A Scalable Data Engineering Approach to AI-Powered Health Risk Prognosis

Michael Anderson
Asst Professor

Abstract

Predictive healthcare analytics aims to anticipate future clinical events in the form of predictions from individual patients or group cohorts by devising and applying sophisticated analytical models to relevant data sets. The supporting data engineering enables the acquisition of ready-to-use, progress-ready data that meet the quality, temporal, and licensing requirements of analytics and predictive models. Multi-cloud data engineering tools and services play a crucial role in this context and can be fully exploited to develop a cloud-agnostic solution for the central data engineering life cycle. Such an approach helps to ingest, organize, and harmonize data from multiple sources and supports both the descriptive and the predictive phases of healthcare analytics.

The use of Multi-Cloud environments, defined as cloud solutions that combine services from multiple providers belonging to different service models and technology stacks, reduces the risk of vendor lock-in while providing access to the best data engineering cloud capability for a specific use case, promoting the idea of Cloud for Data Engineering. The subsequently applied model development and deployment — typically concentrated in dedicated environments within the cloud — focus on predictive healthcare analytics, which aims to anticipate clinical conditions. Risk stratification models derived from these analytical methodologies are intended to provide clinicians and the healthcare organization with useful information for subsequent decision-making and care activities.

Keywords : Predictive analysis, healthcare, multi-cloud computing, distributed data engineering, privacy-preserving artificial intelligence.

<H1> Introduction

Healthcare problems, processes, and services generate incredibly large amounts of multi-faceted data. Data are abundant in all forms, ranging from Centre of Disease Control (CDC) health alert e-mails, who.int newsletters, World Bank financial statistics and digitalisation index data, computer simulations, frontline workers' observations, human behaviour, global output of COVID-19-vaccine research, social media postings about health threats and successful strategies, timely and untimely death statistics, and many more.

In a crisis situation, it is often impossible for an analyst to look at and assess all data at the same time. Therefore, Artificial Intelligence (AI) and Machine Learning (ML) technologies can be leveraged to predict healthcare outcomes. These technologies can also be useful for predictive modelling whenever external data sources are limited. Potential Deliverables can be formed by creating AI/ML models capable of delivering accurate results on validation datasets and available clinical expertise in comprehending them. If the models achieve good Enough performance on particular datasets, especially clinical trials, they can be deployed in practice.

However, building a reliable and predictive AI/ML model is challenging for several reasons. Lack of Data-data points in terms of cases, years of data availability and number of input features can affect the prediction quality. Thus, it is useful to create ensembles of multiple predictions based on prediction Diversity– prediction variation among models. Another approach involves sharing the burden of prediction across institutions: Federated Learning involves sharing the Model architectures and features among institutions while keeping the Data Private in each institution. Hence, if Data from Hospitals 1 and 2 are not sufficient for developing a useful Model, Data from Hospital 3 can be used for a different Model, and if Model 1 and Model 3 share Similar features, they can be fused to form a more reliable prediction. Further, if the Model result on a Limited validation Dataset is supported by ClinicalTrials.gov evidence, the model is more likely to be accurate.

<H2> Background and Significance

Predictive analytics for healthcare and other domains rely on large datasets and high-performance computing facilities for model training and prediction. While many computational models have been developed and tested using historical datasets, the availability of sufficient local data to support the training of reliable predictive models is often a limitation in real-world applications. Multi-cloud data engineering approaches can address these constraints by facilitating the combined use of disparate data sources while enabling the management and orchestration of model training in the cloud.

The analytics-as-a-service design can guide the process of developing and deploying sophisticated predictive models without requiring substantial local digital infrastructure. Sensitivity analyses can help select the most appropriate modelling technique based on the underlying data distribution rather than algorithmic complexity, with performance targets set according to local application needs. The approach enables historical datasets, even with missing features, to be analysed for model training and validation; predictive performance then assessed using local data from other sources.

