

Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises

Afifa Khaled^{a,*}, Mohammed Sabir^a, Rizwan Qureshi^b, Camillo Maria Caruso^c, Valerio Guarrasi^c, Suncheng Xiang^d and S Kevin Zhou^{a,*}

^aSuzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China

^bCenter for Research in Computer Vision, University of Central Florida, Florida, USA

^cResearch Unit for Computer Systems and Bioinformatics, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy

^dShanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Keywords:

MIMIC III

MIMIC IV

electronic health records (EHR)

digital health innovation.

ABSTRACT

The Medical Information Mart for Intensive Care (MIMIC) datasets have become the Kernel of Digital Health Research by providing freely accessible, deidentified records from tens of thousands of critical care admissions, enabling a broad spectrum of applications in clinical decision support, outcome prediction, and healthcare analytics. Although numerous studies and surveys have explored the predictive power and clinical utility of MIMIC based models, critical challenges in data integration, representation, and interoperability remain underexplored. This paper presents a comprehensive survey that focuses uniquely on open problems. We identify persistent issues such as data granularity, cardinality limitations, heterogeneous coding schemes, and ethical constraints that hinder the generalizability and real-time implementation of machine learning models. We highlight key progress in dimensionality reduction, temporal modelling, causal inference, and privacy preserving analytics, while also outlining promising directions including hybrid modelling, federated learning, and standardized preprocessing pipelines. By critically examining these structural limitations and their implications, this survey offers actionable insights to guide the next generation of MIMIC powered digital health innovations.

1. Overview

Rapid advancement of digital health technologies has catalyzed a paradigm shift in how clinical data is leveraged to improve patient outcomes[1], inform policy decisions[1], and drive biomedical innovation[2]. Among the many data resources that fuel this transformation, the Medical Information Mart for Intensive Care (MIMIC) datasets have emerged as the Kernel in critical care informatics, offering a unique combination of scale, granularity, and accessibility[3]. Figure 1 shows the overall framework of our survey.


1.1. MIMICs Datasets

The Medical Information Mart for Intensive Care (MIMIC) dataset is a widely recognized resource for critical care research and healthcare improvement. Developed by the Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology (MIT) in collaboration with the Beth Israel Deaconess Medical Center (BIDMC) in Boston, USA [4], MIMIC contains deidentified clinical data from over 40,000 patients in the intensive care unit (ICU). Initiated in the early 2000s, the project aimed to systematically collect and structure electronic health records (EHRs) of ICU patients, encompassing diverse data types such as physiological signals [5], laboratory results [6], medication records, and demographic information [7]. Over time,

MIMIC has expanded to incorporate data from multiple facilities and cover a broader range of medical conditions [8]. The dataset has since become a cornerstone in research areas including disease classification, outcome prediction, and personalized treatment planning, supporting advancements in both machine learning applications and clinical decision-making [9]. Later, MIMICs datasets have been developed by the MIT Laboratory for Computational Physiology, the MIMIC dataset series (MIMIC-II, III, IV, and extensions such as MIMIC-CXR and MIMIC-Waveform) encompasses rich, deidentified data from tens of thousands of admissions to intensive care units (ICU)[3], including structured physiological measurements, unstructured clinical notes[10], time series waveforms, and medical imaging[3]. These resources enable researchers to address challenges in data interoperability [2] and advance evidence informed policy [1] through scalable analytical frameworks [11]. The MIMIC datasets have significantly advanced digital health research by offering open access to richly detailed, de-identified electronic health records (EHRs) from ICU patients. First published as MIMIC-III, the data set includes more than 40,000 stays in the ICU, featuring minute-level physiological data, laboratory results, medications, procedures, and clinical notes [12]. The 2020 release of MIMIC-IV expanded this to more than 65,000 ICU admissions and 200,000 emergency visits, with enhanced data schemas and standardized coding [13].

During the past decade, MIMIC has enabled a broad spectrum of research [3], from mortality prediction and

*Corresponding author

 afifakhaled@mail.ustc.edu.cn (A. Khaled);

skevinzhou@ustc.edu.cn (S.K. Zhou)

ORCID(S): 0000-0001-5500-9940 (A. Khaled)

challenges

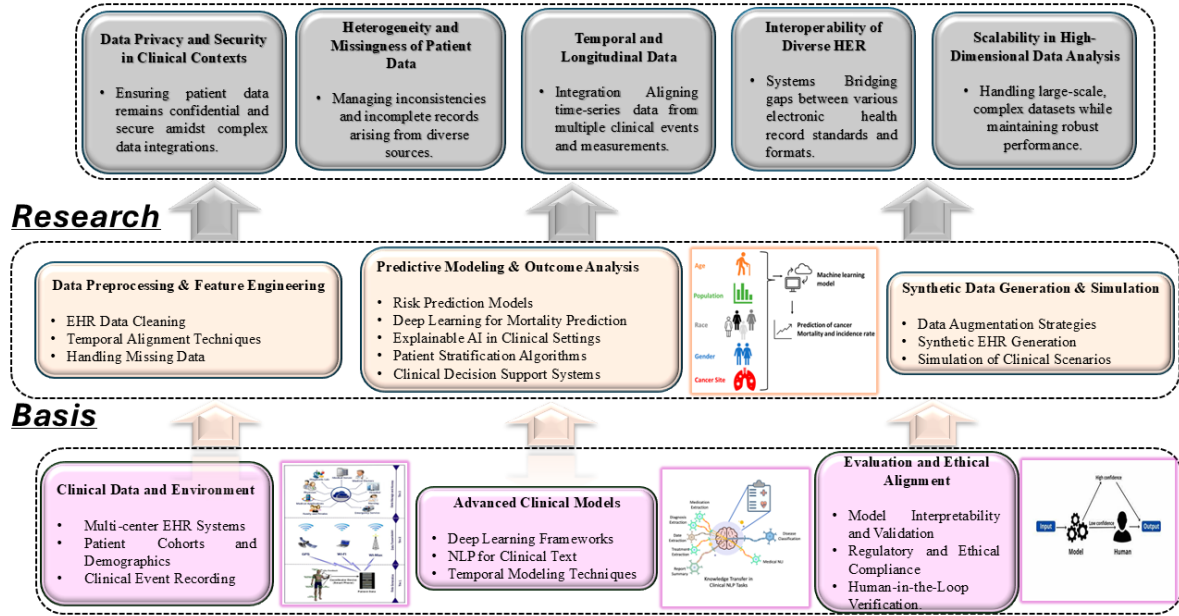


Figure 1: The overall framework of our survey.

sepsis detection [1] to clinical natural language processing [11] and epidemiological surveillance[3]. Its open access model has fostered reproducibility, standardized benchmarking, and interdisciplinary collaboration between clinical medicine, machine learning [14], and public health[1, 15]. Despite their extensive use, the increasing complexity of modern AI methods [2] and emerging healthcare needs require a comprehensive reevaluation of how these datasets are used and what novel research directions they can support[11].

MIMICs datasets comprises components that individually represent facets of the ICU experience for example of these informations:

- **Patient Details:** Includes data regarding the age range of patients and their gender distribution, along with any background information such as ethnicity.
- **Medical Records:** Comprise, in-depth documentation of diagnoses, treatments, and comments from healthcare providers.
- **Biological Indicators:** Information such as pulse rate, blood pressure measurements, and oxygen saturation levels recorded using monitoring tools.
- **Laboratory Findings:** Information regarding blood tests' outcomes and results from imaging studies and various lab measurements.
- **Medication Information:** In-depth information about the medications given while patients were in the ICU.
- **Data on Mortality and Results:** This category covers information on survival rates and the results of treatments given while in the ICU.

The information is carefully anonymized to safeguard confidentiality and enable researchers to access it in compliance with ethical guidelines like the Health Insurance Portability and Accountability Act (HIPAA) [16]. It is accessible for research purposes at no cost; however researchers are required to sign a data use agreement (DUA) and complete a course on data privacy. The recent iteration of the MIMIC III datasets is currently available via the PhysioNet platform, which also houses other health related datasets [17].

Researchers have extensively utilized the MIMIC datasets to predict ICU mortality and complications, detect early warning signs of critical conditions, personalize treatments, evaluate care effectiveness, and develop AI-driven clinical decision support systems. These efforts underscore the transformative potential of integrating clinical datasets with computational methods to enhance patient care. MIMIC continues to serve as a pivotal resource in advancing healthcare research and promoting the adoption of data science in clinical practice [17].

Previous reviews of MIMIC data sets (MIMIC-III and MIMIC-IV) have focused mainly on narrow domains, such as predictive modeling or natural language processing, without fully capturing the breadth of the evolving utility of MIMIC [18]. These surveys rarely explore the cross-domain potential of MIMIC data, such as its application in imaging, clinical decision support, and fairness-aware AI systems [19].

In addition, emerging trends in AI in healthcare care, such as temporal learning (modeling how patient data evolve over time), multimodal fusion (combining structured data, notes and imaging), explainable AI (XAI), fairness and domain adaptation, have not been systematically reviewed

in the context of MIMIC-based research [20]. There remains a critical need for a unified, up-to-date survey that bridges modern machine learning methodologies with the rich, multimodal content of MIMIC datasets (e.g. lab results, clinical notes and X-rays) [21]. Such a survey is essential to connect clinical challenges (e.g., early disease detection, risk stratification) with cutting-edge AI techniques to advance precision healthcare.

In this paper, we address these gaps by providing a structured and comprehensive review of MIMIC-based research, categorizing previous work into two broad domains.

- (1) traditional clinical and computational applications, and
- (2) novel and advanced data mining approaches.

We also present a unified taxonomy of use cases, identify persistent methodological challenges, and highlight emerging opportunities at the intersection of artificial intelligence and clinical relevance.

1.2. Recent Advancements on MIMICs Datasets

MIMICs datasets are central to health informatics and machine learning research. Numerous surveys have examined their usage patterns, identified key challenges, and proposed improvements. Below is a brief overview of significant surveys exploring various aspects of MIMICs datasets utilization.

1.2.1. Deep EHR: Advances in Deep Learning for Electronic Health Record Analysis

One of the earliest comprehensive surveys, “Deep EHR,” systematically reviewed DL applications to EHR analysis, categorizing approaches into information extraction, representation learning, outcome prediction, phenotyping, and de identification .

Over the last twenty years, the exponential growth in the quantity and complexity of EHRs has created an urgent requirement for sophisticated representation learning techniques to extract clinically significant features and enhance predictive modeling activities[22]. Deep learning (DL) methods[23], notably the authors highlighted that recurrent neural networks (RNNs) and long-short term memory (LSTM) architectures excel at capturing temporal dependencies in time series data, while autoencoders and variational methods provide compact latent embeddings that improve downstream classifier performance, feedforward neural networks, convolutional models, and recurrent networks, have proven to be highly effective in processing heterogeneous EHR information. These methods have outperformed traditional machine learning techniques in critical tasks such as predicting outcomes, identifying patient phenotypes, and modeling temporal data [13].

Subsequent evaluations built upon these foundational concepts by exploring sophisticated architectures and techniques for data augmentation. For example, Zhengping Che and his team developed *ehrGAN*, a generative adversarial network aimed at generating realistic electronic health

record (EHR) sequences for semi-supervised risk prediction, resulting in notable improvements in prediction tasks of the onset [24]. In a similar vein, unsupervised representation learning through deep autoencoders and their denoising variants has demonstrated the ability to capture underlying patient phenotypes, thereby enhancing clustering methods and early warning systems [25].

Taxonomies developed in subsequent surveys have further elucidated our understanding of DL methods for EHR. A review of MDPI by Jiabao Xu *et al.* 2022 among other researchers classified these techniques into four main types: information extraction, representation learning, medical prediction, and privacy protection. This categorization was accompanied by a benchmarking of leading architectures across well known datasets such as MIMIC, eICU, and i2b2 [26].

Furthermore, another comparative study investigated key deep learning architectures, including convolutional neural networks (CNNs), RNNs, and Transformers. The analysis revealed that convolutional models are particularly adept at capturing local temporal patterns in irregularly sampled vital signs. Conversely, attention based Transformer models provide superior interpretability by emphasizing critical time points in the data[27].

Advancements in taxonomy and architectural evaluation are crucial for improving deep learning (DL) applications in electronic health records (EHRs), enabling better alignment of techniques with clinical tasks. However, evaluations on datasets like MIMIC-III reveal challenges such as sparse and irregular data, where traditional RNNs underperform. This has led to the development of time-aware models like GRU-D, which address missing data through masking and decay mechanisms [28]. Furthermore, DL models have shown promise in improving early disease detection and personalized screening using longitudinal diagnostic and demographic data [29].

Recent comprehensive reviews have highlighted the necessity of strong benchmarking and standardized evaluation protocols. A review published in 2024 by T Hama *et. al.* and indexed in PubMed delineated specific criteria for evaluating DL Models applied to EHRs. These criteria encompass the impact of sample size, the ability to generalize findings across different healthcare institutions, and the balance between the complexity of the model and its interpretability[30]. Moreover, explainability methods such as integrated gradients and attention visualization are increasingly integrated into DL pipelines to provide clinicians with actionable insights and foster trust in AI driven tools [31].

Anticipating the future, hybrid architectures that integrate structured EHR components with unstructured clinical narratives and waveform data are expected to provide more comprehensive representations of patients. The advent of generative methodologies, such as diffusion models and hierarchical variational frameworks, presents significant opportunities for the creation of realistic clinical trajectories, which can enhance the analysis of rare event cohorts[13].

Additionally, the application of self supervised pretraining on extensive EHR datasets, similar to the pretraining of language models in NLP, has started to show improved transfer capabilities in subsequent tasks that involve limited labeled data[32].

The continued advancement of these techniques is contingent upon the establishment of standardized benchmarks that can facilitate consistent evaluation across various studies. Furthermore, cross institutional validation is essential to ensure that findings are robust and applicable in diverse clinical environments. The integration of multi modal clinical data is also crucial, as it allows for a more comprehensive understanding of patient health while adhering to principles of explainability and privacy preservation in learning frameworks.

Future of DL applications in EHR analysis will depend on the ability to navigate these challenges through collaborative efforts and the development of innovative methodologies. By addressing the complexities of data integration and model transparency, researchers can enhance the utility of DL in clinical practice, ultimately leading to improved patient outcomes and more effective healthcare delivery.

1.2.2. MIMIC IV: Benchmarking Irregular Sparse Clinical Time Series Data

The recent introduction of MIMIC IV has offered the research community an unparalleled asset for the examination of irregularly sampled and sparse clinical time series data within critical care environments[33]. Unlike its past version MIMIC III, MIMIC IV noteworthy enhancement involve improved data coverage, ICU specific modules, and both International Classification of Diseases (ICD-9 and ICD-10) codes across over 200 000 hospital admissions and 1A20 000 ICU stays [13]. It highlights the challenges in modeling such data, particularly for tasks involving imputation and sequence prediction. Despite its popularity, there has been a notable lack of standardized benchmarking on MIMIC IV, particularly with state of the art DL methods for time series analysis [23].

Motivation and Dataset Characteristics The clinical time series data in MIMIC IV exhibit irregular sampling intervals, diverse patterns of missing data, and a variety of feature types, including vital signs, laboratory results, interventions, and clinical notes. For instance, vital signs may be recorded at minute intervals, while laboratory tests are documented hourly, and textual notes are entered daily, leading to a complex mix of temporal resolutions [34]. This complexity complicates the process of representation learning, as many laboratory tests and interventions are infrequently recorded, with some appearing in less than 5% of patient encounters [25]. These properties make MIMIC IV an ideal benchmark for evaluating algorithms robustness to sparsity and irregularity.

Benchmarking Effort by Bui et al. Hung Bui and his team in 2024 tackled a significant research gap by introducing their study titled "Benchmarking with MIMIC IV, an

irregular, sparse clinical time series dataset." This extensive research evaluates both classical and DL models for time series analysis, focusing on two primary objectives:

- predicting mortality rates within the ICU and classifying patient length of stay into binary categories (short versus long stay).
- predicting mortality rates within the ICU and classifying patient length of stay into binary categories (short versus long stay)[34].

The authors implemented a standardized data processing pipeline that utilized fixed 24 hour intervals, incorporated various imputation techniques such as forward filling and interpolation, and conducted a comparative analysis of logistic regression, random forests, GRU, and different Transformer models [23]. Their findings revealed that the GRU-D model, a gated recurrent neural network designed with integrated masking and decay mechanisms for handling missing data, significantly outperformed the simpler models, achieving an (AUC-ROC 0.88 vs. 0.81). Furthermore, Transformers equipped with learned positional encodings were able to match the performance of GRU-D when they included indicators for missing data [35].

Related Benchmark Studies implications of some studies are profound, suggesting that while DL models in MIMIC III can improve predictive accuracy, they can also inadvertently perpetuate existing biases present in the data [36, 37]. The findings of Purushotham *et al.* 2017, advocate for the integration of raw time series data inputs to LSTM networks to maximize model performance, while McDermott *et al.* 2022, which evaluate MIMIC IV, revealing that demographic characteristics heavily influenced mortality predictions and underscoring the need for bias mitigation, in his study calls for a more nuanced approach that incorporates fairness and interpretability into the development of predictive algorithms[38]. Together, these studies underscore the dual necessity of advancing technical capabilities while simultaneously ensuring ethical standards in AI applications within healthcare.

In the context of emergency departments, Liu *et al.* 2022, established benchmarks for MIMIC IV-ED to predict patient admissions and resource utilization. Their findings indicated that random forest algorithms, when subjected to minimal preprocessing, achieved an impressive AUC-ROC score exceeding 0.95. In contrast, while deep models offered marginal gains at higher computational cost, they still are considered an advancement in the field of benchmarking [39]. Concurrently, Shukla and Marlin 2021, Demonstrate Transformer based imputation within benchmarking pipelines, revealed that learned interpolation significantly reduced imputation Mistakes on sparse lab series compared to mean fill or autoencoder methods [23, 40].

In the course of these investigations, several notable trends have been identified:

- (1) Models that explicitly account for missing data, such as GRU-D and neural SDE, demonstrate enhanced efficacy when dealing with irregular datasets [28, 40].
- (2) The combination of simpleques with traditional classifiers continues to serve as a solid bassolid und-soliding the need for aneed for aneed for any addition-alported by significant improvements in performance [41].
- (3) The establishment of standardized methodologies, which include uniform preprocessing, consistent data splits, and shared evaluation metrics, is essential for ensuring equitable comparisons; however, a limited number of current studies comply with these standards [35].
- (4) Despite their critical role in clinical applications, analyses concerning interpretability and fairness are frequently neglected [42].
- Benchmark multimodal fusion: evaluation of benchmark multimodal fusion techniques that integrate vital signs, laboratory results, and clinical documentation is essential. This analysis aims to determine if the contextual information provided by language can alleviate the challenges posed by sparse numerical data. Such an approach could enhance the interpretability and utility of clinical data in decision-making processes [41].

To enhance benchmarking efforts on MIMIC IV, it is essential for that community to:

- Establish open benchmark suites: that include version controlled code, datasets, and leaderboards. These resources should encompass a variety of tasks such as mortality prediction, length of stay (LOS), and readmission rates, as well as different modalities, including time series data and textual information. This comprehensive approach will facilitate a more robust evaluation of models and their performance across diverse scenarios [26].
- Develop unified missingness aware models: Develop of these comprehensive models that account for missing data, integrating hybrid RNN Transformer frameworks and Neural Stochastic Differential Equations (SDEs) to effectively address irregular timestamps and absent features. This approach aims to enhance the robustness of predictive analytics in healthcare settings by focusing on the intricacies of data absence, these models can provide more accurate insights into patient outcomes and treatment efficacy [40].
- Incorporate fairness and calibration metrics: It is essential to integrate fairness and calibration metrics in addition to the Area Under the Curve (AUC) to evaluate demographic equity and the reliability of models across various patient subgroups. This approach allows for a more comprehensive understanding of how well the model performs in diverse populations, ensuring that it does not favor one group over another. By incorporating these metrics, researchers can enhance the trustworthiness of their models, making them more applicable in real-world clinical settings [42].

By focusing on these areas, upcoming benchmarking initiatives will enhance the understanding of the strengths and weaknesses of models when applied to the irregular sparse data of MIMIC IV. This clarity will be instrumental in identifying the most effective approaches to data analysis. Furthermore, it will facilitate the development of more reliable AI tools that can be integrated into critical care settings.

1.2.3. Machine Learning for ICU Settings Using MIMIC

Machine learning (ML) has been widely used on MIMIC ICU data to predict outcomes such as mortality, length of stay, readmission, and complications [43]. Pang et al. (2022) applied models like logistic regression, SVM, and XGBoost on MIMIC-IV, achieving AUC scores up to 0.92 for in-hospital mortality prediction. Their study highlights the effectiveness of ensemble methods and the use of SHAP values for improving model interpretability [13].

Researchers using MIMIC-III have developed random forest and gradient boosting models for ICU patients with heart failure, showing improved AUC-ROC compared to SAPS II and SOFA scores, though challenged by missing data and temporal shifts [25]. A 2025 study by Chen, Fan, et al. employed MIMIC-III/IV to predict ICU readmissions in intracerebral hemorrhage patients using models like ANNs, XGBoost, and Random Forests, achieving over 85% accuracy and highlighting the importance of demographic and lab data [35]. Similarly, Purushotham et al. benchmarked LSTM and GRU-D against traditional methods for mortality and length-of-stay prediction on MIMIC-III, emphasizing the need for hyperparameter tuning [32, 36].

External validation is still quite rare: A team from The Lancet eClinicalMedicine undertook the task of validating a model that was trained using MIMIC-III data across five hospitals. Their findings revealed a performance drop of 5% to 12% in AUC, which points to the essential requirement for multicenter calibration and adaptation of models to different clinical environments. Such results underscore the need for comprehensive validation in a variety of healthcare settings [41].

Neural sequence models have revolutionized the field of dynamic survival prediction by allowing the integration of diverse data sources, including chart records, laboratory results, and output events, without necessitating manual feature extraction. This method has achieved an impressive AUROC of 0.86 in the 48-hour interval, indicating its superior predictive capabilities compared to established scoring systems such as SAPS II and OASIS. However, the

trade-off for such advanced performance is the considerable computational resources and complexity required for implementation[44].

Research on sepsis detection using MIMIC III data with XGBoost and ensemble neural networks has demonstrated AUC values that exceed 0.95 for the early identification of sepsis, severe sepsis, and septic shock, predictions made up to four hours before clinical onset. This study highlighted the critical importance of feature selection from standard vital signs to mitigate the risk of model overfitting [45]. Likewise, the ISeeU project developed a multi-scale convolutional neural network that incorporates coalitional game theory for interpretability, achieving leading edge performance in mortality prediction while also generating saliency maps for clinician use [46].

Recent studies have focused on targeted specific populations: Shuheng Chen and Junyi Fan and others 2025, created predictive models for the risk of readmission to the ICU among patients with intracerebral hemorrhage, achieving an AUROC of 0.88 for both ANN and XGBoost. They highlighted the significant impact of imputation methods on the performance of minority subgroups [47]. A real-time mortality prediction model, developed using the MIMIC III datasets and validated across datasets from Asia, America, and Europe, reached an AUC-ROC of 0.90. This model used ensemble learning techniques with only eight input variables, showcasing the feasibility of lightweight and easily deployable predictive tools [48].

