roles, tools, and challenges – a systematic literature review. Presented at: 24th International Conference on Enterprise Information Systems. SCITEPRESS - Science and Technology Publications. 308-320; Online Streaming, --- Select a Country ---. 2022.[doi: 10.5220/0010997300003179]
18. Garg S, Pundir P, Rathee G, Gupta PK, Garg S, Ahlawat S. On continuous integration / continuous delivery for automated deployment of machine learning models using mlops. Presented at: 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE); Dec 1-3, 2021:25-28; Laguna Hills, CA. [doi: 10.1109/AIKE52691.2021.00010]
19. MLOps: continuous delivery and automation pipelines in machine learning. Google Cloud. 2023. URL: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning [Accessed 2023-11-01]
20. Machine learning operations maturity model - azure architecture center. Learn Microsoft. 2023. URL: https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/mlops-maturity-model [Accessed 2023-11-02]
21. Zayas-Cabán T, Haque SN, Kemper N. Identifying opportunities for workflow automation in health care: lessons learned from other industries. Appl Clin Inform. May 2021;12(3):686-697. [doi: 10.1055/s-0041-1731744] [Medline: 34320683]
22. Babic B, Gerke S, Evgeniou T, Cohen IG. Algorithms on regulatory lockdown in medicine. Science. Dec 6, 2019;366(6470):1202-1204. [doi: 10.1126/science.aay9547] [Medline: 31806804]
23. Lee S, Yin C, Zhang P. Stable clinical risk prediction against distribution shift in electronic health records. Patterns (N Y). Sep 8, 2023;4(9):100828. [doi: 10.1016/j.patter.2023.100828] [Medline: 37720334]
24. Sparrow R, Hatherley J, Oakley J, Bain C. Should the use of adaptive machine learning systems in medicine be classified as research? Am J Bioeth. Oct 2024;24(10):58-69. [doi: 10.1080/15265161.2024.2337429] [Medline: 38662360]
25. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. Lancet Digit Health. Jun 2021;3(6):e337-e338. [doi: 10.1016/S2589-7500(21)00076-5] [Medline: 33933404]
26. Khalil H, Peters MD, Tricco AC, et al. Conducting high quality scoping reviews-challenges and solutions. J Clin Epidemiol. Feb 2021;130:156-160. [doi: 10.1016/j.jclinepi.2020.10.009] [Medline: 33122034]
27. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050]
28. Peters MDJ, Godfrey C, McInerney P, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth. Apr 1, 2022;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
29. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. JBI Manual for Evidence Synthesis. JBI. 2020. URL: https://synthesismanual.jbi.global [Accessed 2025-03-05]
30. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. Oct 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
31. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth. Mar 1, 2023;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
32. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850]
33. Aromataris E, Riitano D. Constructing a search strategy and searching for evidence. A guide to the literature search for a systematic review. Am J Nurs. May 2014;114(5):49-56. [doi: 10.1097/01.NAJ.0000446779.99522.f6] [Medline: 24759479]
34. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs. Apr 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
35. Bahaa S, Ghalwash AZ, Harb H. DataOps lifecycle with a case study in healthcare. Int J Adv Comput Sci Appl. 2023;14(1). [doi: 10.14569/IJACSA.2023.0140115]
36. Bai E, Song SL, Fraser HSF, Ranney ML. A graphical toolkit for longitudinal dataset maintenance and predictive model training in health care. Appl Clin Inform. Jan 2022;13(1):56-66. [doi: 10.1055/s-0041-1740923] [Medline: 35172371]
37. Granlund T, Stirbu V, Mikkonen T. Towards regulatory-compliant MLOps: Oravizio's journey from a machine learning experiment to a deployed certified medical product. SN Comput Sci. Sep 2021;2(5):342. [doi: 10.1007/s42979-021-00726-1]
38. Kanbar LJ, Wissel B, Ni Y, et al. Implementation of machine learning pipelines for clinical practice: development and validation study. JMIR Med Inform. Dec 16, 2022;10(12):e37833. [doi: 10.2196/37833] [Medline: 36525289]
39. Karacsony T, Loesch-Biffar AM, Vollmar C, Noachtar S, Cunha JPS. DeepEpil: towards an epileptologist-friendly AI enabled seizure classification cloud system based on deep learning analysis of 3D videos. Presented at: 2021 IEEE

EMBS International Conference on Biomedical and Health Informatics (BHI). Jul 27-30, 2021:IEEE. 1-5; Athens, Greece. [doi: 10.1109/BHI50953.2021.9508555]