Table 1: Major Components of Multi-Cloud Predictive Healthcare Analytics Framework

Component	Description	Purpose
Data Ingestion	Collects raw data from hospitals, EHRs, public datasets, IoT devices	Unified data acquisition
Data Transformation	Cleansing, normalization, feature engineering	Improve data quality
Data Storage	Data lakes, warehouses, feature stores	Scalable storage
Data Governance	Metadata, lineage, quality control	Trustworthy analytics
AI Model Training	ML/DL model development	Prediction
Federated Learning	Distributed model training without data sharing	Privacy
Model Deployment	REST APIs, cloud-native services	Clinical use
Validation & Benchmarking	Accuracy, AUC, F1, calibration	Reliability

<H1> Background and Rationale

Predictive healthcare analytics can enhance healthcare delivery and policy decision-making. While several AI models for clinical outcome prediction exist, the healthcare industry still encounters roadblocks when implementing

predictive models in clinical practice. These barriers are likely attributable to the lack of adequate data engineering processes to support the large-scale training, importance monitoring, and performance validation of predictive models across healthcare organizations. Overcoming this challenge entails addressing four main blocks: the establishment of data engineering processes suitable for the training of AI models; the identification of suitable AI model architectures for the predictive task at hand and their performance benchmarking against well-established reference models; the provision of AI model training in a federated setting, with privacy-preserving data-sharing schemes that ensure patient confidentiality, legal compliance, and respect for ethical concerns surrounding patient data; and the design and development of a complete end-to-end pipeline for the provisioning of use-case-specific predictive healthcare analytics.

Healthcare organizations typically use multiple cloud service providers for data storage and processing due to legal and policy rules that force them to store and process sensitive patient data in specific countries or regions. A Multi-Cloud setup allows a cloud-native design of Predictive Healthcare Analytics Data Engineering (PHADE) processes that encompass Data Ingestion, Transformation, Storage, Access, Feature Generation, Quality Control, Metadata Management, and Data Lineage functions. Such a specialized design enables Predictive Healthcare Analytics models to meet clinical requirements for predictive performance, robustness, and calibration and provide respective modules for uncertainty estimation, diagnosis, and explanation.

<H2> Research design

Three fundamental gaps serve as the foundation for the work. First, while AI-facilitated predictive analytics are gaining popularity in healthcare, they primarily operate at single-institution level using local data, contracts and infrastructure. Second, a paradigm shift towards a service-oriented, cloud-native data-engineering framework has been proposed more than a decade ago, yet the potential benefits have been neither acknowledged nor demonstrated. Third, there is a notable absence of cloud-agnostic data-engineering pipelines that span the full data life cycle from ingestion through storage to integration and preparation for AI-based analytics.

The theoretical framework supporting the work is constituted by the Digital, Service, and Cloud Computing theories, underpinning a multi-cloud data-engineering framework specifically designed for predictive healthcare analytics. The objective is to fulfil the aforementioned research gaps and investigate the predicted aspect of the health policy challenge in the presence of a lethal pandemic by using Multi-cloud Data Engineering and Intelligent AI Models for Predictive Healthcare Analytics in order to allow Analytical Information Forecasting in Multi-Institutional Research Projects During Pandemics.

<H1> Data Engineering on Multi-Cloud Platforms

A cloud-agnostic strategy for data ingestion, storage, and integration across different services of different providers lays the groundwork for building high-quality predictive healthcare analytics solutions in a multi-cloud environment. Data engineering pipelines covering data standardization, governance, and quality assurance further structure the process.

Deploying predictive pipelines for healthcare delivery demands engineering the underlying data at the infrastructure level. Data-Engineering Pipelines in a Cloud-Agnostic Multi-Cloud Configuration address these needs by covering ingestion, storage, transformation, and integration. Examined next are ingestion and storage strategies based on an overarching cloud-agnostic vision, along with data-standardization, governance, and quality-assurance processes, bolstered by provenance-engineering capabilities. Data pipelines and orchestration choices conclude the conversation.

<H2> Data Ingestion and Integration Across Clouds

Data-agnostic data ingestion processes are developed to facilitate the aggregation of raw data from multiple disparate sources and deployments within a federated multi-cloud environment. A cloud-native Data Mesh strategy is employed for data storage and integration in a cloud-agnostic manner. Cloud-agnostic Data Mesh data storage is integrated into a cloud-native orchestrated Data Mesh architecture that natively handles all data transformation, feature engineering, and machine learning model-training orchestration while preserving data locality and resident compliance with regulatory laws related to National, Regional, and Organizational Data Residency and Sovereignty, such as GDPR and HIPAA.

The orchestrated Data Mesh architecture is integrated with ModelDB and experiment tracking databases such as DVC, MLflow, and SAC, responsible for managing experiment tracking, model inputs and outputs along with their lineage, and the automatic instantiation on-demand of cloud-native Data Mesh pipelines for orchestrated data transformation and feature engineering when training models across the multi-cloud federated environment. The cloud-agnostic Data Mesh constructs the identity of a cloud-native Multi-Cloud Data Engineering service that natively preserves data residency.