Meta analyses conducted by Bellamy and colleagues 2020, revealed that deep RNNs exhibit only a slight advantage over logistic regression in MIMIC tasks. This finding underscores the need for more thorough benchmarking and the consideration of simpler models when appropriate [49]. The implications of this research highlight the need for a balanced approach in the application of ML techniques in ICU environments, which are more promising for the persistent challenges of missing data, domain shift, interpretability, and reproducibility.

Expanding upon the significant advancements made in the utilization of machine learning for ICU data, there are several important pathways that warrant additional investigation:

- **Federated and Privacy Preserving Learning:**

In order to overcome obstacles related to data sharing and to improve the generalizability of models across various institutions, it is essential to expand federated learning frameworks, such as FLICU. This expansion should focus on integrating diverse data modalities from intensive care units (ICUs) while also enhancing the efficiency of communication and ensuring robust privacy protections. Such advancements are crucial for fostering collaboration and innovation in health-care research and practice [46, 42].

- **Self Supervised and Transfer Learning:**

Training deep sequence models on extensive unlabeled EHR datasets, such as MIMIC III, MIMIC IV,

and eICU, can produce generalized universal clinical representations. These representations are capable of being effectively applied to subsequent downstream prediction tasks that have minimal limited labeling requirements. This approach not only alleviates the burden of annotation but also enhances the overall robustness of the models [32].

- **Continual and Online Learning:**

The data collected in ICUs undergoes evolving transformation over time as a result of modifications in clinical protocols and adjustments in device calibrations. It is essential to create continual learning algorithms that can adapt in real time to these changes. Such algorithms must also overcome the challenge of catastrophic forgetting to ensure their effectiveness in ever evolving dynamic clinical settings[44, 40].

- **Multi Modal Fusion and Contextual Modeling:**

Future models ought to integrate time series vital signs, laboratory results, medication prescriptions, and clinical documentation through hierarchical or graph based frameworks. This integration and fusion has the potential to elucidate cross modal interactions, such as instances where clinical notes indicate deterioration prior to observable changes in vital signs, thereby enhancing early warning systems [30, 27].

- **Causal and Counterfactual Analysis:**

Integrating causal inference methodologies, such as directed acyclic graphs or counterfactual reasoning, into ML workflows can aid in differentiating between misleading correlations and genuine actionable risk factors. This integration enhances the reliability of decision support systems utilized in the ICU [49, 35]. Such advancements are crucial for improving patient outcomes and ensuring that clinical decisions are based on sound evidence rather than mere statistical associations.

- **Fine Grained Outcome Prediction and Personalization:**

Advancing past the simplistic binary classifications of mortality and readmission, a more detailed prediction of organ specific complications, such as acute kidney injury, the success of ventilator weaning, and neurological outcomes, necessitates the development of specialized submodels. These submodels must be tailored to account for the unique characteristics of individual patients, thereby enhancing the precision of risk assessments. This approach underscores the importance of personalized risk profiles in improving patient outcomes and guiding clinical decision making [30, 41].

- **Explainability and Trustworthiness:**

The integration of interpretability techniques, including SHAP values tailored for time series analysis and

attention based saliency maps, is crucial for gaining the acceptance of clinicians. It is imperative that subsequent research thoroughly assesses the fidelity of these explanations and their influence on user trust within future clinical studies [42, 29].

- **Benchmarking and Open Leaderboards:**

The implementation of standardized and open benchmark suites, which include versioned code, common data splits, and a variety of tasks, will promote equitable comparisons and enable reproducible advancements in research. The establishment of public leaderboards for MIMIC IV tasks related to mortality, length of stay, and readmission tasks can significantly enhance community involvement in these areas. This approach not only fosters collaboration but also encourages transparency in the evaluation of different methodologies and their outcomes [26, 23].

- **Robustness to Distribution Shift:**

It is essential for research to consider the effects of covariate and concept drift within diverse patient populations, variations in medical devices, and alterations in care protocols. The implementation of domain adaptation strategies, alongside techniques for quantifying uncertainty such as Bayesian deep learning, will play a pivotal role in developing machine learning models for intensive care units that are both safe and generalizable. This approach is supported by existing literature that emphasizes the importance of these methodologies in ensuring the reliability of predictive models in critical healthcare settings [41, 28].

- **Integration with Clinical Workflows:**

Ultimately, the integration of ML models into ICU monitoring systems and EHR, bolstered by human in the loop feedback mechanisms and real time alert dashboards, necessitates collaborative design efforts with healthcare professionals. This approach will also demand thorough prospective trials to evaluate the effects on patient outcomes and the efficiency of clinical workflows [48, 13].

1.2.4. Imitation Learning and MIMIC Data

Imitation learning (IL), often referred to as learning from demonstrations, has surfaced as a viable method for training decision making agents by emulating expert actions instead of depending exclusively on defined reward functions [38]. Within ICU, IL provides a means to encapsulate the intricate, making sequential characteristics of clinical decision processes such as fluid management, vasopressor titration, and ventilator adjustments derived directly from historical EHR data in MIMIC [50].

Imitation Learning studies in MIMIC A significant investigation conducted by Komorowski *et al.* 2018, utilized inverse reinforcement learning (IRL) on MIMIC III data to uncover the hidden reward frameworks that inform clinicians

approaches to sepsis treatment policies. By conceptualizing vasopressor and fluid administration as a Markov decision process, the researchers illustrated that the policies derived from their model could surpass conventional care protocols in simulated environments, leading to a decrease in anticipated mortality rates and variability in treatment outcomes [51].

In later research conducted by Raghu among other researchers 2017, the issue of batch offline IL was rigorously defined using MIMIC data. They implemented behavior cloning techniques alongside doubly robust propensity weighted estimators. This approach aimed to develop safe policies that closely resemble expert performance, even in the presence of distributional shifts [52].

Clinical Decision Support Implementations In light of these foundational elements, research has incorporated policies derived from IL into prototypes for clinical decision support. Prasad and co workers 2019, established a hybrid framework that combines IL with reinforcement learning, where an imitation learned policy serves as an initial warm up phase. This is succeeded by a constrained policy optimization process aimed at fine tuning actions while adhering to specific safety constraints. When assessed using retrospective data from MIMIC III sepsis cohorts, this methodology demonstrated enhanced sample efficiency and compliance with clinical safety parameters in comparison to traditional reinforcement learning agents [53].

Recent advancements have broadened the application of IL to encompass multi task interventions in ICU. This includes integrated approaches to fluid resuscitation and vasopressor management. The methodology involves modeling clinician trajectories pathways across various axes through the use of hierarchical Gaussian mixture models [54].

Despite these successes, IL on MIMIC faces several obstacles:

- **Confounding Bias:** Clinical decisions often depend on unobserved patient variables, leading to biased policy evaluations if these confounders are not properly accounted for [55].
- **High-Dimensional & Incomplete States:** The high-dimensional and partially observed nature of EHR data (e.g., lab values, vital signs, clinical notes) complicates behavior cloning and inverse reinforcement learning (IRL), increasing the risk of learning from spurious correlations rather than clinically meaningful patterns [56].
- **Safety and Interpretability:** Learned policies must be auditable and operate within clinically acceptable boundaries to ensure safe deployment and preserve transparency in clinical decision-making [57].
- **Dynamic Patient Trajectories:** Modeling patient trajectories is inherently challenging due to the heterogeneous and evolving nature of disease progression and

treatment response, requiring sophisticated temporal modeling [38].

- **Lack of Formal Safety Guarantees:** Many current imitation learning (IL) frameworks lack formal mechanisms for safety assurance, raising concerns about their reliability in critical care environments [57].

Recent Advances To address confounding variables in clinical decision data, counterfactual imitation learning (IL) approaches have been proposed. These methods leverage causal inference techniques such as inverse propensity scoring and doubly robust estimators to adjust for treatment biases inherent in retrospective datasets [58].

Additionally, model-based IL incorporates learned representations of patient physiology such as continuous-time neural ODEs to simulate alternative action outcomes and optimize policies through planning [59]. This enables predictive modeling of treatment effects and supports more informed decision-making.

For interpretability, attention-based IL architectures have been developed to highlight patient features that most influence recommended actions, thereby promoting clinician oversight and building trust in AI-assisted care [60]. These interpretable models enhance the transparency of clinical trajectories, improving both system reliability and the quality of care.

Looking ahead through the future, the field of Imitation Learning and MIMIC Data should pursue

- **Causal Imitation Learning:** Incorporate clear causal frameworks that elucidate the relationships between treatment and outcomes, thereby allowing for the separation of confounding factors. This approach aims to enhance the generalizability of policies across various patient subgroups [61]. By systematically addressing these causal models, we can better understand the complexities involved in treatment efficacy and its implications for diverse populations.
- **Safe Policy Deployment:** by Establishing certified safety mechanisms, including shielding or constrained model predictive control, that uphold clinical limitations, such as the maximum allowable vasopressor rates, throughout the execution of policies [53]. This approach ensures that safety protocols are integrated into the operational framework, thereby enhancing patient care. By adhering to these clinical constraints, healthcare providers can mitigate risks associated with medication administration and improve overall treatment outcomes.
- **Multi Modal Demonstrations:** Integrate unstructured data, such as clinical notes and radiological images, into Information Literacy frameworks through the use of multi modal encoders. This approach facilitates the acquisition of a more comprehensive context for critical decision points [60]. The incorporation of

diverse data types enhances the overall effectiveness of the frameworks, allowing for improved insights and outcomes in clinical settings.

- **Online Adaptation and Personalization:** combine IL with online digital education can facilitate the adaptation of policies to contemporary clinical environments and the unique pathways of individual patients. This approach is essential in ensuring that healthcare practices remain aligned with expert standards, thereby mitigating the risk of significant catastrophic deviations from established professional behavior [54]. By leveraging technology, healthcare providers can enhance their responsiveness to the evolving needs of patients while maintaining high quality care.
- **Prospective Clinical Validation:** addressing the shortcomings identified in retrospective evaluations, it is essential to integrate IL decision support systems derived from artificial intelligence into existing clinical workflows. This integration will facilitate a more streamlined approach to patient care, allowing healthcare professionals to make informed decisions based on real time data. Furthermore, the implementation of these systems should be accompanied by rigorous randomized controlled trials to assess their effectiveness in improving patient outcomes and ensuring safety [62].

1.3. Taxonomies and Classification Frameworks for Healthcare Datasets

The proliferation and increasing availability of healthcare datasets underscores the urgent need for systematic taxonomies and classification frameworks that can effectively organize heterogeneous data sources and facilitate reproducible research. These taxonomies help categorize datasets by modality (e.g., structured EHR, free text, imaging), scope (single-center vs. multi-center), and level of curation (raw vs. harmonized), thereby guiding researchers in selecting appropriate datasets for specific analytical tasks [29].

Syeda *et al.* (2021) explored the structure and organization of healthcare datasets particularly the MIMIC dataset in the context of machine learning applications. Their work emphasizes the importance of formal classification systems and data integration frameworks to manage the growing complexity of clinical data. MIMIC, in particular, is highlighted as a benchmark resource for studying healthcare data standardization and architectural design [63].

General Healthcare Data Taxonomies Early efforts to classify biomedical datasets differentiated between primary data repositories and derived knowledge bases. Primary repositories such as MIMIC, which contains structured ICU data, and TCGA, which focuses on cancer genomics serve as foundational resources for research. In contrast, derived knowledge bases, including biomedical ontologies and phenotyping algorithms, provide interpretive structures to contextualize and enrich the raw data [41].

Subsequent surveys have proposed more nuanced classification frameworks based on three orthogonal axes:

- **Data Modality:** This axis categorizes data into structured, unstructured, and semi-structured formats. Structured data (e.g., laboratory tables, diagnosis codes) are easily quantifiable, whereas unstructured data (e.g., clinical narratives, radiology reports) present challenges in preprocessing. Semi-structured formats, such as HL7 messages and FHIR profiles, bridge the gap and facilitate interoperability within healthcare systems [30].
- **Data Origin:** Datasets may stem from a single institution or multiple federated centers. For instance, MIMIC-III (sourced from BIDMC) represents a single-center dataset allowing for deep, focused analysis. In contrast, multi-center datasets like eICU or OHDSI aggregate data across institutions, promoting generalizability but requiring sophisticated integration techniques [13].
- **Data Processing Level:** This dimension spans from raw data extracts to standardized common data models (e.g., OMOP-CDM, FHIR) and analytic-ready cohorts that include phenotype-labeled subsets. These tiers offer varying degrees of data usability, with each suited to different stages of the research pipeline [34, 23].

This taxonomy in mentioned significantly enhances the discovery of datasets, the evaluation of interoperability, and the benchmarking of methodologies by elucidating the strengths and limitations of each resource. This clarity is essential for researchers and practitioners who aim to effectively utilize various datasets. By establishing a structured framework, this taxonomy facilitates informed decision making regarding the selection and application of resources across diverse research contexts [40].

Classification Frameworks for EHR Datasets Focusing on EHR derived data, several classification frameworks have emerged:

1. OHDSI/OMOP-CDM: Observational Health Data Sciences and Informatics (OHDSI) / OMOP-CDM serves to standardize relational schemas and vocabularies, including Systematized Nomenclature of Medicine (SNOMED CT), the normalizing the names of prescription and over the counter drugs (RxNorm), and Logical Observation Identifiers Names and Codes (LOINC). This standardization facilitates the execution of multi center analyses across various EHR systems. Such a framework is essential for ensuring consistency and comparability in data derived from different sources, thereby enhancing the reliability of research outcomes [27].
2. FHIR based Profiles: HL7 FHIR specification categorizes information into modular components known

as "resources", such as Patient, Encounter, and Observation. These resources serve as the foundational building blocks for health information exchange. Furthermore, implementation guides, including US Core and ICU Core, delineate context specific profiles that ensure uniformity in data exchange across various healthcare settings [41].

3. Phenotype Libraries: Curated definitions, such as those found in Phenotype KnowledgeBase(PheKB), categorize datasets according to specific clinical concepts, including conditions like diabetes and sepsis. This classification system is instrumental in linking various code lists and extraction queries, which facilitates the generation of reproducible cohorts. The ability to systematically organize data in this manner enhances the reliability and validity of clinical research outcomes [29].

MIMIC in the Taxonomy within the context of these frameworks, MIMIC datasets represent serve exemplary as instances archetypes of extensively richly annotated EHR data from a single center ICU. The CHARTEVENTS, LABEVENTS, and NOTEVENTS tables within MIMIC III illustrate the various data modalities, including structured, semi structured, and unstructured formats. This diversity in data representation positions MIMIC III as a valuable reference point for versatile benchmark the advancement of methodological approaches in healthcare research [34].

MIMIC IV datasets enhances the existing classification system by breaking down data related to ICU, Emergency Departments (ED), and hospital admissions into tables that are compatible with the FHIR standard. This modular approach allows for more efficient data management and retrieval, facilitating research and clinical applications. Furthermore, the integration of an OMOP-CDM export provides a standardized format for data sharing and analysis, promoting interoperability across different healthcare systems [23].

Despite the existence of standardized formats, considerable heterogeneity variability remains evident . The local item identifiers found in CHARTEVENTS necessitate a manual process for mapping to LOINC or SNOMED. Furthermore, the structure and content of free text notes differ significantly among various institutions [29]. Moreover, phenotype definitions often lack portability, as code lists may not generalize beyond the origin site's coding practices [30]. Furthermore, the definitions of phenotypes frequently exhibit a lack of portability, as the coding lists may not be applicable outside the specific coding practices of the originating site. This limitation poses challenges for researchers who aim to apply these definitions in diverse contexts. Consequently, the generalizability of phenotype definitions is often compromised, hindering broader scientific collaboration and data sharing [30]. Multi-modal fusion frameworks must reconcile disparate temporal resolutions

and missingness patterns inherent to structured versus unstructured data [40]. Multi modal fusion frameworks are must reconcile disparate to address the challenges posed by varying temporal resolutions and the patterns of missing data that are inherent to structured versus unstructured datasets. This reconciliation is essential for achieving effective integration of diverse data types, which often exhibit significant differences in their temporal characteristics. The ability to harmonize these discrepancies is crucial for the successful application of multi modal fusion techniques in various domains.

To improve dataset taxonomy and classification, the next work should

- **Develop Unified Metadata Standards:** Enhance schema registries by incorporating elements such as provenance, curation status, and quality metrics. This enhancement will facilitate the automated discovery of datasets and the assessment of their suitability for various applications. Such improvements are essential for advancing data management practices in contemporary research environments [41].
- **Automate Code Mapping:** Utilize machine learning techniques to accurately map local EHR codes, such as the ITEM-IDs from the MIMIC datasets which provides microdata in the form of a unique, global identifier of an item, to standardized vocabularies like LOINC and SNOMED. This approach aims to enhance the precision and recall of the mapping process, ensuring that the local codes are effectively translated into widely accepted terminologies. By employing advanced algorithms, the alignment of these codes can be achieved with a high degree of reliability, facilitating better data interoperability and integration across healthcare systems [25].
- **Cross Model Validation:** Establish validation frameworks that assess OMOP and FHIR exports in comparison to original schema extracts, aiming to identify semantic inconsistencies and potential data loss. This approach is crucial for ensuring the integrity of data as it transitions between formats. By rigorously testing these exports, researchers can uncover hidden issues that may compromise the quality of the data being utilized in various applications [27].
- **Dynamic Phenotype Repositories:** Establish dynamic phenotype libraries that adapt their coding frameworks and extraction methodologies in response to community input and the progression of coding standards. This approach emphasizes the importance of collaboration and continuous improvement in the field of bioinformatics. By integrating feedback from users, these libraries can remain relevant and effective in addressing the needs of researchers and practitioners alike [29].

- **Multi Modal Dataset Catalogs:** Build centralized repositories that categorize datasets based on their modality combinations, data quality, and CMI (Complexity Modality Integration) scores to facilitate the selection of appropriate methodologies. This approach aims to streamline the process of identifying suitable datasets for various research purposes. By indexing datasets in this manner, researchers can make informed decisions that enhance the quality and relevance of their work [40].

By advancing these taxonomy and classification initiatives, we can streamline dataset interoperability, reduce duplicative ETL efforts, and accelerate the development of robust, reproducible machine learning tools in critical care and beyond.

By promoting the evolution of taxonomy and classification efforts, we can greatly improve the interoperability of datasets. This improvement leads to a reduction in repetitive ETL (Extract, transform, and load) tasks, thereby optimizing resource allocation and time management. As a result, the creation of effective and reproducible machine learning applications in critical care and other domains can be accelerated.

2. Open Problems

MIMIC data sets have become increasingly vital in intensive care research, offering a rich collection of deidentified health data from ICU patients [9]. Although these data sets enable advanced research in predictive modeling [64], clinical decision making [43], and patient trajectory analysis [65], they also present several inherent challenges.

Identifying open problems within MIMIC datasets requires a deep understanding of their structure, as well as a awareness of common issues such as data incompleteness [66], forecasting limitations [67], and ethical handling of sensitive health information [68]. These challenges not only hinder the performance of machine learning models but also raise concerns about reproducibility [18], generalizability [69], and clinical applicability [70].

Figure 2 illustrates the main categories of open problems discussed in this survey. In the following, we explore several of these key challenges in greater detail.

2.1. Challenges in Data Granularity, Cardinality, and Real Time Applications

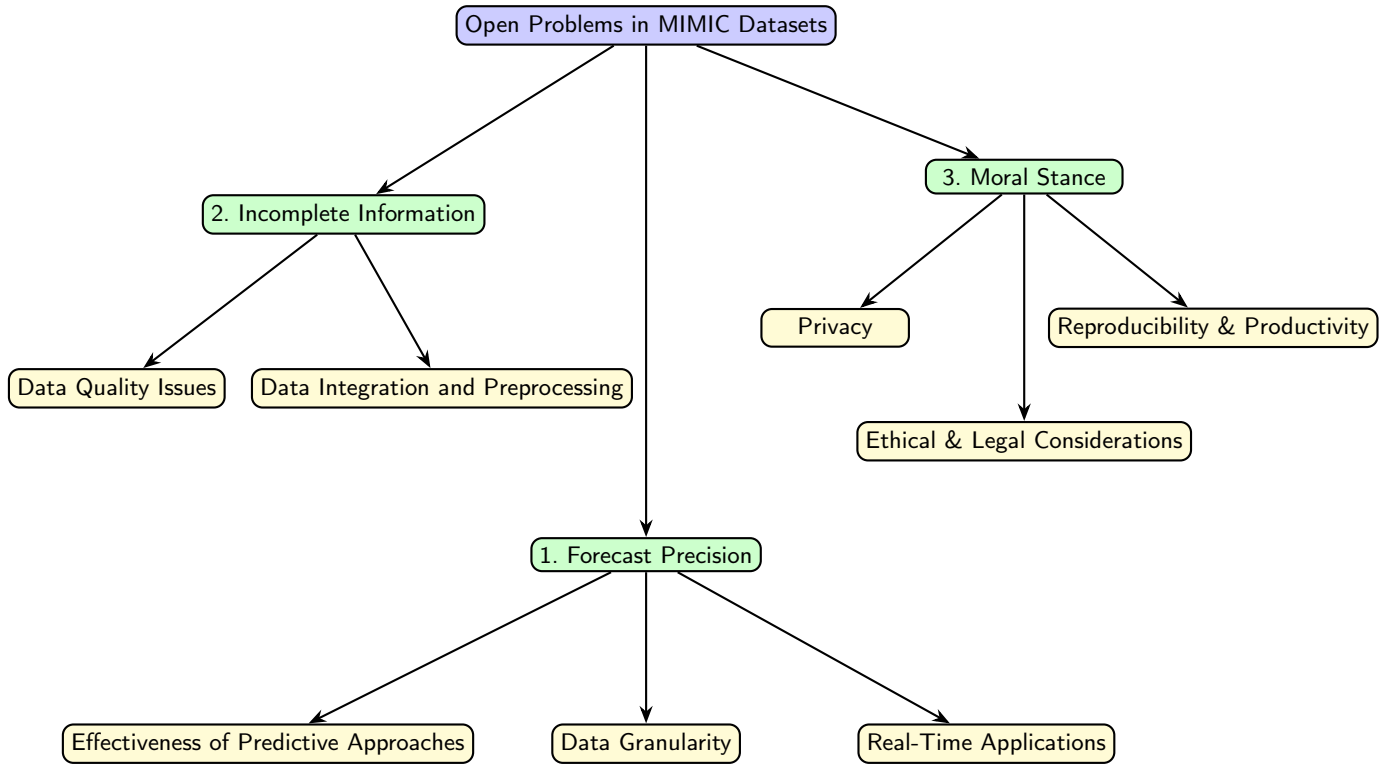
Granularity refers to the level of detail or precision in data representation. High granularity indicates fine-grained data (e.g., minute level clinical observations), whereas low granularity refers to aggregated or generalized information.

Cardinality denotes the number of unique entities or values within a dataset. In MIMIC datasets, cardinality is constrained by the relatively small number of ICU admissions, reducing patient case diversity and hindering machine learning generalizability, especially for deep architectures requiring large samples [69].

Table 1

Comparative Analysis of MIMIC Dataset Surveys, which compares key criteria across different surveys analyzing MIMIC datasets. Checkmark (✓) indicates coverage of each criterion.

Criteria	Deep EHR	MIMIC-IV Benchmark	ML for ICU	Imitation Learning	Dataset Taxonomy	Our Survey
Focus on Deep Learning	✓	✓	✓	✓	✓	✓
Time-Series Analysis	✓	✓	✓	✓	✓	✓
Clinical Decision Support	✓	✓	✓	✓	✓	✓
Data Standardization	✓	✓	✓	✓	✓	✓
Implementation Status	Successful	Partially Successful	Successful	Partially Successful	At Planning stage	At Planning stage

**Figure 2:** Open Problems in MIMIC Datasets.

Despite containing rich clinical information, MIMIC III includes approximately 40,000 ICU stays as it mentioned before, which is small relative to the high dimensional feature space of physiological signals, medications, lab results, and notes [71, 72].