40. Kleftakis S, Mavrogiorgou A, Mavrogiorgos K, Kiourtis A, Kyriazis D. Digital twin in healthcare through the eyes of the vitruvian man. In: Chen YW, Tanaka S, Howlett RJ, Jain LC, editors. Innovation in Medicine and Healthcare. Springer Nature; 2022:75-85. [doi: 10.1007/978-981-19-3440-7_7]
41. Krishnan A, Subasri V, McKeen K, et al. CyclOps: cyclical development towards operationalizing ML models for health. medRXiv. Preprint posted online on Dec 8, 2022. [doi: 10.1101/2022.12.02.22283021]
42. Kundu A, Bilgaiyan S. Automatic enhancement of deep neural networks for diagnosis of COVID-19 cases with x-ray images using mlops. In: Noor A, Saroha K, Pricop E, Sen A, Trivedi G, editors. Proceedings of Emerging Trends and Technologies on Intelligent Systems. Springer Nature; 2023:155-165. [doi: 10.1007/978-981-19-4182-5_13]
43. Meel V, Bodepudi A. Melatect: a machine learning model approach for identifying malignant melanoma in skin growths. arXiv. Preprint posted online on Sep 21, 2021. [doi: 10.48550/arXiv.2109.03310]
44. Mirza B, Li X, Lauwers K, et al. A clinical site workload prediction model with machine learning lifecycle. Healthc Anal. Nov 2023;3:100159. [doi: 10.1016/j.health.2023.100159]
45. Tougui I, Jilbab A, Mhamdi JE. Machine learning smart system for Parkinson disease classification using the voice as a biomarker. Healthc Inform Res. Jul 2022;28(3):210-221. [doi: 10.4258/hir.2022.28.3.210] [Medline: 35982595]
46. Tseng TW, Su CF, Lai F. Fast Healthcare Interoperability Resources for inpatient deterioration detection with time-series vital signs: design and implementation study. JMIR Med Inform. Oct 13, 2022;10(10):e42429. [doi: 10.2196/42429] [Medline: 36227636]
47. Ghosh A, Chaki J. Fuzzy enhanced kidney tumor detection: integrating machine learning operations for a fusion of twin transferable network and weighted ensemble machine learning classifier. IEEE Access. 2025;13:7135-7159. [doi: 10.1109/ACCESS.2025.3526272]
48. Imrie F, Cebere B, McKinney EF, van der Schaar M. Autoprognosis 2.0: democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. PLOS Digit Health. Jun 2023;2(6):e0000276. [doi: 10.1371/journal.pdig.0000276] [Medline: 37347752]
49. Lombardo G, Picone M, Mamei M, Mordonini M, Poggi A. Digital twin for continual learning in location-based services. Eng Appl Artif Intell. Jan 2024;127:107203. [doi: 10.1016/j.engappai.2023.107203]
50. Lutnick B, Ramon AJ, Ginley B, et al. Accelerating pharmaceutical R&D with a user-friendly AI system for histopathology image analysis. J Pathol Inform. 2023;14:100337. [doi: 10.1016/j.jpi.2023.100337] [Medline: 37860714]
51. Markowitz AL, Ostrow DG, Yen CY, Gai X, Cotter JA, Ji J. Machine-learning operations streamlined clinical workflows of DNA methylation-based CNS tumor classification. medRxiv. Preprint posted online on Oct 4, 2024. [doi: 10.1101/2024.01.25.24301176] [Medline: 39464257]
52. Mathew N, Joseph CT. Applying transfer learning on 3D brain images and an MLOPS study for deployment. Presented at: 2023 9th International Conference on Smart Computing and Communications (ICSCC); Aug 17-19, 2023:541-547; Kochi, Kerala, India. [doi: 10.1109/ICSCC59169.2023.10335014]
53. Moskalenko V, Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. Front Public Health. 2024;12:1342937. [doi: 10.3389/fpubh.2024.1342937] [Medline: 38601490]
54. Crawford F, Cezard G, Chappell FM. A systematic review and individual patient data meta-analysis of prognostic factors for foot ulceration in people with diabetes: the international research collaboration for the prediction of diabetic foot ulcerations (PODUS). NIHR Journals Library. 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK305598/ [Accessed 2024-05-22]
55. The Pew Charitable Trusts. How FDA regulates artificial intelligence in medical products. Pew. 2021. URL: https://pew.org/3yglbCS [Accessed 2024-05-23]
56. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (samd). US Food & Drug Administration. 2019. URL: https://www-fda-gov.login.ezproxy.library.ualberta.ca/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf [Accessed 2025-07-11]
57. Harris S, Bonnici T, Keen T, Lilaonitkul W, White MJ, Swanepoel N. Clinical deployment environments: five pillars of translational machine learning for health. Front Digit Health. 2022;4:939292. [doi: 10.3389/fdgth.2022.939292] [Medline: 36060542]
58. Moorthie S, Hayat S, Zhang Y, et al. Rapid systematic review to identify key barriers to access, linkage, and use of local authority administrative data for population health research, practice, and policy in the United Kingdom. BMC Public Health. Jun 28, 2022;22(1):1263. [doi: 10.1186/s12889-022-13187-9] [Medline: 35764951]
59. Tolera A, Firdisa D, Roba HS, Motuma A, Kitesa M, Abaerei AA. Barriers to healthcare data quality and recommendations in public health facilities in Dire Dawa city administration, eastern Ethiopia: a qualitative study. Front Digit Health. ;6. [doi: 10.3389/fdgth.2024.1261031]