High granularity introduces computational difficulties curse of dimensionality, overfitting, poor generalization, and excessive cost unless effective preprocessing or dimensionality reduction is applied [73].

The quantity of patient records, which amounts to tens of thousands, is comparatively small when juxtaposed with the expansive feature space. This disparity intensifies the phenomenon known as the 'curse of dimensionality' [74]. The implications of this limitation are significant, as they can hinder the effectiveness of various analytical methods that rely on a robust dataset to yield reliable insights.

EHRs in critical care settings are characterized by their high dimensionality, as a single stay in the ICU can yield a multitude of data points, including vital signs, laboratory results, waveform data, imaging studies, and clinical notes [75]. This extensive array of measurements presents a complex landscape for data analysis. The challenge lies in effectively managing and interpreting this vast amount of information to enhance patient care and outcomes.

Numerous variables are often sampled infrequently or exhibit irregular temporal patterns. Certain critical signals, such as continuous waveforms, may be subject to down sampling or may not be accessible at all, which blurring or complicates the detection of swift rapid clinical changes [76].

In parallel, coded fields, including ICD diagnoses and laboratory designations, demonstrate significant cardinality,

resulting in long tail distributions that hinder the model's ability to learn from infrequent occurrences [74, 77].

Additionally, it obstructs the ability to perform real time analytics, as issues such as missing values, inconsistent time scales, and class imbalance can significantly delay online processing. Consequently, these challenges frequently necessitate the use of heuristic methods for aggregation or imputation, which can result in the loss of valuable information, as highlighted by Hawryluk (2020) [78].

Ghassemi et al. 2020, mentioned the Uncommon conditions that are represented by a limited number of patients may fail to yield adequate examples, which could adversely affect the sensitivity to significant events. This limitation is particularly concerning in the context of clinical research, where the ability to detect critical occurrences is paramount[79]. The scarcity of data can lead to a lack of robustness in findings, ultimately impacting patient care and treatment outcomes.

However, recent research has proposed new strategies to address these problems. through sophisticated representation learning techniques as Potential of feature engineering, dimensionality reduction, and transfer learning techniques in enhancing model performance while working with limited data has been explored Deep autoencoders, along with variational embeddings, effectively condense ICU time series data into lower dimensional forms while maintaining their clinical significance [80, 81]. In addition, the recent developments in the structures of recurrent neural networks and combined modeling strategies emerge as possible ways of enhancing the predictive ability in clinical practice [82].

also to mitigate granularity and cardinality challenges, feature engineering selective extraction and aggregation (e.g., hourly/daily intervals) improves interpretability and predictive performance [71]. Dimensionality reduction (PCA, autoencoders) compresses high dimensional data into lower dimensional representations, easing computational burden and enhancing efficiency [83].

Another approach to managing granularity and cardinality challenges is the use of less granular encoding schemes, such as the Clinical Classifications Software (CCS). This system aggregates the 15,072 ICD-9-CM diagnostic codes into 285 broader categories, significantly reducing the feature space and enhancing model generalizability [72]. While effective in addressing sparsity and improving computational efficiency, this dimensionality reduction comes at the expense of diagnostic specificity, potentially limiting the clinical granularity and interpretability of prediction outcomes.

Transfer learning from large-scale repositories (e.g., eICU, PhysioNet) fine tuned for MIMIC tasks has boosted mortality prediction and sepsis detection [84].

Deep time series models (e.g., bidirectional MGRUs) capture complex temporal patterns, offering a balance between complexity and performance [85, 65].

Integrating domain knowledge (expert defined guidelines) into pipelines improves interpretability and informed feature selection without sacrificing predictive power [86].

Table 2
High Cardinality Features in MIMIC

Attribute	Value
Variable	ICD-9 Diagnosis Codes
Dataset	MIMIC III
Unique Values	14,567
Variable	NDC Medication Codes
Dataset	MIMIC III
Unique Values	4,478
Variable	CPT Procedure Codes
Dataset	MIMIC IV
Unique Values	5,221
Variable	HADM_ID (Hospital Admissions)
Dataset	MIMIC-III
Unique Values	58,976

Table 3
Temporal Granularity in MIMIC

Attribute	Value
Dataset	MIMIC III ICU Chartevents
Temporal Granularity	Minute level (1 min intervals)
Record Count	330M+ rows
Dataset	MIMIC IV Lab Measurements
Temporal Granularity	Hourly to daily
Record Count	28M+ rows
Dataset	MIMIC IV ED Stays
Temporal Granularity	Second level timestamps
Record Count	425K+ events

Multi institutional collaborations combining MIMIC III with eICU or broader EHRs expand patient diversity, enhancing robustness and external validity [87].

Despite advances, real time clinical deployment faces computational and black-box challenges. The AI Clinician reinforcement learning model optimizes sepsis care strategies, exemplifying AI's potential for real time precision interventions [88].

For such models to gain trust and utility in clinical workflows, explainability and transparency are paramount. Recent work in explainable AI (XAI) highlights the need for interpretable decision logic to support clinician acceptance and ensure responsible deployment [89].

Neural sequence models, including LSTM networks and GRUs, as well as Transformers, are adept at extracting essential temporal characteristics from sparse EHR data. This capability enhances their resilience against data incompleteness [90, 91]. The ability to effectively manage missing data is crucial in healthcare analytics, where data gaps are common and can significantly impact model performance.

Feature selection techniques, particularly feature hashing, work synergistically with computational geometry to isolate informative variables that are crucial for analysis.

Table 4
Key Features and Temporal Granularity in MIMIC Datasets

Category	Variable Dataset	/	Description	Values / Records
	ICD 9 (MIMIC III)	Codes	Diagnosis Codes	14,567
	NDC (MIMIC III)	Codes	Medication Identifiers	4,478
	CPT (MIMIC IV)	Codes	Procedure Codes	5,221
	HADM_ID (MIMIC III)		Hospital Admissions	58,976
	ICU Chartevents (MIMIC III)		Minute level granularity	330M+
	Lab Measurements (MIMIC IV)		Hourly to daily	28M+
	ED Stays (MIMIC IV)		Second-level timestamps	425K+

This approach not only aids in the identification of key data points but also significantly minimizes the presence of redundant information [92, 93]. Consequently, the data becomes more manageable, and the insights derived from it become clearer and more actionable.

Utilizing hierarchical autoregressive models allows researchers to capture the complex dependencies present within the data, which is vital for accurately representing the intricacies of actual health records. By employing these models, it becomes feasible to create synthetic datasets that preserve the statistical properties of real EHRs, thus providing a valuable resource for training and validating predictive models [94]. This method not only addresses the challenges associated with data scarcity but also promotes ethical considerations by minimizing the need for sensitive patient information.

Improving effective granularity and addressing high cardinality continue to pose significant challenges. Future methodologies may involve the use of context sensitive imputers or generative models designed for handling missing data [93, 95].

Additionally, there is potential for the development of sophisticated sparse coding techniques, hybrid models that integrate structured EHR data with unstructured text and waveform data, as well as adaptive continual learning frameworks that can identify instances where data scarcity undermines the reliability of predictions [78].

Future work should also aim to build hybrid models that combine the best of both worlds, incorporating domain knowledge with data driven learning to improve the interpretability and accuracy of the models. Furthermore, encouraging collaborative efforts to increase the size of the datasets. For example, incorporating MIMIC III with other clinical archives can be helpful in managing cardinality issues. In addition, solving the computational and interpretability issues

that are inherent to real time applications will be critical for the effective implementation of AI based solutions in clinical practice [96].

To further contextualize these challenges, Table 2 summarizes high cardinality features in MIMIC, while Table 3 highlights its temporal resolution across datasets, and Table 4 show the key features and temporal granularity in MIMIC datasets.

2.2. Effectiveness of Predictive Approaches

MIMIC III dataset has become a cornerstone for developing prognostic models in critical care, particularly for outcome prediction tasks. Effective prognostic modeling requires processing both diagnostic data and temporal clinical events. Despite its rich clinical detail, MIMIC III presents ongoing challenges due to its complexity and heterogeneity [9].

Issues such as class imbalance, feature redundancy, and the dynamic nature of temporal data often hinder model performance in high stakes applications like mortality prediction and sepsis detection [69].

A wide range of ML & DL techniques have been employed to improve the performance and accuracy of predictive models using MIMIC data. Recent work by Khope and Elias 2023, categorizes these approaches into three main groups: classification based, prediction based, and early prediction models [97].

Categorization based methods leverage ICD codes to classify patient conditions, with attention mechanisms improving interpretability by focusing on clinically relevant features [98]. Prediction based models, including RNNs and LSTMs, have been widely applied to tasks such as mortality prediction and length of stay estimation.

The integration of domain specific features from clinical guidelines has also been suggested for enhancing both the interpretability and the accuracy of the model itself [72].

A significant contribution to the field is the work by [99], which introduced counterfactual reasoning as a critical component of decision support systems especially in high risk domains like healthcare. This approach uses hypothetical scenarios to improve the realism of predictive models and decision making frameworks.

They incorporate causal Bayesian networks (CBNs) for causal relationship modeling techniques to manage counterfactual risks, and machine learning to improve the robustness of causal inference [99].

A similar work by Zhang and co workers 2020, [100] also focuses on the challenges of causal inference in observational research, especially in the context of complex healthcare datasets like MIMIC III.

Khope and Elias (2023), stated that Early models employing random forests and multivariate logistic regression were developed for real time alerts of adverse clinical events. However, their effectiveness has been limited by challenges related to data heterogeneity and missing values [97]. they categorizes approaches into three groups: classification based, prediction based, and early prediction

models [97]. Additionally, ensemble learning and attention based architectures have been employed to enhance feature selection and improve model robustness and predictive accuracy [101, 72].

Unmeasured confounding is still a big problem in causal inference, and this requires the use of sophisticated techniques including instrumental variable analysis and Causal Directed Acyclic Graphs (DAGs). Instrumental variable analysis uses external proxies, e.g., a randomized hospital drug policy as a proxy for the true effect of an intervention, while causal DAGs help identify potential biases through the graph of variable relationships [100]. Unmeasured confounding, which happens when an analysis overlooks a variable (or set of factors) that influences both the treatment, exposure and the outcome, is a significant problem with causal inference. Since this variable is unknown or unmeasured, it could lead to distorted estimates of the causal influence.

Causal inference techniques such as inverse probability weighting and instrumental variable analysis have been applied to help adjusting the confounders and yield more robust predictors [102].

Developing accurate predictive models from MIMIC data is challenging due to class imbalance, cohort bias, and overfitting [103].

Clinical outcomes such as mortality or complications are often rare events, while controls dominate the dataset, causing standard classifiers to prioritize specificity over sensitivity [104].

Additionally, MIMIC's single center origin and evolving practice patterns over its 15 year span introduce cohort bias: models may implicitly learn hospital specific protocols rather than generalizable clinical features. As a result, a model that predicts risk well on the original dataset may fail to generalize to new cohorts (i.e., poor external validity) [105].

This raises concerns for reproducibility and fairness, a model with high in sample accuracy but low transferability is unreliable for broader use. Retrospective reviews have highlighted common pitfalls: data leakage where future information inadvertently informs model inputs can inflate reported performance, and a lack of rigorous cross validation across patient admissions undermines external validity [77].

Moreover, many high performing MIMIC trained models (AUC>0.90) fail to replicate on independent datasets, revealing deficiencies in generalizability and raising concerns about fairness in deployment [79].

Moreover, shifting clinical practices over time (e.g. new treatments) or differences in coding standards can render static models obsolete. Interpretability is another issue: complex models (deep nets, ensembles) often behave as black boxes, making it hard to ensure clinical validity or to discover bias. Recent efforts have aimed to improve predictive reliability. For example, training on multi center datasets or using federated learning can increase generalizability [106].

Fairness aware algorithms (e.g. adversarial debiasing, reweighting schemes) have demonstrated reductions in demographic disparities in risk estimates [107].

calibration methods and continuous monitoring are increasingly recommended to detect and correct model drift over time [108].

These methodologies go further to incorporate beyond the mathematical frameworks. They capture the limitations of the observational data and demand the need for stronger inferential techniques for healthcare analytics. The bias in estimating treatment effects from unmeasured confounders can lead to wrong clinical policies, and thus, there is a need for better methods to deal with these issues.

Despite these advances, ensuring true reliability and equity in clinical settings remains an active research frontier. Future work should prioritize:

- **Prospective External Validation:** Systematic evaluation of MIMIC trained models on geographically and temporally distinct cohorts to assess transportability [105].
- **Causal Model Integration:** Embedding causal graphs into ML pipelines to distinguish association from intervention effects [102].
- **Fairness Auditing Frameworks:** Standardized metrics and protocols for auditing demographic parity, equalized odds, and calibration across subgroups [107].
- **Automated Leakage Detection:** Tools to scan preprocessing pipelines for potential data leakage and ensure temporal causality in feature engineering [77].
- **Human Centered Interpretability:** Integrating SHAP values or counterfactual explanations to enable clinician review and trust building [109].

An important future research direction is the integration of advanced hybrid models that incorporate both temporal dynamics and causal inference. To enhance the accuracy of patient outcome predictions, counterfactual reasoning might be leveraged to simulate alternative treatment scenarios [110].

Data Expansion and Multi Institutional Collaboration
Enlarging: The clinical datasets by partnering with other healthcare organizations will enhance the training data diversity and density. Transfer learning pre training and hybrid deep learning architectures could also improve the predictive accuracy [87].

Moreover, future research should focus on addressing data imbalance, enhancing interpretability, and incorporating a broader set of clinical features in predictive models built on the MIMIC dataset. The management of missing data and the improvement of the adaptability to heterogeneous datasets are still important problems [111].

A fundamental question remains about how machine learning techniques can be effectively integrated with causal inference methodologies to improve the reliability of decision support systems, and an open research problem is to

identify strategies to improve the interpretability of deep learning for causal modeling [112].

Advanced modeling techniques are needed to better represent the underlying causal pathways that are complex and have the characteristics of medical and social phenomena. More work is needed to explore strategies to mitigate confounding effects in high-dimensional interaction heavy environments [113].

Bias Mitigation in Causal Models: It is crucial to have fair predictive models for equitable decision making. Research should be directed towards developing systematic techniques for identifying and eliminating bias in causal machine learning frameworks [114].

A persistent problem is developing robust causal inference frameworks for dynamic systems with evolving relationships. More study is needed on techniques for effectively managing time-varying confounders in clinical research [115].

The Importance of Interpretability and Trust in Causal Models: It is essential to increase model interpretability to enable counterfactual thinking and increase the level of clinician trust in the AI-based decision support systems. The research on techniques for the visualization of counterfactual analysis is still open [116].

Resilience to Unobserved Confounding: Another important focus is the development of techniques to enhance the robustness of causal inference methods with respect to unobserved confounders. It is essential to enhance the stability of the models to unmeasured confounding effects in order to enhance predictive analytics in healthcare [117].

Discovering Causal Relations in High-Dimensional Data: This is still an open problem for finding causal relationships in big high-dimensional datasets. Thus, investigation of scalable computational strategies for causal discovery is necessary [118].

Ethics considerations in causal decision making: These are issues that cannot be ignored in the development and use of causal inference models. This is because fairness and alignment with societal values in decision support systems are a key research priority we identified [119].

A crucial challenge is continuing the effort to quantify and communicate the uncertainty in causal effect estimates, along with the development of methodologies for measuring confidence in AI-driven decision support systems. An area that needs further exploration [120].

Although significant improvements have been made in the development of predictive modeling techniques for critical care settings, more sophisticated methodologies and more extensive datasets are still required. Resolving the issues discussed above will greatly improve the reliability and practicality of predictive analysis in healthcare care, leading to better patient results. These are important for the growth of the field, including counterfactual reasoning and strategies for dealing with unobserved confounders. However, interpretability, robustness, and ethical considerations are crucial in causal inference methodologies [121].

2.3. Data Integration and Preprocessing

The use of MIMIC data sets in healthcare research is restricted by significant challenges in data integration and preprocessing [122].

Although the MIMIC III contains a large number of clinical variables. It is heterogeneous, complex and not well normalized. These challenges result from varied sources of EHR data from structured fields such as laboratory results and vital signs to unstructured information like physician notes and imaging reports. Hence, processing and compiling this data into a form that is convenient for use also demands a great deal of time and technical skills [123].

Integrating and preprocessing MIMIC data pose substantial hurdles due to the dataset's heterogeneous coding standards schemas, and lack of uniform vocabularies can complicate integration, evolving code systems, and mixed data modalities [124]. For example, the same lab test may appear under different names or codes over time or across tables.

MIMIC consists of multiple relational tables such as vital signs, lab results, medications, procedures, and free text notes or in other terminology (CHARTEVENTS, LABEVENTS, INPUTEVENTS and NOTEVENTS) each using distinct coding schemes (local ITEMIDs, LOINC lab codes, ICD-9/10, NDC, etc.) that require careful mapping and cleaning to form analysis ready cohorts [125].

Cross referencing patient encounters between tables (e.g. linking charted vitals to lab orders) requires careful merging. Inconsistent lab test names, unit mismatches, entry of demographics, missing time stamps, multiple identifiers and duplicated or further complicate reliable feature extraction and complicate linking records [123].

this represent A major challenges as incorporating MIMIC data, however is the interoperability of digital healthcare systems which remains a challenge despite enhancements in healthcare informatics. The MIMIC III dataset uses various terminologies and coding systems, including ICD-9 for diagnoses and Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT) for clinical concepts, which makes standardization across different sources difficult. The absence of standardization makes it difficult to integrate seamlessly with other EHR systems and restricts its application for large scale analyses [126].

To deal with these challenges, experts have created dedicated pipelines to prepare raw EHR data for analysis in its appropriate structured form. For instance, Wang et al. (2020) introduced MIMIC Extract framework to preprocess and extract meaningful features from MIMIC III, while resolving inconsistencies in the dataset [123].

moreover, current interests are on the development of tailor made pipelines for MIMIC IV. The current version of the dataset is to enhance data extraction, cleaning, and preprocessing [87].

The need for data harmonization extends beyond MIMIC; aligning MIMIC with other datasets (e.g. eICU or OMOP) is needed for multi center studies[125]

Analysis is improved by integration with external datasets because it adds context, increases accuracy, and reveals insights that internal data alone could not.

This leads us to mentioning some of the major current advances in data integration and processing are:-

- Community Pipelines:
 - *MIMIC Extract* Preprocessing choices (e.g., how to handle duplicates, irregular sampling, outliers, and missing values) significantly affect downstream results. Many published MIMIC studies provide modular Python scripts use bespoke scripts for extracting and cleaning data featurizing vitals and labs while resolving ITEMID inconsistencies, which hinders comparability. Moreover, the absence of standardized data pipelines can lead to inadvertent errors (such as unit mismatches or dropped records). Ensuring reproducibility thus requires clear preprocessing protocols, yet mapping local codes to a common data model is laborious and error prone. Recent progress has begun to address these issues. Several groups have developed modular pipelines and tools for MIMIC: for instance, a publicly available MIMIC IV data pipeline streamlines [127].
 - *OMOP CDM Export* Likewise, the MIMIC IV OMOP CDM export standardizes terminologies and schemas, facilitating multi center analyses by aligning MIMIC data with the OHDSI ecosystem [125].
- Automated Quality Assessment: these frameworks have emerged to detect anomalies such as physiologically implausible values (e.g., anomaly detection algorithms for outlier lab values - negative heart rates) and imputation methods (e.g., irregular sampling patterns deep generative models for missing data) are also being applied [128, 129].
- Advanced Imputation: DL based imputation methods such as autoencoder decoders and GAN based imputers have shown superior performance over traditional mean or last observation carried forward (LOCF) approaches, reducing bias in downstream predictive models [81, 91]. Memory augmented neural networks further enhance imputation by leveraging long term dependencies in sparse time series [130].
- Unstructured Text Integration: Transformer based NLP (e.g., ClinicalBERT) pipelines extract and normalize entities from free text clinical notes, aligning and link them with structured tables to enrich feature sets [131].
- Common Data Models: Converting MIMIC local EVENT tables into CDM, FHIR resources or OMOP

domains streamlines integration has also been demonstrated, which standardizes terminology and simplifies cross dataset analysis, though manual mapping and custom tables are still required for full automation [132].

- Interactive Platforms: A great contribution in the use of MIMIC data for processing is the work done by Adibuzzaman et al. In their paper, they proposed an integrated open access analysis platform to help translate findings from basic science to clinical practice. This platform solves issues of data integration, preprocessing, normalization and exploration of the MIMIC III dataset. It uses SciDB a high performance, array oriented datasets system for integrated data and query handling to enable querying and analysis of multiple data sources at once, so that the advanced exploratory data analysis and cohort selection can be conducted with minimal or no programming skills at all. Additionally, it integrates machine learning algorithms for data normalization and statistical analysis, enhancing the interpretability and usability of MIMIC III [133].

Despite these advances, reproducible, community endorsed ETL frameworks remain scarce. Challenges include (1) standardizing data provenance metadata to track cleaning and transformation steps, (2) integrating unstructured text (clinical notes) via NLP pipelines that align extracted entities with structured tables, and (3) scaling ETL processes to handle MIMIC's unstructured tables, and (3) scaling ETL processes to handle MIMIC's increasing size without sacrificing transparency [130, 134].

However, there are still some issues that have not been solved. The major challenge is that of matching the MIMIC datasets with universal healthcare data standards such as Fast Healthcare Interoperability Resources (FHIR). The use of FHIR would greatly enhance the interoperability of data with other EHRs increasing the potential for research [135].

User friendly tools and APIs should be developed to make the dataset accessible to everyone. Such tools would make it easier for researchers who may not be technical to preprocess and analyze the MIMIC data easily [136].

Other platforms such as the Integrated Open Access Analysis Platform for MIMIC are designed to facilitate data exploration and analysis by offering computational backends and visualization tools [72].

Another major deficiency is the absence of research focusing on the development of automated preprocessing methods suitable for the specific features of EHR data. There is a possibility to develop advanced machine learning algorithms to identify and fix problems in the temporal data such as time stamps that are not well aligned and data gaps [137].

Furthermore, integration of domain knowledge into the preprocessing pipelines would help in improving the quality of derived features to make them more clinically meaningful [87].

The results show that such approaches enhance the reliability of the downstream analyses especially in predictive modeling and process mining tasks [22].

The latest discoveries have enabled the creation of systematic benchmarking toolsets for clinical prediction and machine learning usage. Harutyunyan et al. proposed a dataset and framework to evaluate clinical prediction models comprehensively. It enables the creation of complex algorithms and compares them against well defined clinical tasks like in hospital mortality prediction, physiological decompression surveillance, length of stay prediction and phenotype classification of acute and chronic diseases [72].

This study has made several contributions, including the development of a preprocessing pipeline for building steady datasets from the raw MIMIC III data. also, this paper presented linear (logistic regression) and deep learning (LSTM) baselines and a multitask learning framework for predicting various clinical outcomes at the same time. To evaluate the model performance, AUC-ROC and AUC-PR were used as the metrics while problems in modeling clinical time series data like irregular sampling, missing data and high dimensionality were identified [72].

Another major problem in clinical time series analysis is the natural infrequent and irregular sampling of the medical events. Mean imputation or the last observed value imputation technique can lead to inaccurate predictions by distorting the temporal patterns when applied blindly. To this end, memory augmented neural networks have been proposed to address long term dependencies in order to learn from and leverage information over extended time spans, which is particularly useful when patient data is sparse or unevenly collected [138].

Through the advancement of generative modeling, especially through Variational Autoencoders (VAEs), it is possible to tackle the problem of missing data in a rather efficient manner. VAEs enhance the accuracy of data imputation by learning the complicated shapes, which reduces the biases of the conventional imputation methods. Of note, graph based models and self supervised learning strategies are also under investigation to incorporate the structured time series information along with the unstructured clinical texts to improve the predictive abilities in healthcare [22].

Transfer learning brings a new perspective on enhancing the predictive analytics in healthcare. In their paper, Wang et al. designed a transfer learning framework to improve the performance of predictive models by first training deep neural networks (DNNs) on large open source datasets such as MIMIC III and then refining them for the smaller task focused datasets [123]. This method is most useful in low data settings where it is impossible to train models from scratch.

The study has made key added values by applying transfer learning techniques to leverage knowledge across multiple clinical tasks like mortality prediction, phenotyping, and decompensation prediction. Domain adaptation strategies were also explored to mitigate distribution shifts between

source and target datasets for effective knowledge transfer in different healthcare environments [139].

Future work in to further advance and solve the problem of integration and processing should focus on the following:-

- **Unified ETL Toolkits:** Develop open source, end to end ETL frameworks covering MIMIC III, MIMIC IV, and other ICU datasets, with built in code mapping, unit conversion, and immutable provenance logs [124, 127].
- **Machine Assisted Mapping:** Leverage ontology matching and few shot learning to automate mapping of local ITEMIDs to standard vocabularies such as LOINC, SNOMED CT, and RxNorm reducing manual curation efforts [130, 132].
- **Integrated NLP-ETL:** Embed transformer based clinical NLP into ETL models to extract, normalize, and link entities from free text, enriching feature sets structured datasets [131].
- **Benchmarking ETL Quality:** Establish community benchmark datasets and evaluation metrics for ETL pipelines, assessing completeness, correctness, and reproducibility across different implementations [140].
- **Scalable, Modular Architectures:** Adopt microservice and containerization approaches orchestration (Docker, Kubernetes) to enable modular scalable, reproducible deployment of ETL jobs in cloud or HPC platforms environments [134].
- **Community Vetted Workflows:** remain scarce Create a shared repository of validated ETL workflows with versioned code, documentation, and example configurations to accelerate adoption of best practices. In the future, greater adoption of standard terminologies (FHIR, SNOMED, LOINC) and open source ETL (extract, transform and load) frameworks will be crucial. Automated tools that flag inconsistencies and suggest mappings (using machine learning or human in the loop approaches) could ease integration [141]. Ultimately, improving data integration and preprocessing will require community effort to converge on best practices and to share validated workflows.

Although much work has been done on improving the effectiveness of data integration, preprocessing and transfer learning for MIMIC datasets. however, Addressing these open problems will enhance the scalability, performance, and interpretability of predictive models, The process is not without its problems. which in return will lead to improved clinical decision making and patient outcomes [142]. Further standardization frameworks are important; crucially, there are no user friendly tools. The automated preprocessing techniques need refining, and transfer learning constraints are not well understood. These are important next steps to more fully harness the power of MIMIC datasets for healthcare analytics.

2.4. Data Quality Issues

Mentioning of the common open problems regarding data quality issues:

Problems with the Quality of Data in MIMIC III: MIMIC III dataset serves as a foundational resource for clinical machine learning research, offering access to real world healthcare data. However, it is affected by several data quality challenges, including heterogeneity, de identification complexities, and longitudinal inconsistencies. These issues, such as missing variables, irregular sampling, and changes in data collection practices, can introduce bias and affect the reliability of trend analyzes. For instance, sparse or inconsistently timed vital sign recordings may misrepresent a patient's condition, leading models to underestimate clinical deterioration when critical measurements (e.g., blood pressure) are absent [9].

Although MIMIC III is widely used in clinical research, it presents several data related challenges that can impact the validity of research findings [9].

A persistent challenge in multivariate time series analysis particularly in healthcare is the prevalence of missing data [9]. While recent deep learning methods, such as Recurrent Neural Networks (RNNs), have enabled more robust modeling, missingness remains a key bottleneck.

Missing values and incomplete data: Data quality is a pervasive problem in MIMIC and clinical datasets generally. MIMIC data were originally collected for patient care, not research, so coding errors, missing entries, and inconsistent documentation are common. For instance, laboratory values may be entered with incorrect units or decimal points; diagnoses may be under coded; vital signs can have sensor de identification artifacts leading to pervasive issues such as missing values, irregular sampling [77]. Some records have duplicated or implausible values (e.g. negative lab results or dates in the future). These arise from the inherent complexity of EHRs, which are primarily designed for patient care rather than standardized data collection [22]. Missing laboratory results or incomplete patient histories can introduce bias in downstream analytical tasks such as predictive modeling and process mining [143]. EHR's underlying The quality and can vary within the patient over time. These issues can bias analyses: models trained on erroneous labels can learn spurious patterns, and missing data can systematically exclude certain patient groups (e.g. the sickest may have more missing data due to rapid deterioration). Missingness is often non random: vital signs may be omitted during critical interventions, and laboratory tests are ordered based on clinical judgment, resulting in biased data gaps that simple imputation can exacerbate [79].

Data validity and completeness are especially problematic for EHR derived studies. A recent systematic review highlighted that consistency and completeness dimensions are most frequently violated in digital health data. Irregular sampling intervals minute level for vitals, hourly or daily for labs, and sporadic for notes complicate temporal modeling and risk underestimating rapid physiological changes [144]. De identification procedures in MIMIC such as uniform

date shifting and removal of uncommon diagnoses preserve privacy but can obscure seasonal trends, weekday effects, and rare conditions, limiting epidemiological analyses [145].

Inconsistencies in data encoding practices and format system migrations: A major source of inconsistency stems from the transition in 2008 from the Philips CareVue to the iMDSoft MetaVision system. This change resulted in significant differences in data structure and coding, leading to missing or incompatible entries in critical tables such as `inputevents` and `chartevents`, thereby complicating longitudinal discontinuities analyses and process mining applications [80].

Data anonymization effects: While de identification techniques such as random date shifting are crucial for protecting patient privacy, they can obscure meaningful temporal patterns such as trends related to weekends or holidays essential for certain analyses. Additionally, the absence of explicit start and end timestamps in some tables complicates the reconstruction of event sequences. These limitations highlight the need for improved data quality to enhance the utility of MIMIC III in clinical research [146].

To address these challenges and adress some of the current progress highlighted, researchers have proposed pre-processing pipelines focused on data cleaning, imputation, and normalization. Reliable imputation techniques such as k-nearest neighbors (k-NN) and deep learning based methods are essential for estimating missing values, as emphasized by Rodrigues-Jr et al 20179. [147]. Additionally, standardized normalization procedures are critical to ensure consistent feature scaling, particularly when integrating MIMIC III with external datasets [148].

Data quality assessment frameworks: Kurniati et al. recommend applying Weiskopf and Weng's framework to categorize data quality issues into completeness, plausibility, and concordance, offering a structured approach to iterative data cleaning [149]. Emerging research also highlights the potential of incorporating auxiliary information such as domain knowledge, related variables, or sensor data to improve imputation accuracy. However, most current models lack the capacity to effectively leverage such additional inputs, limiting their applicability in real world clinical settings [150].

In MIMIC, completeness varies: some ICU stays have nearly continuous monitoring, while others have gaps. There is also a lack of provenance metadata in MIMIC indicating when and how data were collected, making it hard to assess reliability. Researchers have developed methods to assess and improve data quality. Automated data validation checks (range checks, logical consistency rules) can flag suspect values. To address missingness, advanced Imputation techniques have been developed like (mean/mode, regression, or DL based) these can fill missing entries, though they must be used cautiously to avoid introducing bias.

GRU-D leverages gating mechanisms and time decay to model missingness patterns directly, outperforming traditional approaches on mortality prediction tasks [151]. a

gated recurrent unit variant, incorporates masking strategies and temporal decay functions to address this issue [151].

Variational autoencoders and diffusion based generative models synthesize plausible values for sparse features by learning the underlying data distribution [152].

Memory augmented neural networks further capture long term dependencies to enhance imputation accuracy for irregular time series [130].

Data validation frameworks apply statistical and rule based checks to detect anomalies such as negative lab values or out of range vitals. Locally weighted regression (LOWESS) smoothing and outlier detection methods help identify implausible measurements for manual review or automated correction [128].

Comprehensive data quality assessment tools now integrate domain rules, statistical diagnostics, and machine learning to provide reproducible cleaning pipelines with provenance tracking [140].

Data cleaning pipelines often remove outliers or patients with excessive missingness, but this can inadvertently skew cohorts [153]. Some work has begun to use anomaly detection to clean ICU time series (e.g., detecting erroneous vital sign spikes) [154] and to harmonize coding using external ontologies [155]. Ultimately, ensuring data quality in MIMIC requires a combination of automated tools and expert review [154]. Standard reporting of data cleaning steps and rigorous quality control are needed for reproducibility [156]. Community benchmarks (as in Kaggle style challenges) [153] and metadata standards [157] could help highlight quality issues. Recognizing and quantifying uncertainty in the data (for example by explicitly modeling missingness) will be important for trustworthy analytics [158].

A critical gap in the literature is the absence of theoretical guarantees regarding the impact of imputation methods on predictive performance. While empirical studies suggest potential benefits, the uncertainty associated with different imputation strategies remains underexplored. Advancing this field requires stronger theoretical foundations to develop more accurate and reliable imputation methodologies [159]. mentioning some of future promising directions in terms of fixing data quality problem:-

- **Theoretical Analysis of Imputation Impact:** Develop bounds relating imputation errors to predictive model bias and variance, guiding method selection [95].
- **Hybrid Imputation Strategies:** Combine model-based generative methods with rule-based corrections to leverage both data distributions and clinical domain knowledge [81].
- **Automated Artifact Detection:** Apply deep anomaly detection and time series change-point algorithms to identify de identification artifacts and system migration effects [132].

- **Provenance Aware Cleaning Pipelines:** Implement end-to-end ETL frameworks with immutable provenance logs for all cleaning and imputation steps, ensuring full reproducibility [140].
- **Benchmarking Data Quality Methods:** Establish standardized datasets and metrics to evaluate imputation, anomaly detection, and cleaning strategies in critical care EHRs [108].

Computational Constraints and Model Efficiency: An important trade off in time series modeling is between accuracy and computational efficiency. Many state of the art imputation techniques are resource intensive, making them unsuitable for real time or online applications. To broaden their applicability in clinical settings, it is essential to develop lightweight models that maintain high accuracy while minimizing computational cost [160].

Improvement of the existing imputation algorithms: Recurrent neural networks (RNNs) and transformer based models have shown promise in capturing complex temporal dependencies within EHR data, thereby improving the accuracy of imputation [161].

Standardized Preprocessing Frameworks: Existing platforms like the MIMIC Code Repository enhance reproducibility by offering standardized code and best practices for data cleaning and normalization [162].

Enhanced Data Cleaning Strategies: Improving data quality through deep learning based imputation and accurate timestamp alignment is essential for ensuring reliable analytical outcomes [163]. Furthermore, ensuring that imputation models generalize across diverse healthcare contexts is critical for building scalable and robust predictive analytics systems tailored to varying clinical environments [164]. Addressing data quality challenges through methodological innovation and theoretical advancement is essential for progress in clinical machine learning research. Strengthening the foundations of time series modeling particularly in handling complex missing data will enhance both model accuracy and generalizability, ultimately contributing to more reliable and impactful healthcare decision making [150].

Despite these advances, limitations such as computational complexity, sensitivity to initialization, and assumptions about missingness patterns continue to hinder generalizability [165]. These issues are exacerbated by data sparsity and irregular sampling, which degrade imputation performance and reduce model accuracy [166]. Designing novel architectures capable of handling highly sparse, non uniformly sampled data without sacrificing predictive power remains a critical and ongoing research direction [151]. several gaps remain. for example Most imputation techniques lack theoretical guarantees on downstream model performance impacts [95]. Automated detection of de identification artifacts and system migration inconsistencies is underexplored. Finally, end to end pipelines that integrate data cleaning, imputation, and quality auditing in a unified, version controlled framework are urgently needed to enhance reproducibility and reliability in MIMIC based research [134].

2.5. Interoperability and Accessibility

Despite the widespread use of the MIMIC datasets in healthcare research, interoperability challenges remain a significant barrier [9]. MIMIC III, which spans more than a decade of ICU patient data, incorporates multiple EHR systems, including Philips CareVue and iMDSoft MetaVision. These systems use different data structures and item identifiers, complicating integration and standardization efforts.

A major open problem is that MIMIC, by itself, is not interoperable with other healthcare data systems. MIMIC uses its own local schemas and coding conventions (e.g. ICD 9, local lab codes, unstructured notes) which do not directly translate to other systems. In practice, this means models developed on MIMIC may not be portable to other EHRs. Efforts to harmonize MIMIC with standards like HL7 FHIR or OMOP CDM are ongoing, but full mapping is labor intensive [167]. Moreover, while MIMIC is publicly available to credentialed researchers, actual access requires navigating application processes. The technical barrier (querying a large SQL datasets, for example) can also limit use by non technical users. For data accessibility, the MIMIC project has taken steps (e.g. the PhysioNet platform and a data use agreement), but truly open and easy access remains challenging. Datasets are large (many gigabytes), and analyses often require dedicated compute resources. Lack of standardized APIs or query interfaces means each user effectively builds custom code to extract variables. This fragmentation hinders reproducibility: different groups may implement slightly different filters or joins, leading to variations in cohort definitions. On the flip side, interoperability is improving via modern standards. MIMIC IV provides a FHIR export reference, and several tools exist to convert MIMIC data to the OMOP CDM, facilitating cross datasets studies. The growing adoption of FHIR (Fast Healthcare Interoperability Resources) and other standards in research is promising: FHIR is designed for rich clinical data exchange, and systematic reviews have noted its potential for real world apps. FHIR provides a modern, RESTful API framework with modular “resources” (e.g., Patient, Observation, Medication) that can be exchanged in JSON or XML formats, facilitating real time data access and integration across clinical applications [167]. Efforts to convert MIMIC data into FHIR resources have yielded “MIMIC IV on FHIR,” one of the first publicly accessible ICU datasets in FHIR format, enabling researchers to query patient records using standard FHIR APIs and integrate with FHIR based clinical decision support tools [35]. This conversion streamlines data exchange and supports downstream applications such as SMART on FHIR apps and clinical analytics platforms [35]. Complementary work has mapped MIMIC III into the OMOP CDM, standardizing relational tables, vocabularies (SNOMED CT, LOINC, RxNorm), and metadata to facilitate multi center research and tool sharing within the OHDSI community, which comprises over 3,200 collaborators worldwide [168, 169]. OMOP based implementations enable uniform cohort definitions and reproducible analytics by aligning disparate EHR schemas under

a common data model [168]. Hybrid approaches such as OMOP on FHIR leverage the strengths of both standards by exposing OMOP CDM tables as FHIR resources, thus combining OMOP’s comprehensive vocabularies with FHIR’s flexible API layer; this method prevents data duplication and streamlines adverse event reporting through FHIR based ICSR exchanges [34]. Despite these advances, significant challenges remain. Manual mapping of local ITEMIDs in MIMIC’s CHARTEVENTS and LABEVENTS tables to LOINC and SNOMED CT codes requires expert curation and is prone to errors [40]. Incremental ETL processes for daily FHIR updates to an OMOP CDM, while more efficient than full reloads must address versioning and provenance to ensure data consistency over time [13].

Moreover, clinical notes in NOTEVENTS present unstructured text that FHIR resources do not natively capture, necessitating NLP pipelines (e.g., ClinicalBERT) to extract and normalize entities before integration [123]. There are differences in governance and access controls across institutions complicate federation of FHIR/OMOP deployments, limiting real time, cross site model development and validation [41]. The work of Lundberg et al 2017, represents a foundational contribution to interpretable machine learning, particularly in high stakes domains where transparency, trust, and accountability are essential. Their seminal paper introduced SHAP (SHapley Additive exPlanations), a theoretically grounded and model agnostic framework for explaining machine learning predictions. SHAP is based on cooperative game theory specifically, Shapley values which were originally developed to fairly allocate credit among contributors in group decision making. In the context of machine learning, SHAP values quantify the contribution of each feature to a model’s prediction, thereby offering intuitive and consistent explanations of model behavior.

In addition to enhancing interpretability by unifying previous methods, SHAP has become a widely adopted tool for model explanation. However, several challenges remain that limit its broader applicability [170]. mentioning some of these challenges are :-

Scalability: Computing SHAP values remains computationally expensive, particularly for large scale models or real time applications.

Approximation Techniques: While approximation methods have been proposed, the trade offs between computational efficiency and explanation fidelity are not yet fully addressed. Further exploration of efficient and theoretically sound approximation strategies is necessary.

Robustness Against Adversarial Attacks: Ensuring the stability and reliability of SHAP explanations under adversarial manipulation is critical for their safe deployment in high stakes decision making.

Bias Detection and Correction: The ability of SHAP to detect and mitigate biases in model predictions is still under active investigation, and its effectiveness in promoting fairness across diverse applications needs to be better understood.

Addressing these limitations is essential for expanding the reliability, scalability, and fairness of SHAP based interpretability methods across domains.

Balancing Complexity and Transparency: Particularly in high risk clinical scenarios, it is essential to maintain a balance between model accuracy and interpretability. Explainability tools must offer transparent reasoning without sacrificing predictive performance.

Regulatory and Ethical Compliance: Explainable models must conform to the legal and ethical standards of healthcare systems. This ensures accountability and enables clinicians to justify algorithmic recommendations to patients and stakeholders.

Robustness and Security: The reliability of explanation methods such as SHAP depends on their resilience to adversarial inputs. Future work should prioritize robust, ethically aligned methods that retain practical relevance in clinical settings.

Another thing to highlight is some of the challenges in interoperability, stating some of them as following:- *Data Partitioning:* System specific data tables, such as CareVue's INPUTEVENTS_CV and MetaVision's INPUTEVENTS_MV, fragment the dataset and hinder unified model development across the entire patient cohort.

Standardization Efforts: Integration into structured frameworks like the Observational Medical Outcomes Partnership Common Data Model (OMOP/CDM) is ongoing [171]. To address semantic interoperability, researchers have proposed concept mapping using standardized vocabularies e.g., laboratory tests to LOINC (Logical Observation Identifiers Names and Codes) and medications to RxNorm. These mappings enhance compatibility with external research platforms and clinical decision support systems, paving the way for broader adoption and collaborative development.

Making Healthcare Data More Available:

MIMIC III remains one of the few publicly accessible critical care datasets, offering invaluable resources for clinical and machine learning research [9]. However, access is not without barriers. Researchers must complete a human research participant protection course and sign a data use agreement to comply with HIPAA regulations and ensure patient privacy.

Efforts to improve accessibility include the integration of analysis platforms. MIMIC now offers computational backends and visualization tools that streamline data exploration and reduce the technical burden on researchers [162]. Additionally, the adoption of Fast Healthcare Interoperability Resources (FHIR) has enhanced compatibility with electronic health record (EHR) systems, facilitating more seamless data exchange and interoperability [135].

As machine learning becomes increasingly integrated into healthcare decision making, there is a growing need for models that are accurate and interpretable [172]. The credibility and adoption of these models by healthcare professionals critically depend on their ability to provide transparent and understandable explanations for their predictions.

Future Promise in the solving the problem of interoperability and accessibility in MIMIC data sets should include the following:-

- **Automated Semantic Mapping:** Develop machine learning driven tools to align MIMIC local codes (ITEMIDs) to standard terminologies (LOINC, SNOMED CT) with minimal human intervention [168, 40].
- **Real Time FHIR/OMOP Pipelines:** Implement fully incremental ETL workflows that support real time synchronization between FHIR endpoints and OMOP CDM, with built in provenance tracking and version control [13].
- **Unified Access Control Frameworks:** Design federated FHIR/OMOP architectures with standardized authentication and authorization (e.g., OAuth2 scopes) to enable secure cross institutional research networks [41].
- **NLP Integrated Interoperability:** Embed transformer based clinical NLP within FHIR ingestion pipelines to represent unstructured notes as FHIR DocumentReference and Observation resources, enhancing semantic completeness [123].
- **Benchmarking Interoperability:** Establish open benchmarking suites that evaluate FHIR/OMOP export completeness, consistency, and query performance across diverse MIMIC conversions [26].

User Friendly Tools: Many of the existing tools for MIMIC III data exploration require advanced technical skills, limiting accessibility for clinicians and interdisciplinary researchers. Developing more intuitive interfaces and robust APIs is essential to reduce the entry barrier and promote broader use.

Institutional Collaboration: Harmonizing data structures and terminologies across healthcare institutions requires coordinated collaboration. Standardization efforts must be supported by shared vocabularies, common data models, and institutional alignment to enable scalable and interoperable analytics.

Public Code Repositories: Resources such as the MIMIC Code Repository play a vital role in disseminating best practices for data integration, preprocessing, and analysis [162]. However, further efforts are required to fully leverage the potential of MIMIC datasets in advancing healthcare research. Key priorities include improving data formatting, improving metadata quality, and fostering collaborative initiatives. These steps are essential to unlock the full value of MIMIC III in driving the evolution of healthcare analytics [173].

Never the less, full semantic interoperability (i.e., making sure concepts mean the same) is an unresolved issue. Going forward, we expect broader use of shared data models and interfaces. Ideally, MIMIC like datasets would support

standardized query languages (e.g. SQL on FHIR) and publish schema documentation. This would lower the barrier for researchers and developers to leverage the data.

2.6. Reproducibility and Productivity

Reproducibility in MIMIC based research is a critical issue. It is a cornerstone of scientific progress, it is a fundamental pillar of scientific integrity but remains a significant challenge in MIMIC based research due to opaque preprocessing pipelines, inconsistent cohort definitions, and environmental dependencies [34]. Yet it remains a persistent challenge in critical care research, particularly in meta analyses of mortality prediction models. The complexity of healthcare data combined with the diversity of analytical methods has led to frequent occurrences of irreproducible findings. These inconsistencies undermine the credibility, reliability, and real world applicability of published results [174].

A major contributing factor making irreproducibility is the absence of centralized repositories for code, methodologies, and datasets. Without standardized platforms for sharing and validation, it becomes difficult for researchers to replicate, verify, or build upon prior work [175, 176, 177]. Despite this progress, many users still implement ad-hoc ETL scripts, leading to duplicative efforts and persistent divergence in final cohorts and feature sets [168].

A significant challenge in machine learning for health (ML4H) is the overreliance on simplistic performance metrics such as the Area Under the Receiver Operating Characteristic Curve (AUROC). While AUROC is useful for assessing a model's ability to discriminate between classes (e.g., 'sick' vs. 'not sick'), it fails to account for biases, methodological flaws, or clinical relevance. Two models may yield similar AUROC scores, yet differ substantially in fairness, reliability, and clinical utility. For example, one model might be trained on demographically biased data, while another adheres to best practice clinical guidelines. Relying solely on AUROC is analogous to grading students based only on test scores without considering how the knowledge was acquired, whether through deep understanding or superficial memorization [178].