60. Cabitza F, Campagner A, Soares F, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. Sep 2021;208:106288. [doi: 10.1016/j.cmpb.2021.106288] [Medline: 34352688]
61. Campagner A, Carobene A, Cabitza F. External validation of Machine Learning models for COVID-19 detection based on Complete Blood Count. Health Inf Sci Syst. Dec 2021;9(1):37. [doi: 10.1007/s13755-021-00167-3] [Medline: 34721844]
62. Eken B, Pallewatta S, Tran NK, Tosun A, Babar MA. A multivocal review of MLOps practices, challenges and open issues. arXiv. Apr 16, 2025. [doi: 10.48550/arXiv.2406.09737]
63. Minssen T, Gerke S, Aboy M, Price N, Cohen G. Regulatory responses to medical machine learning. J Law Biosci. 2020;7(1):lsaa002. [doi: 10.1093/jlb/lsaa002] [Medline: 34221415]
64. Stirbu V, Granlund T, Mikkonen T. Continuous design control for machine learning in certified medical systems. Software Qual J. Jun 2023;31(2):307-333. [doi: 10.1007/s11219-022-09601-5]
65. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health. Jun 2020;2(6):e279-e281. [doi: 10.1016/S2589-7500(20)30102-3] [Medline: 33328120]
66. Li J, Jin L, Wang Z, et al. Towards precision medicine based on a continuous deep learning optimization and ensemble approach. NPJ Digit Med. Feb 3, 2023;6(1):18. [doi: 10.1038/s41746-023-00759-1] [Medline: 36737644]
67. Pianykh OS, Langs G, Dewey M, et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology. Oct 2020;297(1):6-14. [doi: 10.1148/radiol.2020200038] [Medline: 32840473]
68. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digit Med. Jun 14, 2023;6(1):113. [doi: 10.1038/s41746-023-00858-z] [Medline: 37311802]
69. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. Dec 2022;6(12):1330-1345. [doi: 10.1038/s41551-022-00898-y]
70. Anderson D, Bjarnadottir MV, Nenova Z. Machine learning in healthcare: operational and financial impact. In: Innovative Technology at the Interface of Finance and Operations. Vol 1. Springer International Publishing; 2022:153-174. [doi: 10.1007/978-3-030-75729-8_5] ISBN: 978-3-030-75729-8
71. Callahan A, Shah NH. Machine learning in healthcare. In: Sheikh A, Cresswell KM, Wright A, Bates DW, editors. Key Advances in Clinical Informatics Academic Press. 2017:279-291. [doi: 10.1016/B978-0-12-809523-2.00019-4] ISBN: 978-0-12-809523-2
72. Sauer CM, Pucher G, Celi LA. Why federated learning will do little to overcome the deeply embedded biases in clinical medicine. Intensive Care Med. Aug 2024;50(8):1390-1392. [doi: 10.1007/s00134-024-07491-8] [Medline: 38829532]
73. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. Jan 2021;14(1):49-58. [doi: 10.1093/ckj/sfaa188] [Medline: 33564405]
74. Fohner AE, Volk KG, Woodahl EL. Democratizing precision medicine through community engagement. Clin Pharmacol Ther. Sep 2019;106(3):488-490. [doi: 10.1002/cpt.1508] [Medline: 31206610]
75. Holzer JK, Ellis L, Merritt MW. Why we need community engagement in medical research. J Investig Med. Aug 2014;62(6):851-855. [doi: 10.1097/JIM.0000000000000097] [Medline: 24979468]
76. Chi S, Tian Y, Wang F, Zhou T, Jin S, Li J. A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. Artif Intell Med. Mar 2022;125:102256. [doi: 10.1016/j.artmed.2022.102256] [Medline: 35241261]
77. Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc. Dec 1, 2019;26(12):1448-1457. [doi: 10.1093/jamia/ocz127] [Medline: 31397478]
78. Tricco AC, Lillie E, Zarin W, et al. A scoping review on the conduct and reporting of scoping reviews. BMC Med Res Methodol. Feb 9, 2016;16:15. [doi: 10.1186/s12874-016-0116-4] [Medline: 26857112]
79. Peters MDJ, Marnie C, Colquhoun H, et al. Scoping reviews: reinforcing and advancing the methodology and application. Syst Rev. Oct 8, 2021;10(1):263. [doi: 10.1186/s13643-021-01821-3] [Medline: 34625095]

Abbreviations

CL: continual learning
CM: continuous monitoring
ML: machine learning
MLOps: Machine Learning Operations
PCC: “Population,” “Concept,” and “Context”
PRISM-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by Andrew Coristine; peer-reviewed by Daniel Gartner, L Raymond Guo, Li-Hung Yao; submitted 16.09.2024; final revised version received 03.06.2025; accepted 04.06.2025; published 19.09.2025

Please cite as:
Li Y, Tian J, Xu A, Greiner R, Hayward J, Greenshaw AJ, Cao B
Maturity Framework for Operationalizing Machine Learning Applications in Health Care: Scoping Review
J Med Internet Res 2025;27:e66559
URL: https://www.jmir.org/2025/1/e66559
doi: 10.2196/66559