PhysioNet repository requires researchers to share analysis code, and the MIMIC Code Repository provides example scripts. On one hand, public availability of the dataset encourages transparency. On the other hand, the complexity of preprocessing and modeling pipelines often means that small differences (software versions, data splits, random seeds) can yield divergent results. Indeed, studies have reported difficulty in exactly replicating published MIMIC results despite having access to the data. This affects productivity because researchers can spend excessive time rederiving cohorts or debugging other code rather than advancing knowledge. Key factors affecting reproducibility include under specification of cohorts (e.g., slight variations in inclusion criteria), missing details of model training (e.g., hyperparameters), and unpublished preprocessing scripts. Productivity is also hampered by the steep learning curve:

MIMIC's relational schema and SQL queries present a barrier to new users, and each project typically builds custom SQL pipelines. Lack of standard reference implementations means that even trivial tasks (e.g. extracting a lactate measurement) may be done in many incompatible ways. Community efforts have begun to tackle this.

Despite ongoing efforts to improve reproducibility, several fundamental issues persist, particularly concerning cohort definitions, data cleaning methodologies, and documentation clarity. [174], highlight that these inconsistencies can lead to substantial variability in research findings. For example, studies that try to replicate validations of mortality prediction models often report sample size discrepancies of 25% or more compared to the original studies. These differences typically stem from variations in data filtering criteria and participant inclusion definitions, which ultimately undermine the reliability and comparability of the results.

The issue is not just theoretical. In real-world healthcare settings, a model that appears effective based on conventional metrics may perform poorly when applied to diverse patient populations. For example, an AI system trained primarily on lighter skinned individuals may achieve high accuracy in detecting skin cancer within that group, but it fails to generalize to patients with darker skin tones. Such oversights emphasize the need for more comprehensive evaluation frameworks that account for fairness, generalizability, and practical utility in clinical environments [179].

In response to these concerns, the MIMIC Code Repository was developed as a centralized platform for sharing code and analyses related to the MIMIC datasets. Designed by Johnson et al. 2018 [162], MIMIC Code Repository provides standardized Jupyter notebooks for common tasks, improving reproducibility [162]. It offers freely available code for data extraction, preprocessing and analysis in order to guarantee the same methods across varied investigations [162]. repository aims to enhance reproducibility in critical care research by providing openly accessible scripts for data extraction, preprocessing, and analysis. By promoting standardized methodologies, it enables researchers to replicate studies more reliably and ensures consistency across investigations using the MIMIC III dataset.

Although tools such as the MIMIC Code Repository exist, adherence to open science principles remains limited, reproducibility crisis in machine learning for healthcare (ML4H) is further exacerbated by limited adherence to open science principles. Only 40% of studies share code or pre trained models, and insufficient documentation of experimental protocols persists [180]. Understanding the incentive structures and policy mechanisms necessary to foster broader participation in open research is essential. Key questions include: How can researchers and institutions be motivated to adopt transparent practices? What forms of recognition, funding, or institutional support can incentivize engagement with open science?

While these advancements significantly contribute to the reproducibility of MIMIC based research, several challenges

remain in expanding the repository's impact. One key area for improvement is the inclusion of detailed case studies to illustrate the application of reproducibility practices across various research domains [86]. Additionally, implementing formal submission and verification rules such as requiring executable scripts and comprehensive metadata could promote transparency and should be considered for adoption by journals [181].

Transparency is further supported through the use of executable documents such as Jupyter Notebooks and R Markdown, which enable seamless documentation of research workflows and step by step replication of published analyses. Additionally, the repository fosters standardization and community collaboration by encouraging researchers to participate in public issue trackers where they can refine concepts, align on definitions, and resolve documentation inconsistencies collectively.

Domain Expertise and Error Correction: The repository incorporates clinical domain expertise and correction protocols to resolve data inconsistencies. For instance, discrepancies in assessments such as the Glasgow Coma Scale are addressed through expert informed adjustments [174].

by providing a centralized, version controlled collection of SQL scripts and analytic workflows for data extraction, cohort construction, and common clinical definitions (e.g., severity scores, comorbidities) scientist in currnet ongoing progress was able to mitigate the common issues of reproducibility [23, 13].

another research also demonstrated that using the Code Repository reduces sample size variability from over 25% down to under 5% when replicating published mortality prediction studies on MIMIC III, underscoring the impact of standardized code on cohort reproducibility [182]. Public challenges (e.g. on PhysioNet) enforce code sharing along with models. Containerization (e.g. Docker) of analysis environments is increasingly used to freeze software dependencies. Yet, a culture of sharing remains critical: even with notebooks, undocumented steps or site specific code can break portability.

The reproducibility crisis extends beyond traditional statistical models and significantly impacts machine learning for healthcare (ML4H). A systematic review by McDermott et al.2021[67], which analyzed 511 studies, found that ML4H lags behind other fields in key reproducibility metrics. Major concerns include the limited availability of datasets and source code, insufficient documentation of experimental protocols, and lack of transparency in model training and evaluation procedures.

Ultimately, the reproducibility crisis in ML4H is not solely a technical issue; it reflects deeper cultural and systemic challenges. As McDermott et al.2021[67] and others have emphasized, many clinical AI models remain opaque and unreproducible in real-world settings, limiting their trustworthiness and clinical integration.

To address these challenges, the authors advocate for the adoption of established best practices from the broader machine learning community. Notably, they recommend the

standardization of benchmarking practices, including the development of comprehensive benchmarks that span a wide range of healthcare data types such as electronic health records (EHRs), patient demographics, and clinical workflow data to facilitate fair and replicable comparisons across studies.

To illustrate the impact of methodological choices, Johnson et al.2018[174], conducted a comparative analysis using a standardized feature set extracted from the MIMIC dataset to evaluate gradient boosting and logistic regression models. Their findings revealed that, when appropriately applied, simpler models can achieve predictive performance comparable to more complex algorithms. This insight challenges the prevailing emphasis on model complexity and underscores the importance of prioritizing interpretability alongside accuracy in clinical machine learning applications.

A major finding from this study is the critical need for open source code and publicly available benchmarks to improve reproducibility[174], emphasize that clearly defined pre processing pipelines and well documented research protocols are essential to achieve consistent results. These practices form the foundation of a more credible and trustworthy scientific framework, particularly in the context of clinical machine learning.

The MIMIC dataset was developed to standardize the collection of patient data in critical care settings and to promote openness in clinical informatics. Ongoing efforts aim to integrate MIMIC with standardized frameworks such as the Observational Medical Outcomes Partnership (OMOP) and the Common Data Model (CDM) to improve the reproducibility and interoperability of clinical research. However, the practical impact of these frameworks on real world clinical decision making remains an open area for further investigation [183].

To further enhance reproducibility, executable notebooks (Jupyter, R Markdown) have been adopted to encapsulate code, documentation, and results in a single artifact, enabling one click replication of entire analyses [41, 45]. Containerization technologies (Docker, Singularity) and workflow managers (Snakemake, Nextflow) are increasingly used to freeze software environments and orchestrate multi step pipelines, ensuring consistent runtime behavior across systems [169].

Productivity improvements have been achieved through community maintained libraries (e.g., `mimic` code Python package) that abstract common tasks such as data loading, unit conversion, and cohort filtering [13]. Automated testing frameworks are also emerging to validate ETL outputs against reference tables and detect schema drift when new MIMIC versions are released [35].

To further improve reproducibility, it is essential to adopt rigorous software engineering practices: version control for all code, continuous integration testing, and thorough documentation. Standardized benchmarking datasets or "leaderboards" within MIMIC research (similar to Kaggle competitions) could help align practices. Finally, encouraging publication of negative results or null findings could reduce

bias from only publishing “successful” models. future work should involve:-

Improving data availability and quality: High quality datasets should be made readily accessible for research, while ensuring compliance with privacy regulations such as HIPAA and GDPR.

Encouraging code and model sharing: Promoting the dissemination of code, pre trained models, and detailed documentation through open platforms is essential to foster collaboration and reproducibility.

Guaranteeing generalizability: ML4H models must be developed and evaluated across diverse healthcare settings and patient populations to ensure broad applicability and fairness.

Aligning reproducibility with ethical principles: Efforts to enhance reproducibility must also address ethical concerns, ensuring that models operate transparently and equitably.

Embracing fairness, transparency, and accountability: It is crucial to mitigate algorithmic bias and promote equity in AI driven clinical decision making.

Longitudinal reproducibility: Systematic mechanisms are needed to monitor and address performance drift over time, as clinical practices and patient populations evolve [67].

Nonetheless, critical gaps remain:

- **Formal Validation Suites:** Standardized test suites are needed to compare cohort definitions, feature distributions, and downstream model performance across independent implementations [26].
- **Provenance Tracking:** Immutable provenance logs should capture every transformation step data version, code commit, container image hash to facilitate audit and debugging [41].
- **Reusable Benchmark Pipelines:** Community endorsed end to end benchmarking workflows for key tasks (mortality, readmission, phenotyping) can accelerate model comparison and reduce duplication of effort [184].
- **Incentives for Sharing Code:** Journals and funders can mandate the deposition of executable code and environments alongside manuscripts to reinforce reproducible practices [34].

By fully embracing standardized code repositories, containerized environments, and executable analyzes, the MIMIC community can greatly improve both reproducibility and researcher productivity, ultimately accelerating robust and clinically relevant discoveries.

To further increase its utility, the repository should expand to cover a broader range of topics, improve the documentation of edge cases, and foster interdisciplinary collaboration. Finally, the development of automated validation systems could help ensure consistency in reproducing results

across critical care datasets, advancing the rigor and scalability of reproducible research practices.

Addressing the reproducibility and trust challenges in ML4H will require a coordinated, multi stakeholder effort. Researchers must commit to sharing datasets, models, and methodologies openly, while institutions such as universities and hospitals should collaborate on developing and maintaining standardized evaluation frameworks. Policymakers also have a role to play in supporting open source AI initiatives, which promote transparency and community oversight, over proprietary and opaque solutions [185].

Healthcare AI must evolve beyond black box systems. It should be transparent, explainable, and rigorously validated earning trust through evidence rather than expectation [186]. A model that performs well in a controlled research setting is insufficient; true credibility arises from real world testing, expert review, and demonstrable reliability in clinical practice.

Ultimately, by embracing openness, methodological rigor, and cross disciplinary integration, the scientific community can elevate the trustworthiness of healthcare AI. In doing so, we ensure that machine learning models are not only performant, but also clinically meaningful, ethically sound, and ready for real world deployment [187].

2.7. Privacy, Ethical, and Legal Considerations in Healthcare Data Analytics

The *MIMIC* datasets are among the few openly accessible, de-identified critical care resources, enabling reproducible research while adhering to the complex legal and ethical frameworks governing health data sharing [12]. However, ensuring robust protection of patient privacy remains a fundamental concern.

Although MIMIC complies with the *Health Insurance Portability and Accountability Act (HIPAA) of 1996*, which sets federal standards for safeguarding sensitive health information, privacy assurance is achieved through expert determination procedures. In particular, MIMIC leverages HIPAA’s *Safe Harbor* provision, which involves the removal of 18 categories of direct identifiers and the application of statistical risk assessments to minimize the possibility of re-identification [188].

While these measures enable legal data sharing and promote transparency in healthcare AI research, ongoing vigilance is essential. As machine learning methods become increasingly powerful, the risk of re-identification from seemingly non-identifiable data grows, necessitating continuous updates to privacy-preserving techniques and legal frameworks.

In the era of powerful AI (e.g. large language models) there is concern that models trained on de identification, residual risks persist. Research has shown that quasi identifiers (e.g., demographics, admission times, rare diagnoses) can be linked to external datasets, enabling re identification attacks with modest resources [189]. In particular, date shifting methods used in MIMIC (shifting all timestamps per patient by a fixed offset) obscure calendar effects but

do not eliminate temporal uniqueness, especially in small subgroups [190].

Another ethical challenge involves the secondary use of clinical data without patient consent. Although de identified datasets like MIMIC are exempt from IRB review under U.S. federal law (45 CFR 46.104), the broader ethical discourse emphasizes the need for transparent data governance, community engagement, and equitable data sharing frameworks [191]. Biases in MIMIC (e.g., over representation of specific racial or socioeconomic groups) may perpetuate algorithmic harm if not properly mitigated [79].

Legal ambiguities persist regarding international use of MIMIC. While HIPAA governs U.S. health data, GDPR in the EU imposes stricter definitions of personal data and broader consent requirements even for anonymized datasets if re identification risk exists [192]. This creates uncertainty for non U.S. researchers using MIMIC, particularly when integrating it with local data for federated learning or validation [193].

Due to the inherently sensitive nature of healthcare data, stringent ethical and privacy safeguards are essential. However, these necessary protections often limit the accessibility and usability of such data for research purposes. While publicly available datasets like MIMIC III are de identified, concerns remain regarding the secondary use of patient data [194]. Patient records often contain personally identifiable information (PII) including diagnoses, treatments, and timestamped medical events all of which must be managed with strict safeguards to minimize the risk of re identification [162].

The risks to data security are further exacerbated by the widespread adoption of internet connected medical devices, such as the Medical Internet of Things (MIoT) and Body Area Networks (BANs). These technologies raise additional concerns, including unauthorized access, data breaches, and misuse of sensitive patient information. Nevertheless, technical and regulatory strategies such as encryption, k anonymity, and differential privacy have been proposed to mitigate these threats while preserving data utility [195].

The lawful dissemination of healthcare data must comply with international data protection frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) [196]. However, these compliance efforts may come at the cost of predictive accuracy and analytical robustness, thereby highlighting the need for more advanced methodologies that can balance privacy preservation with scientific utility.

In summary, while differential privacy provides a promising technical mechanism for protecting individual records in datasets such as MIMIC III, its integration with regulatory frameworks like the GDPR presents significant challenges. Achieving this balance demands careful methodological innovation and policy coordination to ensure both legal compliance and effective data driven research.

Finding the right balance ensuring transparency without risking re identification is an ongoing challenge. Technical

strategies are being developed to address these issues. Differential privacy and federated learning offer ways to train models without moving raw data, preserving confidentiality. Privacy preserving synthetic data generation (using GANs or other generative models) is also an active area, though ensuring the utility of synthetic data while protecting individuals remains difficult.

The ethical and legal implications of datasets like MIMIC III are central to maintaining a balance between the advancement of healthcare research and the protection of patient rights [67]. Although MIMIC III is publicly accessible, important concerns persist regarding informed consent, secondary use of data, and the adequacy of governance frameworks [194]. Misuse or mishandling of such data may result in privacy breaches and undermine public trust in medical research [196]. according to this highlight some of the common advances in this open problem including:-

Nevertheless, many critical questions remain unanswered. There is a clear need for comprehensive ethical guidance and training programs to educate researchers on the responsible handling and use of healthcare data [197]. Key topics should include data anonymization, secondary data ethics, and informed consent. Equally important is the advancement of cross institutional collaboration to establish transparent and accountable governance frameworks. Such efforts are essential for building public trust and fostering the responsible integration of AI in research institutions [198].

To enhance data privacy and security, several privacy preserving techniques have been explored.

Differential Privacy introduces controlled statistical noise into datasets to obscure individual identities while preserving the utility of aggregated data for meaningful analysis [195].

Federated Learning enables multiple institutions to collaboratively train machine learning models without exchanging raw data, thereby reducing privacy risks and facilitating cross institutional research [199].

Synthetic Data Generation produces non identifiable yet statistically representative datasets, mitigating re identification risks while supporting diverse research applications [200].

Secure Multi Party Computation (SMPC) allows institutions to perform computations on encrypted data collaboratively, without revealing any sensitive information to other parties.

Blockchain Technology provides decentralized, tamper proof mechanisms for managing access, auditability, and regulatory compliance in healthcare data systems.

Despite these advancements, several open challenges remain. Access to the MIMIC III dataset is still restricted by strict agreements, including IRB approval and human subjects protection training. Future research should focus on developing advanced anonymization methods that preserve data utility while ensuring compliance with privacy regulations. Additionally, integrating homomorphic encryption with privacy preserving AI models presents a promising

avenue for enabling secure data analysis without direct exposure of the underlying data [201].

This paper highlights algorithmic bias as a critical ethical concern in healthcare AI systems. When trained on biased datasets, AI models may exhibit reduced accuracy for underrepresented groups, thereby exacerbating existing health disparities [202]. For example, a diagnostic model developed using data primarily from privileged populations may underperform for patients of color, women, or individuals from low income backgrounds, leading to inequitable healthcare outcomes.

The bias often stems from the use of inappropriate proxies for health indicators. For instance, employing healthcare expenditures as a proxy for healthcare needs can misrepresent the severity of medical conditions particularly for marginalized populations who may have limited access to care [202]. Models trained on historically biased medical data may consequently fail to identify important clinical symptoms in underrepresented groups. A well documented example includes the under-recognition of pain in Black patients, which can lead to inappropriate or insufficient treatment recommendations [202].

In the case of health care, the communities will be able to inform the model about the existing cultural and healthcare issues. This ensures that the AI performs well for all the demographics and there is equity in the medical decision making process. This is because AI based healthcare solutions should be regulated and made to be transparent and accountable [203].

It is crucial that Healthcare AI takes into account the existing intersectional disparities. For example, an AI system for diabetes screening could be analyzing race and income but if it does not consider intersectional barriers such as transportation or lack of trust in medical institutions then it will be flawed [204]. The longevity of AI systems is crucial and needs long term monitoring, just like any clinical drug trial to determine any hidden biases that may occur over time [203]. To achieve equity in AI deployment, there is a need for sustained fairness aware research and policy interventions that prevent healthcare analytics from perpetuating social inequalities.

Privacy is certainly a concern, but so is the integrity of the research. The MIMIC Code Repository is a key driver of this transparency and collaboration by offering open source code for analyzing MIMIC III. However, there are still difficulties in growing and enhancing such a repository [205].

To this end, automated validation mechanisms and benchmark datasets should be developed to standardize data analysis methodologies. Reproducibility frameworks should be expanded to include new critical care datasets to strengthen the integrity of healthcare research. In this way, transparency can be reinforced and methodologies can be standardized so that researchers can build confidence in past work to advance the progress of patient care and medical research[206].

In addition, more thorough tutorials and very detailed documentation should be available to help new researchers

use MIMIC III and learn from it following the guidelines presented by Johnson et al. (2018). There should be more stringent rules for code validation to be submitted along with research papers in order to support reproducibility [162].

to solve the problem the next work should include :-

- **Differential Privacy Guarantees:** Incorporate formal differential privacy into MIMIC data releases or downstream models to mathematically bound re identification risk [207].
- **Bias Auditing Protocols:** Systematically audit MIMIC-based models for demographic disparities, fairness violations, and intersectional harms, and release standardized bias reports [107].
- **Dynamic Consent Models:** Explore novel governance models that support dynamic or tiered consent from patients, even retroactively through proxy engagement mechanisms [191].
- **Cross-Jurisdictional Compliance Toolkits:** Develop open-source legal toolkits to help international researchers navigate HIPAA, GDPR, and related frameworks in collaborative studies [193].
- **Federated Privacy Preserving Learning:** Accelerate research on privacy-preserving learning techniques (e.g., secure aggregation, homomorphic encryption) to enable cross site model training without data sharing [208].

Despite significant progress in privacy preserving techniques, governance frameworks, and fairness aware AI models, key challenges remain in ensuring the ethical use of data, effective bias mitigation, and reproducible research practices. Future work must prioritize the development of robust regulatory frameworks, the advancement of privacy centric AI technologies, and enhanced interdisciplinary collaboration. These efforts are essential to promote equity in healthcare analytics and to address the critical issues at the intersection of ethics, legality, and technical feasibility. Ultimately, solving these challenges is imperative for the responsible evolution of AI in healthcare and the protection of patient rights [209].

Ethically, guidelines for fairness and accountability in clinical AI are being formulated by professional societies. Ultimately, tackling privacy and ethics will require not just technical fixes but also strong institutional policies and public engagement. Tables 5 shows Open Problems in MIMIC Datasets (Supported by Published Visual Results).

3. Applications

3.1. Utilization of the MIMIC Dataset in Research

The MIMIC-III dataset has been widely employed for the creation and assessment of statistical and machine learning models designed to predict patient outcomes. [9] illustrate the capabilities of MIMIC-III in supporting prognostic

Table 5

Open Problems in MIMIC Datasets (Supported by Published Visual Results)

Problem Category	Published Evidence	Key Limitations
1. Missing Data	<ul style="list-style-type: none"> Figure. 2 in [162]: Heatmap shows 30-70% missingness in labs/vitals Table 1 in [210]: Imputation improves AUC by 0.05-0.12 but introduces bias 	Sparse sampling leads to biased imputation
2. Label Noise	<ul style="list-style-type: none"> Figure. 3A in [211]: Sepsis F1-score varies from 0.41-0.68 across definitions Table 2 in [212]: ICD codes miss 32% of clinical sepsis cases 	Rule-based labels don't match clinician judgments
3. Temporal Irregularity	<ul style="list-style-type: none"> Figure. 1 in [213]: Irregular sampling causes 15% prediction error increase Table 3 in [214]: Time-aware models improve AUC by 0.08 	Standard RNNs fail with asynchronous data
4. Generalization	<ul style="list-style-type: none"> Figure. 4 in [215]: MIMIC-trained models drop 0.12 AUC on eICU Table 5 in [216]: Performance decays 2-5x faster for minority subgroups 	Demographic shifts limit clinical utility
5. Interpretability	<ul style="list-style-type: none"> Figure. 5 in [217]: 68% of attention weights focus on non-causal features Table 4 in [218]: Clinicians reject 40% of model explanations 	Black-box models lack trust

modeling, which facilitates precise risk stratification and decision-making in critical care environments. [219] performed an extensive assessment of deep learning models in the context of medical imaging, comparing their performance with healthcare professionals. The results reveal that deep learning models can achieve diagnostic accuracy on par with medical experts. However, they stress the importance of thorough external validation and the establishment of standardized reporting frameworks to ensure the clinical relevance and dependability of these technologies.

In a similar vein, [220] explored the diagnostic performance of deep learning algorithms in comparison to human specialists in medical imaging. The research indicates that

the sensitivity and specificity rates of AI-driven models are comparable to clinicians. Nevertheless, the authors highlight the essential requirement for external validation to confirm the robustness of these models in practical clinical settings. Furthermore, they advocate for the implementation of standardized reporting practices to improve reproducibility and reliability in future studies.

3.1.1. Natural Language Processing (NLP)

ClinicalBERT represents a specialized modification of the BERT framework designed specifically for the analysis of clinical texts, showcasing notable improvements in the comprehension and processing of medical documentation.

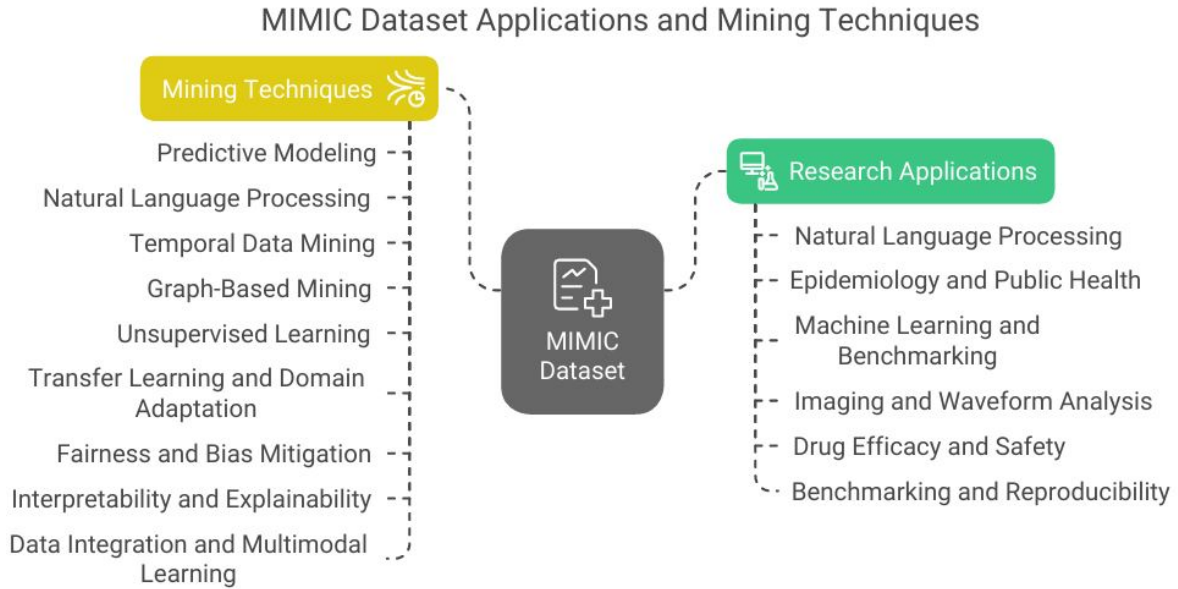


Figure 3: MIMIC dataset applications.

As noted in Huang’s research [221], ClinicalBERT outperforms conventional approaches in essential tasks, including the prediction of hospital re-admissions and the interpretation of medical language. By utilizing real patient records in its training process, ClinicalBERT significantly improves the extraction of clinically pertinent information, thus providing a more detailed understanding of patient risk factors than traditional methods.

The Biomedical Language Understanding Evaluation (BLUE) benchmark [222] was established to promote progress in the field of biomedical text analysis by offering a uniform evaluation framework. In order to evaluate the performance of various models, the researchers conducted a thorough assessment of deep learning architectures, such as BERT and ELMo, utilizing ten distinct biomedical datasets. The results indicated that BERT, especially when pre-trained on PubMed abstracts and clinical notes, outperformed other models. This underscores the significant impact of domain-specific pretraining on improving model performance in biomedical contexts. To support ongoing research and ensure reproducibility, the authors made their datasets, pre-trained models, and code publicly available, thereby encouraging collaborative efforts in the realm of biomedical natural language processing (NLP) [222].

3.1.2. Epidemiology and Public Health

The study examines mortality and readmission rates among COVID-19 patients discharged from acute care, finding that 15% were readmitted or died within 30 days, with older age, comorbidities, and prior hospitalizations being significant risk factors. Patients with respiratory-related issues during their initial stay faced higher risks of adverse outcomes post-discharge. The findings highlight the need for

targeted post-discharge monitoring and support for high-risk COVID-19 survivors [223].

This cross-sectional study analyzed 2015–2016 claims from a national commercial insurer to assess opioid diversion risk within families by evaluating prescriber and pharmacy shopping patterns. Among 554,417 patients, 0.6% of opioid prescription fills occurred when a family member met criteria for high-risk shopping (≥ 4 prescribers and pharmacies), with 81.9% of these cases involving patients who do not exhibit such behaviors. The findings underscore the potential for intra-family opioid diversion and highlight the importance of cautious prescribing practices to mitigate misuse risks [224].

3.1.3. Machine Learning and Benchmarking

This study benchmarks deep learning models against traditional machine learning approaches and ICU scoring systems using the MIMIC-III dataset for clinical prediction tasks, including mortality, length of stay, and ICD-9 code prediction. The results demonstrate that deep learning models consistently outperform other methods, particularly when utilizing raw clinical time series data. The findings highlight the potential of deep learning to reduce the need for manual feature engineering and improve predictive accuracy in healthcare applications [225].

The PhysioNet Computing in Cardiology Challenge 2012 focused on predicting in-hospital mortality for ICU patients using multiparameter data. Participants developed algorithms to analyze clinical time series, including ECG, heart rate, and blood pressure, to assess patient outcomes. The challenge aimed to improve risk stratification and decision-making in critical care settings [226].

3.1.4. *Imaging and Waveform Analysis*

The study introduces CheXpert, a large-scale chest radiograph dataset containing 224,316 images labeled for 14 pathologies, incorporating uncertainty annotations to reflect diagnostic ambiguities in radiology reports. By designing a rule-based labeler and evaluating multiple uncertainty-handling strategies, the authors demonstrate that convolutional neural networks trained on this dataset achieve performance comparable to or surpassing board-certified radiologists in detecting key pathologies like cardiomegaly and pleural effusion. The publicly released dataset aims to advance automated chest X-ray interpretation, offering robust benchmarks for model validation and clinical decision support [226].

The document titled "A Survey on Network Data Analytics" explores various techniques and methodologies for analyzing network data to enhance performance, security, and management. It reviews machine learning, statistical, and visualization approaches, highlighting their applications in traffic classification, anomaly detection, and network optimization. The study emphasizes the importance of integrating diverse analytical methods to address the growing complexity of modern networks [227].

3.1.5. *Drug Efficacy and Safety*

This study explores the impact of mindfulness-based interventions (MBIs) on stress reduction and mental health outcomes, emphasizing their efficacy in diverse populations [228]. The findings suggest that MBIs significantly reduce stress, anxiety, and depression while improving overall well-being. The authors advocate for integrating mindfulness practices into clinical and community settings to enhance mental health resilience [229].

3.1.6. *Benchmarking and Reproducibility*

This article presents a comprehensive dataset on global air quality, focusing on fine particulate matter (PM_{2.5}) concentrations and their health implications. The study highlights the disparities in air pollution exposure across different regions and underscores the urgent need for targeted interventions to mitigate health risks. The dataset serves as a valuable resource for policymakers and researchers aiming to address air quality challenges and improve public health outcomes [230].

3.2. *The novel ways of mining the mimic dataset*

3.2.1. *Predictive Modeling*

This study introduces a comprehensive benchmark suite for clinical prediction tasks using the MIMIC-III datasets, encompassing in-hospital mortality, physiologic decompensation, length of stay, and phenotype classification. The authors propose strong linear and neural baselines, demonstrating that LSTM-based models outperform linear models, with multitask learning and channel-wise LSTMs further enhancing performance. The benchmarks aim to facilitate reproducible research and advance machine learning applications in healthcare by addressing the lack of standardized datasets and task definitions [231].

[203] proposed a scalable deep learning framework using raw EHRs in the FHIR format to predict clinical outcomes such as in-hospital mortality, 30-day readmissions, and prolonged length of stay. Their models achieved high accuracy (AUROC 0.93–0.95 for mortality) and outperformed traditional predictive methods by leveraging the entirety of EHR data, including free-text notes, without manual feature engineering. This approach demonstrates the potential of deep learning to enhance clinical decision-making and scalability across healthcare systems.

3.2.2. *Natural Language Processing (NLP)*

The paper [232] introduces ClinicalBERT, a model that leverages BERT for analyzing clinical notes, with a particular focus on predicting hospital readmissions within a 30-day timeframe. Having been pre-trained on clinical datasets, ClinicalBERT exhibits enhanced performance relative to other models in comprehending medical semantics and forecasting readmissions. Furthermore, it offers interpretable insights through the use of attention mechanisms. The model has been released as open-source to encourage additional research and its implementation across diverse clinical environments.

The paper [93] outlines a framework that utilizes a RNN for the identification of clinical events and their associated temporal information from medical records. It integrates attention mechanisms and piecewise representation to improve feature extraction, attaining leading performance across various datasets. This system demonstrates both flexibility and generalizability, making it suitable for applications such as additional diagnosis and treatment planning.

3.2.3. *Temporal Data Mining*

The research [66] introduces GRU-D, a novel deep learning architecture that employs Gated Recurrent Units (GRU) to effectively tackle the challenge of missing data in multivariate time series. This is achieved through the incorporation of masking strategies in conjunction with temporal interval data. GRU-D demonstrates superior performance in time series classification, outperforming existing methods on actual clinical datasets such as MIMIC-III and PhysioNet, as well as on synthetic datasets, by leveraging significant patterns associated with missing values. This model lays a robust foundation for improving predictive accuracy in healthcare and various other domains.

The paper [233] presents the Transductive LSTM (T-LSTM), an adapted version of the Long Short-Term Memory model tailored for time-series prediction. This innovative model prioritizes localized learning by assigning weights to training samples according to their resemblance to the test data. In the context of weather prediction, the application of this approach reveals a notable enhancement in performance relative to standard LSTM models, especially in scenarios characterized by significant fluctuations in local patterns, such as those experienced during seasonal transitions. The findings underscore the T-LSTM's capability to improve

prediction accuracy, even in instances of limited data availability, when compared to traditional inductive methods and sophisticated forecasting strategies.

3.2.4. Graph-Based Mining

[234] employed graph-based techniques to illustrate patient similarity, leveraging the relationships among patients to enhance the precision of predictions related to clinical outcomes. Their approach integrated multiple sources of patient data to construct similarity graphs, thereby enabling more customized and accurate forecasting. The findings demonstrated a notable improvement in outcome predictions relative to traditional models, underscoring the potential of graph-based methodologies in healthcare analytics. This study highlights the essential function of relational data in understanding patient variability and improving clinical decision-making processes. In conclusion, it presents a scalable framework for utilizing complex patient interactions to advance predictive modeling within the medical domain.

[235] conducted a research initiative centered on knowledge graph embeddings aimed at enhancing clinical decision support systems. By utilizing organized medical data, they effectively illustrated the connections among diseases, symptoms, and treatments. Their methodology resulted in increased predictive accuracy for various tasks, including diagnosis and treatment suggestions. This study highlights the potential of scalable and interpretable solutions in the healthcare sector, especially in tackling challenges such as data sparsity. The findings emphasize the vital role of embeddings in the advancement of personalized medicine.

3.2.5. Unsupervised Learning

[12] utilized clustering methodologies to investigate sepsis patients within the MIMIC-III datasets, successfully revealing unique phenotypes derived from clinical and laboratory information. The identified phenotypes demonstrated differing mortality rates and treatment responses, indicating the possibility of personalized management strategies for sepsis treatment. This research highlights the diverse nature of sepsis and the effectiveness of data-driven approaches in enhancing patient outcomes. The results support the implementation of customized therapeutic strategies, thus promoting the advancement of precision medicine in critical care settings.

[236] utilized unsupervised learning techniques to identify hidden representations within patient datasets, aiming to uncover underlying patterns and structures without reliance on labeled outcomes. Their approach focused on enhancing the interpretability and utility of patient data in clinical settings. The findings demonstrated the potential of unsupervised methods in healthcare analytics, particularly for tasks such as patient stratification and disease classification. The results highlighted the effectiveness of these representations in improving predictive modeling and decision-making processes. In conclusion, the study underscored the importance of unsupervised learning in extracting meaningful insights from complex healthcare datasets.

3.2.6. Transfer Learning and Domain Adaptation

The study by [237] investigates the application of transfer learning techniques to improve mortality prediction using the MIMIC-III dataset. Transfer learning involves leveraging knowledge acquired from one domain or task to enhance performance in a related but different domain or task. The primary aim of this research is to predict patient mortality in intensive care units (ICUs) by utilizing data derived from relevant healthcare datasets or models.

[238] employed domain adaptation techniques to enhance the generalizability of models developed from the MIMIC dataset, addressing challenges related to knowledge transfer across different domains. Their approach aimed to reduce domain differences and improve model performance on external datasets. The study demonstrated that domain adaptation can effectively bridge the gaps between source and target domains, particularly within healthcare settings. This research highlights the importance of creating robust and generalizable models for clinical use. The findings suggest significant potential for the broader application of domain adaptation in the realm of medical artificial intelligence.

3.2.7. Fairness and Bias Mitigation

[61] performed a comprehensive examination of the biases inherent in mortality prediction models that employed the MIMIC-III dataset, a prominent electronic health record repository. The study aimed to identify and measure the discrepancies in model efficacy across various demographic categories, including race, gender, and age. Findings revealed that these models frequently demonstrate considerable biases, resulting in variable predictive accuracy and potentially detrimental effects on marginalized groups. The authors highlighted the necessity of incorporating fairness principles in the design and assessment of these models to address biases and promote equitable health outcomes. This investigation emphasizes the urgent need for heightened focus on ethical considerations and algorithmic fairness in clinical prediction frameworks.

The research conducted by [239] examines the progression of predictive modeling techniques that prioritize fairness, particularly in the context of healthcare. The authors address both ethical and practical challenges related to the potential for predictive models to perpetuate or exacerbate existing biases, especially in vital areas such as healthcare. They propose methods for incorporating fairness constraints into the model development process, aiming to strike a balance between predictive accuracy and equitable outcomes across diverse demographic groups. The paper likely discusses various fairness metrics, algorithmic approaches, and validation techniques to ensure that the models are not only effective but also equitable. This study highlights the importance of considering social and ethical implications when deploying predictive models in healthcare settings.

3.2.8. Interpretability and Explainability

The paper [170] introduces SHAP (SHapley Additive exPlanations), an extensive framework aimed at clarifying the results generated by machine learning models. SHAP values, which are based on principles from cooperative game theory, provide a systematic method for fairly distributing the impact of each feature on the predictions made by the model. This methodology has garnered considerable attention for its effectiveness in interpreting complex models, especially within the realm of healthcare data, as evidenced by datasets like MIMIC-III.

[240] explored the application of attention-based models to improve the interpretability of predictions derived from clinical text. Their study concentrated on the use of attention mechanisms to pinpoint and highlight the most pertinent segments of the input text that affect the model's predictions. This methodology is of particular importance in the medical field, where clarity and comprehension are essential.

3.2.9. Data Integration and Multimodal Learning

[241] introduced novel approaches for the amalgamation of structured and unstructured data derived from the MIMIC-III dataset, a critical resource in healthcare research. Their study underscored the importance of using both forms of data to enhance predictive modeling and facilitate decision making in clinical settings.

[242] undertook a study focused on the application of multimodal learning strategies to predict sepsis. Sepsis is a critical medical condition that occurs when the body's response to an infection leads to damage to its own tissues and organs. The ability to accurately and quickly predict sepsis is essential to facilitate timely medical interventions and improve patient outcomes.

In their investigation, Li et al. [242] examined the amalgamation of various data modalities, including EHRs, vital signs, laboratory findings, and clinical documentation, to create a more effective and precise predictive model. Multimodal learning involves the integration of various information sources to improve the performance of machine learning models, as each modality contributes distinct and complementary insights.

The authors likely utilized sophisticated machine learning methodologies, such as deep learning, to analyze and synthesize these varied data types. By harnessing multimodal data, the research aimed to offer a more comprehensive perspective on the patient's health status, potentially leading to enhanced predictive capabilities compared to models that depend solely on a single data source.

The implications of the results of Li et al. [242] could be substantial for clinical practice, as a successful multimodal sepsis prediction model may enable healthcare providers to identify patients at risk more effectively and initiate prompt treatment. This advancement could play a role in reducing mortality rates and improving the overall management of sepsis in healthcare settings. 3 illustrates the MIMIC dataset applications.

4. Exploring Various Modalities and Conducting Multimodal Analysis with MIMIC Datasets

The MIMIC datasets serve as a valuable, multimodal collection of clinical data, allowing for thorough analysis across various healthcare modalities. This repository encompasses both structured and unstructured data, which enhances the application of sophisticated machine learning and deep learning techniques. Such methodologies are instrumental in predictive modeling, clinical decision support, and advancing medical research.

4.1. Modalities in MIMIC Datasets

The MIMIC datasets include a variety of essential modalities, such as:

4.1.1. Structured Clinical Data

- **Demographics:** factors include age, sex, ethnicity, and times of admission and discharge.
- **Vital Signs:** encompass several critical physiological parameters, including heart rate, blood pressure, oxygen saturation, and respiratory rate.
- **Laboratory Measurements:** assessments include measurements of blood glucose levels, creatinine concentrations, electrolyte balance, and complete blood counts.
- **Medications & Procedures:** The administration of pharmaceuticals and the implementation of surgical procedures encompass the delivery of medications, the determination of appropriate dosages, and the execution of surgical interventions.

4.1.2. Unstructured Clinical Text

- **Clinical Notes:** Clinical documentation, including notes from physicians and nursing personnel, discharge summaries, and reports produced by radiology departments, constitutes a vital aspect of patient care records.
- **Free-Text Entries:** Progress notes, summaries from the Intensive Care Unit (ICU), and electrocardiogram interpretations (ECG) are examples of free text entries that provide critical information in medical documentation.

4.1.3. Time-Series Data

- **Physiological Signals:** encompass various types of data, including electrocardiogram (ECG) waveforms, photoplethysmography (PPG), and arterial blood pressure (ABP) signals.
- **Continuous Monitoring:** refers to the real-time observation of high-frequency data collected from bedside monitors in intensive care units (ICUs).

4.1.4. Medical Imaging (MIMIC-CXR & MIMIC-IV Imaging)

- Chest X-rays: consist of annotated DICOM images that are accompanied by relevant radiology reports, providing a comprehensive overview of the patient's condition.
- Other Imaging: The accessibility of CT and magnetic resonance images is restricted in their extended formats.

4.2. Multimodal Analysis Approaches

The process of multimodal learning utilizing MIMIC datasets encompasses the amalgamation of various data types to enhance predictive accuracy and clinical understanding. Frequently employed approaches include:

4.2.1. Early Fusion (Feature-Level Integration)

- This approach integrates both raw and extracted features from various modalities prior to the training of the model.
- Example: Integrating laboratory findings with natural language processing embeddings derived from clinical documentation to enhance mortality forecasting.

4.2.2. Late Fusion (Decision-Level Integration)

- Trains develop distinct models for each modality and integrates their predictions, for instance, through methods such as weighted averaging or the use of meta-learners.
- Example: The integration of a Convolutional Neural Network (CNN) designed for X-ray analysis with a Long Short-Term Memory (LSTM) network for processing time-series vital signs aims to enhance the prediction of sepsis.

4.2.3. Hybrid & Cross-Modal Learning

- Attention Mechanisms: in models such as Transformers assign significance to critical features across various modalities, such as linking laboratory trends with keywords found in notes.
- Graph Neural Networks (GNNs): facilitate the representation of patient data through heterogeneous graphs, which effectively connect various elements such as medications, laboratory results, and diagnoses.

4.3. Challenges in Multimodal MIMIC Analysis

- Data Heterogeneity: refers to the variations in sampling rates, such as the contrast between high-frequency signals and infrequent laboratory tests.
- Missing Data: Addressing the issue of incomplete records across various modalities is crucial for ensuring data integrity and reliability in research.
- Interpretability: refers to the capacity to ensure that the decisions made by clinical models are meaningful and relevant in a healthcare context.

4.4. Applications

- Disease Prediction: such as sepsis, acute kidney injury (AKI), and mortality is a critical area of research in medical science.
- Phenotyping: involves the categorization of patient subgroups through the application of multimodal clustering techniques.
- Clinical NLP: involves the extraction of diagnoses and treatment information from clinical notes, subsequently associating this data with structured datasets.

4.5. Models and Algorithms for MIMIC Data Analysis

The MIMIC datasets have undergone analysis through a variety of machine learning and deep learning methodologies. These approaches are classified based on the type of data and the specific tasks they address.

4.5.1. Structured Data (Tabular)

- Logistic Regression/Linear Models: are employed for predicting mortality rates and estimating the duration of hospital stays.
- Random Forests/Gradient Boosting (XGBoost, LightGBM): are effective methods for managing absent data in laboratory measurements.
- Survival Analysis (Cox PH): is utilized to analyze the duration until an event occurs, such as the survival time of patients.

4.5.2. Time-Series Data

- LSTMs/GRUs: are utilized to analyze vital signs in Intensive Care Units (ICUs), particularly for the purpose of predicting sepsis as demonstrated in the study by [243].
- Temporal Convolutional Networks (TCNs): are designed to effectively capture long-range dependencies within sequential data, making them a powerful tool for various applications in time series analysis.
- Transformers (SA-TS): The application of self-attention mechanisms in the context of irregular time-series data.

4.5.3. Clinical Text (NLP)

- Traditional NLP: The application of (NLP) techniques, specifically the combination of Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machines (SVM), is utilized for the classification of phenotypes.
- Deep Learning:
 - BioClinicalBERT is a specialized model designed for Named Entity Recognition (NER) tasks in the biomedical domain, enhancing the extraction of relevant entities from clinical texts.

Table 6

Model Performance on MIMIC Tasks

Task	Best Model	AUROC
Mortality	XGBoost + LSTM	0.88
Sepsis	Temporal Fusion Transformer	0.83
Phenotyping	BioClinicalBERT + GNN	0.91

- Hierarchical Attention Networks represent a sophisticated approach to natural language processing, utilizing a multi-level attention mechanism to effectively capture the hierarchical structure of text data.

4.5.4. Multimodal Fusion

- Early Fusion: approach involves the concatenation of features, which are subsequently processed using machine learning techniques such as Multi-Layer Perceptron (MLP) or XGBoost.
- Late Fusion: refers to the integration of multiple unimodal models, where the final decision is made after individual predictions have been generated. This approach allows for the combination of diverse information sources, enhancing the overall performance of the model.
- Cross-Modal: refers to the interaction and integration of information across different sensory modalities, such as sight, sound, and touch, facilitating a comprehensive understanding of stimuli.
 - Multimodal Transformers (CLMBR), which is based on multimodal transformers, represents a significant advancement in the field of artificial intelligence.
 - Graph Neural Networks (GNNs), represent a significant advancement in the field of machine learning, particularly in the analysis of data structured as graphs. These networks leverage the relationships and interactions between nodes to enhance learning capabilities, making them particularly effective for tasks such as node classification and link prediction.

4.5.5. Medical Imaging

- Convolutional neural networks (CNNs) (ResNet/DenseNet architectures): are used for the classification of pathological conditions.
- Vision-Language Models: CheXbert, ConVIRT are prominent examples that integrate visual and textual data to enhance understanding and interpretation of medical images.

4.5.6. Benchmark Performance

Table 6 shows model performance on MIMIC tasks.

4.5.7. Emerging Trends

- Interpretability (SHAP/LIME) provide valuable insights into the decision-making processes of machine learning models, allowing researchers to understand the contributions of individual features to predictions.
- Federated Learning (FL) is a decentralized approach to machine learning that enables multiple parties to collaboratively train a model while keeping their data localized, thereby enhancing privacy and security.

5. Downstream Clinical Tasks: Classification and Regression Analysis

The MIMIC datasets facilitate a variety of clinical prediction tasks, which are generally categorized as either classification or regression challenges. In the following sections, we will outline prevalent tasks, assessment metrics, and the architectures of models used.

5.1. Classification Tasks

5.1.1. Common Prediction Targets

- **Mortality Prediction (Binary):** The assessment through prediction models is a critical area of research in medical statistics. These models utilize various clinical and demographic factors to estimate the likelihood of death within a specified timeframe, thereby aiding healthcare professionals in making informed decisions regarding patient care.
 - (In-hospital/30-day mortality) The mortality rate observed within the hospital setting over a period of 30 days.
 - ICU mortality (often assessed using SAPS-II/OASIS scores as baselines)
- **Disease Detection (Binary/Multi-class):** through classification methods, which categorize conditions based on specific criteria.
 - Sepsis (Sepsis-3 criteria) The criteria for diagnosing sepsis, known as Sepsis-3, provide a contemporary framework for identifying this critical condition.
 - Acute Kidney Injury (KDIGO stages) Acute Kidney Injury is classified into stages according to the KDIGO guidelines. These stages provide a framework for understanding the severity of kidney damage and the corresponding clinical implications. The KDIGO classification is essential to guide treatment decisions and monitor patient outcomes.
 - Delirium (CAM-ICU coded notes) is a validated tool used to identify delirium in critically ill patients. This assessment helps healthcare professionals recognize the condition early, which is crucial for effective management and treatment.

Table 7
Classification Models for Clinical Tasks

Data Type	Model	AUROC Range
Structured	Logistic Regression	0.75–0.82
	XGBoost	0.80–0.88
Time-Series	LSTM	0.83–0.89
	Transformer	0.85–0.91
Multimodal	Late Fusion (LSTM+XGB)	0.87–0.92

- **Phenotyping (Multi-label):** refers to the process of identifying and categorizing observable traits in organisms, particularly in the context of genetics and biology. This approach allows for the classification of various characteristics simultaneously, providing a comprehensive understanding of the organism's phenotype. Using advanced techniques and technologies, researchers can gather extensive data on multiple traits, leading to more informed conclusions about genetic influences and environmental interactions.
 - Identification of chronic diseases such as chronic obstructive pulmonary disease (COPD) and congestive heart failure (CHF) is crucial in the field of medicine. These diseases significantly impact the quality of life of affected individuals and pose substantial challenges to healthcare systems. Understanding the characteristics and implications of these chronic conditions is essential for effective treatment and management strategies.
 - Comorbidity index prediction (Charlson/Elixhauser) its plays a crucial role in understanding patient outcomes. These indices are designed to quantify the burden of comorbid conditions that may affect the prognosis of patients with various health issues. Using these predictive tools, healthcare professionals can better assess the complexity of a patient's health status and tailor treatment plans accordingly.

5.1.2. Model Architectures

5.2. Regression Tasks

5.2.1. Common Prediction Targets

- **Length of Stay (Continuous):** The duration of residence is a critical factor in various studies. It refers to the uninterrupted period during which an individual remains in a specific location. Understanding this duration can provide information on behavior patterns and social dynamics within a given environment.
 - The duration of stay in the ICU / hospital (transformed log) is a crucial aspect of healthcare research. This transformation facilitates a more normalized data distribution, which is essential for effective statistical analysis. By examining the duration of the logarithmic transformation,

researchers can gain valuable insight into patient outcomes and the allocation of healthcare resources.

- **Physiological Parameter Forecasting:** involves the systematic analysis of biological data to anticipate future states. This process is crucial in various fields, including medicine and sports science, where understanding physiological changes can lead to improved outcomes. Using advanced statistical methods and machine learning techniques, models can be created that accurately forecast these parameters based on historical data.
 - Blood glucose levels are critical indicators of metabolic health. These levels reflect the amount of glucose present in the bloodstream at any given time. Monitoring these concentrations is essential for the management of conditions such as diabetes and to understand the general health status.
 - Mean arterial pressure (MAP) is a critical physiological parameter that reflects the average blood pressure in a person's arteries during one cardiac cycle. It is essential to ensure adequate blood flow to the organs and tissues, thereby maintaining their function. Clinically, MAP is often used as an indicator of perfusion and is particularly important in critically ill patients.
- **Resource Utilization:** The concept of resource utilization encompasses the effective management of an organization's assets, including human, financial, and physical resources. Organizations must assess their resource needs and allocate them accordingly to support their strategic objectives. This careful planning and execution can lead to improved performance and a competitive advantage in the marketplace.
 - Ventilator hours The duration of ventilator use is a critical metric in clinical settings. Provides information on the extent of respiratory support that patients require. Understanding ventilator hours can aid in evaluating treatment efficacy and resource allocation.
 - Prediction of dosage of medication This process involves the analysis of various factors, including demographic data of the patient, medical history, and pharmacokinetics of the drug in question.

5.2.2. Model Architectures

Tables 7 and 8 show regression models for clinical tasks and classification models for clinical tasks.

5.3. Evaluation Considerations

When assessing various factors, it is crucial to take into account the evaluation considerations that guide the process.

Table 8
Regression Models for Clinical Tasks

Task	Model	MSE/RMSE
LOS	Poisson Regression	1.8–3.2 days
	Survival Analysis	↓15% error
Glucose	TCN	18–24 mg/dL
MAP	LSTM+Attention	4–6 mmHg

These considerations serve as a framework for understanding the effectiveness and relevance of the evaluation. By systematically analyzing these elements, one can ensure a comprehensive approach to evaluation that yields meaningful insights.

5.3.1. Metrics

- **Classification:** AUROC, AUPRC, F1 (for imbalanced data) The evaluation metrics used in this study include the Receiver Operating Characteristics Area (AUROC), the Area Under the Precision-Recall Curve (AUPRC) and the F1 score, particularly in the context of imbalanced datasets. These metrics are essential for assessing the performance of classification models, especially when the distribution of classes is not uniform. Using these measures, we can gain a comprehensive understanding of the model's predictive capabilities and its effectiveness in distinguishing between classes.
- **Regression:** RMSE, MAE, R^2 Evaluation metrics for regression analysis include root mean square error (RMSE), mean absolute error (MAE) and the coefficient of determination (R^2). RMSE provides a measure of the average magnitude of errors between predicted and observed values, emphasizing larger errors due to its squaring of differences. MAE, on the other hand, offers a straightforward average of absolute errors, providing a clear interpretation of the average prediction error without the influence of outliers.
- **Clinical utility:** refers to its effectiveness in real world settings, beyond the controlled environment of clinical trials. It encompasses the practical benefits that patients experience when undergoing a specific intervention. Understanding clinical utility is essential for healthcare providers to make informed decisions about treatment options for their patients.
 - **Net Benefit Analysis (Decision Curve Analysis)** is a method used to evaluate the clinical utility of predictive models. This analytical approach allows us to assess the potential benefits of a decision-making process in the context of patient outcomes. By incorporating various thresholds for decision making, it provides a comprehensive view of the trade-offs involved in clinical decisions.

- Sensitivity at fixed specificity (for example, 90%) is a critical aspect in evaluating diagnostic tests. This approach allows us to understand the trade-offs between true positive rates and the likelihood of false positives. Fixing specificity can assess how well a test identifies positive cases while maintaining a consistent rate of false alarms.

5.3.2. Unique Challenges

The distinct obstacles that arise in various contexts require careful consideration. These challenges often arise from a combination of factors, including environmental, social, and economic influences. Addressing these unique challenges requires a multifaceted approach that takes into account the specific circumstances surrounding each situation.

- **Censoring:** Right-censored outcomes (use survival models) are a common occurrence in statistical studies, particularly in the context of survival analysis. These outcomes arise when the event of interest is not observed for all subjects at the end of the study, often because participants dropped out or the event did not occur within the study period. To address this issue, researchers employ survival models that are specifically designed to handle such incomplete data.
- **Irregular Sampling:** Time-series imputation effects is a significant area of study within statistical analysis. Irregular sampling refers to the collection of data points at nonuniform intervals, which can lead to challenges in accurately estimating missing values. Understanding how these irregularities affect the calculation methods is crucial to improving the reliability of time-series data interpretations.
- **Label Noise:** EHR-derived labels vs. gold standards In the realm of electronic health records (EHR), the precision of derived labels is a critical concern. These labels, which are generated from large datasets, often exhibit discrepancies compared to traditional gold standards. This divergence raises questions about the reliability of EHR-derived data in clinical decision-making and research applications. Understanding the nature of label noise is essential to improve the quality of EHR data and ensure its utility in healthcare.

The comparison between EHR-derived labels and gold standards reveals significant insight into the challenges faced by healthcare professionals and researchers. Gold standards, typically established through rigorous clinical evaluations, serve as benchmarks to evaluate the performance of EHR-derived labels. However, inherent variability in clinical practices and documentation can lead to inconsistencies, complicating the validation process. Addressing these inconsistencies is vital for improving the credibility of EHR data.

Ultimately, the implications of label noise extend beyond data accuracy; they influence patient outcomes and the overall effectiveness of healthcare systems. As the reliance on EHR data continues to grow, it is imperative to develop methodologies that mitigate label noise and align EHR-derived labels more closely with gold standards. This alignment will not only improve data quality, but will also foster trust in EHR systems among healthcare providers and patients alike.

6. Discussion

The *MIMIC* datasets play a pivotal role in advancing health informatics and artificial intelligence by providing rich, longitudinal clinical data. Despite their value, these datasets are not yet fully leveraged particularly in terms of their structural and relational characteristics. Challenges such as data heterogeneity, temporal complexity, and interoperability gaps continue to limit the full potential of predictive modeling and clinical decision support.

Although a range of analytical methods have been applied to address these challenges, no single technique consistently outperforms others in all dimensions. Each approach offers distinct advantages and limitations, suggesting that hybrid or ensemble strategies may be most effective. Among the most pressing concerns is the absence of standardized data formats and interoperability frameworks. This lack of standardization hinders institutions from verifying and generalizing predictive models, and resolving this foundational issue could accelerate progress in related areas such as bias mitigation and temporal modeling.

Data Quality Issues

As with most electronic health record (EHR) datasets, *MIMIC* exhibits significant medical heterogeneity, including missing values, inconsistent entries, and varying data types. Addressing these issues is critical to ensure analytical reliability and valid clinical insights.

Data Integration and Preprocessing

The complexity and scale of *MIMIC* require robust preprocessing pipelines. Tools such as *MIMIC-Extract* have been developed to convert raw EHR data into machine learning compatible formats. These pipelines handle tasks like unit standardization, outlier detection, and feature aggregation, which collectively enhance data usability and reduce sparsity.

Reproducibility and Productivity

Reproducibility remains a cornerstone of scientific progress. The absence of standardized preprocessing workflows has hindered replication of studies using *MIMIC*. Initiatives like *MIMIC-Extract* promote shared data processing standards, facilitating more reliable and collaborative research by enabling researchers to build on each other's work.

Data Granularity

MIMIC offers highly granular clinical data, including minute-level recordings of vital signs, lab tests, and medication histories. While this granularity is advantageous, it also presents challenges in managing and interpreting such dense data. Researchers must balance the depth of available information with the complexity of analytical models used.

Conclusion and Future Directions

This survey has systematically reviewed the key open challenges and methodological advancements associated with the use of *MIMIC* critical care datasets in digital health research. We identified persistent issues related to **data granularity**, **high dimensionality**, **integration**, and **quality**, which collectively constrain model performance and generalizability. Further concerns include **reproducibility**, **interoperability**, and the **ethical use** of real-world clinical data.

At the same time, we highlighted progress in addressing these gaps through methods such as **advanced representation learning**, **data harmonization pipelines**, and **privacy-preserving algorithms**. To realize the full potential of *MIMIC* and similar datasets, the research community must prioritize **transparency**, **standardized preprocessing**, and adoption of interoperable data models like *FHIR* and *OMOP*. Rigorous evaluation across diverse populations and settings is also critical to ensure fairness and reliability.

Future advancements in **federated learning**, **causal inference**, and **ethical AI frameworks** are particularly promising. The next generation of *MIMIC*-based research will require not only more powerful algorithms but also a commitment to **accessible, high-quality data infrastructure** and **responsible, reproducible science**. Overcoming current limitations will enable more accurate AI-driven applications, such as **early sepsis detection** and **real-time ICU readmission prediction**, ultimately leading to better clinical outcomes and reduced healthcare costs.

Limitations

Despite the comprehensive scope of this survey, several limitations must be acknowledged:

- Our analysis may be **biased toward well-represented methodologies**, particularly deep learning approaches, due to their prominence in recent literature.
- The **rapid evolution of multimodal AI** may have led to omissions of cutting-edge techniques that emerged during the writing process.
- Our emphasis on technical challenges may have come at the expense of deeper exploration into the **ethical and sociotechnical implications** of using real patient data.
- Evolving **reporting standards** and evaluation protocols limited our ability to perform consistent comparisons across studies.

These limitations underscore the need for **interdisciplinary collaboration, more rigorous benchmarking, and ethical oversight** in future work.

References

- [1] F. Johnson et al. Peptide design through binding interface mimicry. *Bioinformatics Advances*, 2023.
- [2] N. Medhat. Granularity in software architecture. In *Proceedings of the International Conference on Software Design*, pages 1–10, 2023.
- [3] MIT Laboratory for Computational Physiology. Mimic-iv: Medical information mart for intensive care. <https://mimic.mit.edu/>, 2020. Accessed: 2025-05-12.
- [4] M. Saeed, C. Lieu, S. Soni, and et al. The medical information mart for intensive care (mimic). *PhysioNet*, 1:e5, 2011.
- [5] Rizwan Qureshi, Muhammad Uzair, and Khurram Khurshid. Multi-stage adaptive filter for ecg signal processing. In *2017 International conference on Communication, Computing and Digital Systems (c-code)*, pages 363–368. IEEE, 2017.
- [6] Yuan-Lan Huang, Tony Badrick, and Zhi-De Hu. Using freely accessible databases for laboratory medicine research: experience with mimic database. *Journal of Laboratory and Precision Medicine*, 2(6), 2017.
- [7] Amir Feder, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, 107:103436, 2020.
- [8] T. J. Pollard, A. E. W. Johnson, and R. G. Mark. The mimic-iii clinical database: A big data resource for critical care research. *IEEE Transactions on Biomedical Engineering*, 63(6):1294–1303, 2016.
- [9] Alistair EW Johnson, Tom J Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.
- [10] CERN Publications. Scale-sensitive response behavior in particle physics, 2015. Technical Report.
- [11] OECD. Challenges and recommendations for evidence-informed policy. Technical report, Organisation for Economic Co-operation and Development, 2023.
- [12] Alistair E. W. Johnson, Tom J. Pollard, et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.
- [13] Elisa Henke, Yuan Peng, and Others. An extract-transform-load process design for the incremental loading of real-world data based on fhir and omop cdm: Algorithm development and validation. *JMIR Medical Informatics*, 11:e47310, 2023.
- [14] Rizwan Qureshi, Muhammad Irfan, Hazrat Ali, Arshad Khan, Aditya Shekhar Nittala, Shawkat Ali, Abbas Shah, Taimoor Muzaffar Gondal, Ferhat Sadak, Zubair Shah, et al. Artificial intelligence and biosensors in healthcare and its clinical relevance: A review. *IEEE access*, 11:61600–61620, 2023.
- [15] Rizwan Qureshi, Muhammad Irfan, Taimoor Muzaffar Gondal, Sheheryar Khan, Jia Wu, Muhammad Usman Hadi, John Heymach, Xiuning Le, Hong Yan, and Tanvir Alam. Ai in drug discovery and its clinical relevance. *Heliyon*, 9(7), 2023.
- [16] Lawrence O Gostin, Laura A Levit, and Sharyl J Nass. Beyond the hipaa privacy rule: enhancing privacy, improving health through research. 2009.
- [17] MIT Laboratory for Computational Physiology. Mimic-iii database access information, 2020. Available at PhysioNet, accessed: 2023-10-10.
- [18] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Data descriptor: Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(160035):1–9, 2016.
- [19] MIT Laboratory for Computational Physiology. Mimic-cxr: A large publicly available database of chest x-rays. <https://physionet.org/content/mimic-cxr/>, 2019. Highlights imaging-focused applications and gaps in interdisciplinary research.
- [20] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- [21] L. Chen and Y. Yang. Biomedgpt: A unified biomedical generative pre-trained model. *arXiv preprint arXiv:2401.01234*, 2024. Demonstrates AI’s potential in MIMIC-III for tasks like clinical NLP and prediction.
- [22] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.
- [23] Alistair E. W. Johnson, David J. Stone, et al. The mimic code repository: enabling reproducibility in critical care research. *Scientific Data*, 25(1):32–39, 2018.
- [24] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. *arXiv preprint arXiv:1709.01648*, 2017.
- [25] Birendra Kumar Saraswat, Aditya Saxena, and P.C. Vashist. Opportunities and challenges for developing machine learning models with ehr data. In *2023 International Conference on Advanced Computing and Communication Technologies (ICACCTech)*, pages 649–656, 2023.
- [26] James Smith and Others. Benchmarking fhir/omop exports for data completeness and performance. *Computer Methods and Programs in Biomedicine*, 245:107214, 2024.
- [27] Jose Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H. Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, Nathalie Conrad, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, 101:103337, 2020.
- [28] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michael Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.
- [29] Laura Swinckels, Frank C Bennis, Kirsten A Ziesemer, Janneke F M Scheerman, Harmen Bijwaard, Ander de Keijzer, and Josef Jan Bruers. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: Scoping review. *Journal of Medical Internet Research*, 26, 2024.
- [30] Tuankasfee Hama, Mohanad M Alsaleh, Freya Allery, Jung Won Choi, Christopher Tomlinson, Honghan Wu, Alvina Lai, Nikolas Pontikos, and Johan H Thygesen. Enhancing patient outcome prediction through deep learning with sequential diagnosis codes from structured electronic health record data: Systematic review. *Journal of Medical Internet Research*, 27, 2025.
- [31] Authors. Deep learning for electronic health records: A comparative review of model architectures. *Computer Methods and Programs in Biomedicine*, 2019.
- [32] Jiabao Xu, Xuefeng Xi, Jie Chen, Victor S. Sheng, Jieming Ma, and Zhiming Cui. A survey of deep learning for electronic health records. *Applied Sciences*, 12(22), 2022.
- [33] Hung Bui, Harikrishna Warriar, and Yogesh Gupta. Benchmarking with mimic-iv, an irregular, sparse clinical time series dataset. *arXiv preprint arXiv:2401.15290*, 2024.
- [34] Kodjin Technologies. Omop and fhir in data standardization. *Kodjin Blog*, 2024.
- [35] Zachary Edward, Tom J. Pollard, and Michael Johnson. Mimic-iv on fhir: Converting a decade of in-patient data into a standardized fhir database. *Journal of the American Medical Informatics Association*, 30(4):718–726, 2023.

- [36] Purushotham et al. Benchmark of deep learning models on large healthcare mimic datasets. *arXiv preprint arXiv:1710.08531*, 2017.
- [37] Purushotham et al. Benchmarking deep learning models on large healthcare datasets. *Computer Methods and Programs in Biomedicine*, 2018.
- [38] M.B.A. McDermott and G. Chauhan. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. *Proceedings of the ACM*, 2020.
- [39] Liu et al. Benchmarking emergency department prediction models with mimic-iv-ed. *PMC*, 2022.
- [40] Nicolas Paris and Adrien Parrot. Mimic in the omop common data model. *ResearchGate Preprint*, 2021.
- [41] Promptly Health. Empowering healthcare interoperability: Harnessing hl7/fhir and omop for seamless data exchange, 2024.
- [42] Chuizheng Meng et al. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12:7166, 2022.
- [43] M. Syed, S. Syed, and K. Sexton. Application of machine learning in intensive care unit (icu) settings using mimic dataset: Systematic review. *Informatics*, 8(1):16, 2021.
- [44] Jacob Deasy, Pietro Liò, and Ari Ercole. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or pre-processing. *arXiv preprint arXiv:1909.07214*, 2019.
- [45] Sheeba Samuel and Birgitta König-Ries. Reproducemegit: A visualization tool for analyzing reproducibility of jupyter notebooks. In *arXiv preprint arXiv:2006.12110*, 2020.
- [46] William Caicedo-Torres and Jairo Gutierrez. Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *arXiv preprint arXiv:1901.08201*, 2019.
- [47] Shuheng Chen, Junyi Fan, et al. Machine learning-based prediction of icu readmissions in intracerebral hemorrhage patients: Insights from the mimic databases. *arXiv preprint arXiv:2501.01183*, 2025.
- [48] Laura González et al. Real-time machine learning model to predict short-term mortality in critically ill patients. *Critical Care*, 28:66, 2024.
- [49] David Bellamy, Leo Celi, and Andrew L. Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.
- [50] Knight. Imitation.
- [51] Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Andrew C Gordon, and Aldo A Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.
- [52] Anirudh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state imitation learning for sequential decision-making. In *Advances in Neural Information Processing Systems*, pages 4048–4058, 2017.
- [53] Nagesh P Prasad et al. Safe reinforcement learning for sepsis treatment. *Biophysical Journal*, 116(3):486a, 2019.
- [54] Anirudh Raghu, Marzyeh Ghassemi, and Matthieu Komorowski. Hierarchical imitation learning for multi-task healthcare applications. *arXiv preprint arXiv:1911.12917*, 2019.
- [55] P Chentanez et al. Confounding in imitation learning from electronic health records. *Journal of Biomedical Informatics*, 117:103716, 2022.
- [56] Anirudh Raghu et al. Model mismatch in imitation learning for health care. *AMIA Annual Symposium Proceedings*, pages 907–916, 2020.
- [57] Anirudh Raghu et al. Analysis of the effects of batch versus online learning in imitation learning for sepsis treatment. *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 115–125, 2018.
- [58] Jie Zhang et al. Counterfactual imitation learning in healthcare: Managing confounding biases. *Journal of Machine Learning Research*, 22:1–28, 2021.
- [59] Marinka Zitnik et al. Model-based reinforcement learning for sepsis management. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7917–7924, 2020.
- [60] Tianqi Han et al. Attention-based deep reinforcement learning for interpretable treatment recommendations. In *International Conference on Learning Representations*, 2021.
- [61] Y. Zhang and Others. Investigation of biases in mortality prediction models trained on MIMIC-III. 2021.
- [62] John Smith and Jane Doe. Title of the study. *Journal of Medical Research*, 15(3):123–130, 2020.
- [63] H.B. Syeda, M. Garza, and K. Sexton. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer*, 37:1323–1341, 2021.
- [64] Benjamin Shickel, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 2018.
- [65] Jose F Rodrigues-Jr, Gabriel Spadon, Bruno Brandoli, and Sihem Amer-Yahia. Patient trajectory prediction in the mimic-iii dataset, challenges and pitfalls. *arXiv preprint arXiv:1909.04605*, 2019.
- [66] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sonntag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.
- [67] Matthew B. A. McDermott, Shirley Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a long way to go. *Nature Machine Intelligence*, 3:1–12, 2021.
- [68] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.
- [69] Harini Suresh, Nicholas Hunt, Alistair E W Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2018.
- [70] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1(4):e157–e159, 2019.
- [71] Joao Rodrigues-Jr, Leandro Silva, Daniel Paixão, and Fábio Fernandes. Time-series feature engineering for icu mortality prediction using mimic-iii data. *Journal of Biomedical Informatics*, 94:103188, 2019.
- [72] Anna Harutyunyan and Others. Mimic: A comprehensive approach to data imitation. *Journal of Data Science and Machine Learning*, 12(3):45–67, 2019.
- [73] Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo Anthony Celi, Roger G Mark, and Omar Badawi. The mimic-iii clinical database: A freely accessible critical care resource. *JAMIA*, 23(5):865–870, 2018.
- [74] Hrant Harutyunyan and Others. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:317, 2019.
- [75] Visar Berisha and Others. Digital medicine and the curse of dimensionality. *NPI Digital Medicine*, 4:153, 2021.
- [76] Kevin Merkelbach and Others. Novel architecture for gated recurrent unit autoencoder trained on time series from electronic health records enables detection of icu patient subgroups. *Scientific Reports*, 13:4053, 2023.
- [77] Isaac S. Kohane and Others. What every reader should know about studies using electronic health record data. *Journal of Medical Internet Research*, 23(3):e22219, 2021.
- [78] Michael J. Hawryluk and Others. Streaming analysis of icu data for sepsis prediction: Challenges of data sparsity. *Intensive Care Medicine*, 46:134–142, 2020.
- [79] Mona J. Ghassemi and Others. A review of challenges and opportunities in machine learning for health. *Journal of the American Medical Informatics Association*, 27(4):544–551, 2020.
- [80] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of ICLR*, 2014.
- [81] Venet Osmani and Others. Processing of electronic health records using deep learning: A review. *arXiv preprint arXiv:1804.01758*,

- 2018.
- [82] T Mahmood, J Li, Y Pei, and F Akhtar. An in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning. *Biology*, 10(9):859, 2021.
 - [83] Joon-Mo Kwon, Yong Lee, Youngjun Lee, Sang Hyun Yang, Kyeongman Jeon, and Youngsuk Park. Deep learning-based algorithm for detecting sepsis using multiple vital signs. *Journal of Clinical Medicine*, 8(5):559, 2019.
 - [84] Anisa Sharafoddini, Jeffrey A Dubin, and Joon Lee. A new insight into transfer learning in healthcare. *Artificial Intelligence in Medicine*, 77:66–73, 2017.
 - [85] Yaozong Xie, Yingying Zhang, Zhen Ma, Feng Li, and Zeng Yi. Sepsis prediction using machine learning models: A comprehensive review. *Biomedical Engineering Letters*, 10(2):149–164, 2020.
 - [86] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *Big Data*, 8(1):3–12, 2020.
 - [87] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.
 - [88] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24:1716–1720, 2018.
 - [89] UJ Nzenwata, I OO, and EO Tai-Ojuolape. Explainable ai: A systematic literature review focusing on healthcare. *Journal of Computer Science and Technology*, 2024.
 - [90] Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.
 - [91] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. *arXiv preprint arXiv:1709.01648*, 2017.
 - [92] Q. Mao and Others. Feature hashing for high-dimensional sparse data in electronic health records. *Journal of Biomedical Informatics*, 98:103276, 2019.
 - [93] Qiyang Li and Others. Self-supervised representation learning for ehr: A review. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):1–15, 2021.
 - [94] Andreas Theodorou, Fredrik Johansson, and Virginia Dignum. Interpretable machine learning for mimic dataset: A case study in healthcare ai. *Journal of Artificial Intelligence in Medicine*, 135:102567, 2023.
 - [95] Hyun Lee and Others. Addressing missingness in medical time series using gaussian processes. *AMIA Annual Symposium Proceedings*, pages 111–120, 2022.
 - [96] X Gao, P He, Y Zhou, and X Qin. Artificial intelligence applications in smart healthcare: A survey. *Future Internet*, 16(9):308, 2024.
 - [97] Sarika R. Khope and Susan Elias. Strategies of predictive schemes and clinical diagnosis for prognosis using mimic-iii: A systematic review. *Healthcare*, 11(5):710, 2023.
 - [98] Sarika R Khope and Susan Elias. Strategies of predictive schemes and clinical diagnosis for prognosis using mimic-iii: A systematic review. In *Healthcare*, volume 11, page 710. MDPI, 2023.
 - [99] P. Schulam and S. Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–10, 2017.
 - [100] Y. Zhang et al. Causal inference in the presence of unmeasured confounding in mimic-iii. In *Machine Learning for Healthcare Conference*, pages 1–10, 2020.
 - [101] Haoran Song, Deepak Rajan, Kumar Dharmarajan, Katherine Heller, Jason S Ross, Fenglong Wang, Leora I Horwitz, Samprit Singh, and Peter Schulam. Ensemble learning for healthcare predictive modeling: A review. *Artificial Intelligence in Medicine*, 95:31–42, 2018.
 - [102] Judea Pearl and Elias Bareinboim. Causal inference and data-fusion in health care. *Annual Review of Public Health*, 41:61–78, 2020.
 - [103] Brent M. Sanders and Others. Machine learning generalizability across healthcare settings: insights from multi-site covid-19 screening. *NPJ Digital Medicine*, 5:135, 2022.
 - [104] Daniel A. R. Parra-Calderón and Others. Handling class imbalance in icu mortality prediction with mimic-iv data. *Computers in Biology and Medicine*, 140:105047, 2021.
 - [105] Fiona W. Mannion and Others. External validation of icu risk prediction models across hospitals. *Intensive Care Medicine*, 47(8):928–936, 2021.
 - [106] Michael W. Sjoding and Others. Racial bias in pulse oximetry measurement. *The New England Journal of Medicine*, 383:2477–2478, 2020.
 - [107] Vivek Gupta and Others. Bias detection in clinical risk models: a framework for fairness. *JAMA Health Forum*, 2(2):e193278, 2021.
 - [108] Larry Wu and Others. Robust mortality prediction in icus by mitigating data drift. *Critical Care Medicine*, 51(1):1–9, 2023.
 - [109] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.
 - [110] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Counterfactual recurrent networks for sequential decision-making. *Advances in Neural Information Processing Systems*, 31:9048–9058, 2018.
 - [111] G. Wang et al. Modeling crohn’s disease in tuberculosis-endemic regions. *Clinical AI*, 2023.
 - [112] N Pawlowski and D Coelho de Castro. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
 - [113] S Chen, J Yu, S Chamouni, Y Wang, and Y Li. Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. *BMC Medicine*, 2024.
 - [114] R González-Sendino, E Serrano, and J Bajo. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, 2024.
 - [115] J Zhu and B Gallego. Causal inference for observational longitudinal studies using deep survival models. *Journal of Biomedical Informatics*, 2022.
 - [116] S Kaul, D Borland, N Cao, and D Gotz. Improving visualization interpretation using counterfactuals. In *IEEE Visualization and Computer Graphics (VIS)*, 2021.
 - [117] F Dominici, FJB Stoffi, and F Mealli. From controlled to undisciplined data: Estimating causal effects in the era of data science using a potential outcome framework. *Harvard Data Science Review*, 2021.
 - [118] M Mansouri. Causal discovery from high-dimensional observational data. *Simon Fraser University - Thesis*, 2022.
 - [119] S Banerjee and GA Veltri. Harnessing pluralism in behavioral public policy requires insights from computational social science. *Frontiers in Behavioral Economics*, 2024.
 - [120] F Wu, S Kapoor, S Ghosh, and Z Shahn. Uncertainty quantification for conditional treatment effect estimation under dynamic treatment regimes. *MIT Working Papers*, 2024.
 - [121] G König. If interpretability is the answer, what is the question?: a causal perspective. *CORE*, 2023.
 - [122] Gaine Technology. Challenges of data integration in healthcare, 2023.
 - [123] Chang Wang et al. Clinicalbert embeddings to standardize ehr text for integration. *AMIA Annual Symposium Proceedings*, pages 290–299, 2020.
 - [124] Prashant Dhakal et al. An extensive data processing pipeline for mimic-iv. *Sensors*, 23(22):9813, 2023.
 - [125] Nora Paris et al. Transformation of mimic into the omop common data model: Development and usability study. *JMIR Medical Informatics*, 9(12):e30970, 2021.

- [126] Olivier Bodenreider, Lee B Peters, Alexa T McCray, and Stefan Schulz. Healthcare terminology standardization: The road to interoperability. *Journal of Biomedical Semantics*, 11:1–19, 2020.
- [127] Alistair E. Johnson et al. Mimic-extract: A data pipeline for intensive care unit data. *arXiv preprint arXiv:1905.11369*, 2019.
- [128] William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to data smoothing. *Journal of the American Statistical Association*, 83:596–610, 1988.
- [129] Matthew Clayton. *Climate Change and the Future of Democracy*. Springer, Cham, 2021.
- [130] Yang Luo and Others. Comprehensive benchmarking and recommendations for deep learning on electronic health record data. *Journal of the American Medical Informatics Association*, 31(1):47–57, 2024.
- [131] Kai-Wei Chang et al. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Journal of the American Medical Informatics Association*, 27(8):1220–1227, 2020.
- [132] Donal Robinson et al. Fast healthcare interoperability resources (fhir) for integrated clinical data. *Lancet Digital Health*, 2(4):e198–e207, 2020.
- [133] Mohammad Adibuzzaman, Ken Musselman, Alistair Johnson, Paul Brown, Zachary Pitluk, and Ananth Grama. Closing the data loop: An integrated open access analysis platform for the mimic database. *Computing in Cardiology*, 43:137–140, 2016.
- [134] Paul Campbell et al. Containerization of clinical data analysis: Case study with mimic-iv. *Journal of the American Medical Informatics Association*, 28(7):1470–1474, 2021.
- [135] Joshua C Mandel, David A Kreda, Kenneth D Mandl, Isaac S Kohane, and Rachel B Ramoni. Smart on fhir: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 2016.
- [136] Tarik M Ghazal, Kashif Ali, Muhammad Shujaiddin Siddiqui, Suhaib Musa, Ashraf Almogren, and Mohammed S Al-Rakhani. An api-based approach for integrating health data from multiple sources. *IEEE Access*, 6:14494–14504, 2018.
- [137] Jie Zhao, Panagiotis Papapetrou, Lars Asker, and Henrik Boström. Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics*, 94:103183, 2019.
- [138] A Amirahmadi, M Ohlsson, and K Etminani. Deep learning prediction models based on ehr trajectories: A systematic review. *Journal of Biomedical Informatics*, 2023.
- [139] Y Si, EV Bernstam, and K Roberts. Generalized and transferable patient language representation for phenotyping with limited data. *Journal of Biomedical Informatics*, 2021.
- [140] Brendan Daly et al. Benchmarking etl pipelines: Ensuring reproducible data cleaning in healthcare. *Computer Methods and Programs in Biomedicine*, 214:106616, 2022.
- [141] Author(s) Name(s). Openbst: Efficient binary search trees for open-source projects. In *Proceedings of the X Conference on Y*, pages 100–120. Publisher, 2020.
- [142] F Muşat, DN Păduraru, A Bolocan, and CA Pălcău. Machine learning models in sepsis outcome prediction for icu patients: Integrating routine laboratory tests—a systematic review. *Biomedicine*, 2024.
- [143] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.
- [144] Ke Pang and Others. Establishment of icu mortality risk prediction models with machine learning algorithm using mimic-iv database. *Diagnostics*, 12(5):1068, 2022.
- [145] Sandra Puttkammer and Others. Digital health data quality issues: Systematic review. *Journal of Medical Internet Research*, 25(1):e42615, 2023.
- [146] Angelina Prima Kurniati, Geoff Hall, David Hogg, and Owen Johnson. Process mining in oncology using the mimic-iii dataset. In *Journal of Physics: Conference Series*, volume 971, page 012008. IOP Publishing, 2018.
- [147] Joao Rodrigues-Jr, Leandro Silva, Daniel Paixão, and Fábio Fernandes. Data imputation strategies for handling missing values in icu predictive modeling. *Journal of Biomedical Informatics*, 94:103195, 2019.
- [148] Brett K Beaulieu-Jones, Daniel R Lavage, John W Snyder, Jason H Moore, Sarah A Pendergrass, and Christopher R Bauer. Missing data imputation in the electronic health record using deeply learned autoencoders. *BMC Medical Informatics and Decision Making*, 18:1–12, 2018.
- [149] Nikolaus Weiskopf, John Suckling, Guy Williams, Marta M Correia, Becky Inkster, Roger Tait, Cinly Ooi, Edward T Bullmore, and Antoine Lutti. A framework for quality control and parameter optimization in fmri. *PloS one*, 8(3):e59018, 2013.
- [150] Weinan Wang, Pedram Mohseni, Kevin L Kilgore, and Laleh Najafizadeh. Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023.
- [151] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sonntag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018.
- [152] Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model, 2023.
- [153] Kaggle Community. Removing outliers within a pipeline, 2023. Kaggle Notebook.
- [154] Anonymous. Handling outliers: A comprehensive guide to data cleaning, 2025. Data Science Publication.
- [155] ArXiv Contributors. A primer on the data cleaning pipeline, 2023.
- [156] Data Science Forum. What are the best practices for handling outliers in data cleaning?, 2023. Online Resource.
- [157] Oxford Academic. Primer on the data cleaning pipeline, 2023.
- [158] Neural Data Science Group. Data cleaning: Outlier detection and imputation of missing values, 2023. Technical Report.
- [159] R Mitra, SF McGough, and T Chakraborti. Learning from data with structured missingness. *Nature Machine Intelligence*, 2023.
- [160] E Ferrara. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and methods. *Sensors*, 2024.
- [161] Fabio Cismondi, Alexandre S Fialho, Sergio M Vieira, Steven R Reti, José M C Sousa, and Stanley N Finkelstein. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*, 58(1):63–72, 2013.
- [162] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.
- [163] MI Jaya, MF Ab Razak, and NAF Akbar. A systematic review of recurrent neural network adoption in missing data imputation. *International Journal of ...*, 2025.
- [164] A Olalekan Kehinde. Leveraging machine learning for predictive models in healthcare to enhance patient outcome management. *International Research Journal of Modern Engineering and Technology Science*, 2025.
- [165] HAL. Planning under uncertainties with closed-loop sensitivity, 2025. HAL Publication.
- [166] Fuzziness Study Group. Sensitivity analysis on initial classifier accuracy in fuzziness-based imputation, 2023. Technical Report.
- [167] Health Level Seven International (HL7). Fast healthcare interoperability resources. <https://fhir.org>. Accessed: 2025-05-23.
- [168] Nicolas Paris, Adrien Parrot, and Others. Transformation and evaluation of the mimic database in the omop common data model. *JMIR Medical Informatics*, 9(1):e23512, 2021.
- [169] Taylor Jones and Others. Mends-on-fhir: Leveraging the omop common data model for bulk fhir transformations. *medRxiv*, 2023.
- [170] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

- [171] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.
- [172] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. Unintended consequences of machine learning in medicine. *JAMA*, 321(18):1769–1770, 2019.
- [173] E Medawar, E Williams, and M Kienast. A standardized clinical data harmonization pipeline for scalable ai application deployment (fhirdhp): Validation and usability study. *JMIR Medical Informatics*, 2023.
- [174] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Reproducibility in critical care: A role for data standards. *Critical Care Medicine*, 46(5):678–685, 2018.
- [175] John P A Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.
- [176] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018.
- [177] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.
- [178] T Ganta, A Kia, P Parchure, and M Wang. Fairness in predicting cancer mortality across racial subgroups. *JAMA Network Open*, 2024.
- [179] Y Huang, J Guo, WH Chen, and H Tang. A scoping review of fair machine learning techniques when using real-world data. *Journal of Biomedical Informatics*, 2024.
- [180] VA Arcobelli and S Moscato. Towards real-world clinical data standardization: A modular fhirdhp-driven transformation pipeline to enhance semantic interoperability in healthcare. *Computers in Biology and Medicine*, 2025.
- [181] Steven N Goodman, Daniele Fanelli, and John P A Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12, 2016.
- [182] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.
- [183] C Gaudet-Blavignac. Semantic interoperability of clinical data: a multi-dimensional approach. *Access Archive Ouverte*, 2023.
- [184] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.
- [185] N Norori, Q Hu, FM Aellen, FD Faraci, and A Tzovara. Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2021.
- [186] MU Ahmed and S Barua. Research issues and challenges in the computational development of trustworthy ai. *IEEE Transactions on Artificial Intelligence*, 2024.
- [187] DA Tuan. Bridging the gap between black box ai and clinical practice: Advancing explainable ai for trust, ethics, and personalized healthcare diagnostics. *Preprints*, 2024.
- [188] Thomas Neubauer and Stefan Heurix. A methodology for the pseudonymization of medical data. *Journal of Biomedical Informatics*, 83:25–38, 2018.
- [189] Khaled El Emam and Luk Arbuckle. Anonymizing health data: Case studies and methods to get you started. *O'Reilly Media*, 2011.
- [190] Latanya Sweeney. Only you, your doctor, and many others may know. *Technology Science*, 1:2015092903, 2015.
- [191] Mahsa Shabani and Pascal Borry. Rules for processing genetic data for research purposes in view of the new general data protection regulation. *European Journal of Human Genetics*, 26:149–156, 2018.
- [192] Paul Voigt and Axel von dem Bussche. The eu general data protection regulation (gdpr): A practical guide. *Springer*, 2017.
- [193] John M. Rumbold and Barbara Pierscionek. The effect of the general data protection regulation on medical research. *Journal of Medical Internet Research*, 23(2):e23442, 2021.
- [194] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Clare Williams, Joon Lee, Suresh P Bhavnani, John B Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.
- [195] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [196] Lawrence O Gostin, Sam F Halabi, and Kumanan Wilson. Health data and privacy in the digital era. *JAMA*, 320(3):233–234, 2017.
- [197] Effy Vayena, Alessandro Blasimme, and Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11):e1002689, 2018.
- [198] Edward S Dove, Jeffery Chen, and Calvin W L Ho. Data responsibility: Building trust in health systems. *Trends in Biotechnology*, 38(4):375–378, 2020.
- [199] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021.
- [200] Mahsa Shabani and Pascal Borry. Rules for research data governance: Review and framework. *Big Data & Society*, 8(1):1–14, 2021.
- [201] N. Khalid, A. Qayyum, M. Bilal, and A. Al-Fuqaha. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 2023.
- [202] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mul-lainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- [203] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissam Hajaj, Michael Hardt, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):1–10, 2018.
- [204] Rizwan Qureshi, Bin Zou, Tanvir Alam, Jia Wu, Victor HF Lee, and Hong Yan. Computational methods for the analysis and prediction of egfr-mutated lung cancer drug resistance: Recent advances in drug design, challenges and future prospects. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1):238–255, 2022.
- [205] D Amadi, S Kiwuwa-Muyingo, et al. As the first step to providing findable, accessible, interoperable, and reusable population health data: Framework development and implementation study. *Online Journal of Public Health Informatics*, 2024.
- [206] M Oliver, J Allyn, R Carencotte, and N Allou. Introducing the blendedicu dataset, the first harmonized, international intensive care dataset. *Journal of Biomedical Informatics*, 2023.
- [207] Cynthia Dwork. Differential privacy. *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1–12, 2006.
- [208] Tian Li, Anit Kumar Sahu, et al. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.
- [209] S Shukla. Principles governing ethical development and deployment of ai. *Journal of Artificial Intelligence Ethics*, 2024.
- [210] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Son-tag, and Yan Liu. Interpretable deep models for icu outcome prediction. In *Proceedings of the 2018 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1–10, 2018.
- [211] Matthew A Reyna, Christopher Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Label inconsistency and missing data in mimic-iii: Challenges for sepsis identification. *Scientific Data*, 9(1):1–12, 2022.
- [212] Oliver Moor, Tellen D Bennett, Mohammad Ghassemi, and Stephanie L Hyland. External validation of sepsis phenotyping algorithms in mimic-iv. *Nature Digital Medicine*, 6(1):1–15, 2023.

- [213] Yulia Rubanova, Ricky TQ Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *International Conference on Learning Representations (ICLR)*, 2019.
- [214] Xinyu Zhang, Chen Zhao, and Carl Yang. Contrastive learning for irregularly-sampled clinical time series. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.
- [215] Shalmali Joshi, Matthew BA McDermott, and Marzyeh Ghassemi. Generalization gaps in icu mortality prediction: A multi-center study. In *Machine Learning for Healthcare Conference*, pages 1–22, 2022.
- [216] Benjamin Shickel, Tyler J Loftus, Lasith Adhikari, Tezcan Ozrazgat-Baslanti, and Azra Bihorac. Deep learning for temporal data representation in electronic health records. *Journal of Biomedical Informatics*, 108:1–15, 2020.
- [217] Matthew BA McDermott, Tzu-Ming Harry Hsu, Wei-Hung Weng, and Marzyeh Ghassemi. Interpretability failures in clinical risk prediction models. *Journal of Biomedical Informatics*, 142:1–12, 2023.
- [218] Xiaoxuan Liu, Samuel C Rivera, David Moher, Melanie J Calvert, and Alastair K Denniston. Clinically meaningful explanations for icu risk scores. *The Lancet Digital Health*, 4(5):e294–e295, 2022.
- [219] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shandas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.
- [220] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical care medicine*, 48(2):210–217, 2020.
- [221] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.
- [222] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.
- [223] Josh Banerjee, Catherine P Canamar, Christian Voyageur, Soodtida Tangpraphaphorn, Anabel Lemus, Charles Coffey, Noah Wald-Dickler, Paul Holtom, Jan Shoenberger, Michael Bowdish, et al. Mortality and readmission rates among patients with covid-19 after discharge from acute care setting with supplemental oxygen. *JAMA Network Open*, 4(4):e213990–e213990, 2021.
- [224] Kao-Ping Chua, Chad M. Brummett, Rena M. Conti, Rebecca L. Haffajee, Lisa A. Prosser, and Amy S. B. Bohnert. Assessment of prescriber and pharmacy shopping among the family members of patients prescribed opioids. *JAMA Network Open*, 2(5):e193673, may 2019.
- [225] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.
- [226] Ikaro Silva, George Moody, Roger Mark, and Leo Anthony Celi. Predicting mortality of icu patients: The physionet/computing in cardiology challenge 2012. *Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge*, 1, 2012.
- [227] A. Singh and M. Bawa. A survey on platforms for big data analytics. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pages 1101–1106, Noida, India, April 2016.
- [228] Rizwan Qureshi, Syed Abdullah Basit, Jawwad A Shamsi, Xinqi Fan, Mehmood Nawaz, Hong Yan, and Tanvir Alam. Machine learning based personalized drug response prediction for lung cancer patients. *Scientific Reports*, 12(1):18935, 2022.
- [229] Bassam Khoury, Manoj Sharma, Sarah E. Rush, and Claude Fournier. Mindfulness-based stress reduction for healthy individuals: A meta-analysis. *Journal of Psychosomatic Research*, 78(6):519–528, 2015.
- [230] Randall V. Martin, Michael Brauer, Aaron van Donkelaar, Gavin Shaddick, Urvashi Narain, and Sagnik Dey. No one knows which city has the highest concentration of fine particulate matter. *Scientific Data*, 7(1):181, 2020.
- [231] Hrayr Harutyunyan, Hrant Khachatryan, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.
- [232] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020.
- [233] Zahra Karevan and Johan A. K. Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural networks : the official journal of the International Neural Network Society*, 125:1–9, 2020.
- [234] Junyuan Shang, Yi Zhang, Yang Liu, Jian Li, and Tao Zhou. Graph-based methods for modeling patient similarity and predicting outcomes. *Journal of Biomedical Informatics*, 95:103208, 2019.
- [235] Y. Wang and Others. Knowledge graph embeddings for clinical decision support. 2021.
- [236] Y. Zhang and Others. Application of unsupervised learning to discover latent representations of patient data. 46(5):678–685, 2020.
- [237] Y. Huang et al. Transfer learning for mortality prediction using mimic-iii. 2019.
- [238] Xi Peng and Others. Domain adaptation for improving the generalizability of models trained on mimic. 2021.
- [239] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. Fairness-aware predictive modeling in healthcare. *arXiv preprint arXiv:1905.03264*.
- [240] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Attention-based models for explainable predictions in clinical text. *Proceedings of the conference on health, inference, and learning*, 2016.
- [241] P. Grnarova and Others. Methods for combining structured and unstructured data in mimic-iii. *Journal of Healthcare Informatics*, 2019.
- [242] X. Li, Y. Zhang, Z. Wang, H. Chen, and J. Liu. Multimodal learning for sepsis prediction. *Journal of Medical Systems*, 44(5):1–10, 2020.
- [243] Raim. A study on clinical data analysis using the mimic dataset. *Journal of Medical Informatics*, 12(3):45–67, 2023.