Table 2: Predictive Model Types and Their Characteristics

Model Type	Strengths	Limitations
Logistic Regression	Simple, interpretable	Low non-linearity handling
Decision Tree	Easy visualization	Overfitting
Random Forest	High accuracy	Computational cost
SVM	Good for high dimensions	Hard to tune
Neural Networks	Complex pattern learning	Data hungry
Deep Learning	High predictive power	Black-box nature
Ensemble Models	Robust predictions	Increased complexity

<H3> Data Standardization, Governance, and Quality

Data standardization denotes a set of procedures designed to create a uniform structure for data, promoting consistency across heterogeneous datasets. Developing common schemas for diverse data sources requires much effort, as accessible datasets rarely follow the same convention for attribute naming or structuring. Efficient and reliable standardization is necessary to leverage the collected data for further analysis. Centralized governance is vital to data quality management within organizations; without it, stakeholders lose confidence in the data or spend excessive effort on cleansing and reconciling errors. Defining an automation strategy for quality checks is crucial to maintaining data freshness and reliability over time. A dedicated model evaluates the quality of nodes and edges in the predictive knowledge graph. Metadata management enhances the linking of datasets in feature engineering and model development, while preserving lineage helps track the origin and transformations of features in model training.

Initiatives such as the European Open Data Portal emphasize the importance of high-quality, well-governed open data and foster a culture of oral data within member states. Quality constraints for external data sources are commonly neglected in exploratory science, although poor-quality information can diminish the predictive power of machine-learning

models. In clinical settings, the same data-sharing agreements that safeguard sensitive patient information also prevent the sharing of potentially useful datasets—data-sharing initiatives can help overcome these hurdles, provided that quality quantification assures potential adopters of their reliability. One scalable solution uses multiple private hospitals to determine the quality and completeness of health data, enabling the safe exchange of high-quality datasets.

<H1> AI Models for Predictive Healthcare

Healthcare seeks to improve the prediction of life-threatening diseases. Relevant models should allow healthcare professionals to deliver timely treatment interventions and effective prevention programs to at-risk patients. Among the models that can assist any epidemiologist or healthcare analyst are predictive models that identify patients at risk of developing different diseases. Datasets encompassing multiple diseases, termed multi-label datasets, open two avenues for predictive modeling: independent training of a model per disease, thus exploiting the entire dataset without constraints, or joint training of several models on the same patients where classes are exclusive during prediction. Class balance, dataset size, and degree of similarity between diseases are considered when selecting the approach. These predictive models facilitate triggered preventive strategies in order to rescue a patient's life or reduce healthcare costs. Performance evaluation is crucial in the modeling process to determine how well the model is capable of predicting risk, and two aspects related to performance are hospital mortality and disease risk prediction.

Predictive modeling can be accomplished through several methodologies. In supervised learning, a predictive model is defined using supervised information, i.e., label information, through training. In characterizing a feature as predictive, the goal is to predict the value/behavior of the target under certain conditions. For supervised learning methods to make use of functional information of the physical process being evaluated, the size of the dataset needs to be large. Thus, it is recommended to use a vast dataset, being complete and representative in the training phase. A useful way to monitor a supervised learning model in a hospital is to validate the model every six months and re-train it when necessary. The considered feature space can be smaller if the model is trained as online learning. The evaluation parameters mainly concentrates on prediction sensitivity, specificity, accuracy and precision. Apart from it, factors like area under curve, error rate, and kappa index are also used to analyze the efficiency of some predictive models. These factors are grouped together to give a clear picture of the predictive model's operational power in predictive healthcare analysis.

<H2> Model Selection and Benchmarking

Predictive healthcare analytics can use different predictive models like Machine Learning, Deep Learning Models and many other models. The majority of the predictive models differ in their internal working logic and its operational support. All these predictive analytics models, can provide better predictive solutions to the healthcare industry. A majority of predictive models are diverse, in operational support, advantageous features and disadvantages in order to produce efficient predictive analysis, choose the right predictive model according to the clinical conditions, nature of the input dataset and predefined clinical objective. The performance efficiency of various predictive healthcare models implemented earlier can also help in providing the comparative outline to these predictive models. These comparative outlines can help in choosing the best informational predictive model. The best-performing predictive model can expect to produce an effective and useful predictive solution for the healthcare analytics problem. Based on the predictive performance efficiency of these models, a majority of healthcare analysts choose specific predictive models which can support healthcare organization to achieve healthcare analysis goal and also support the actual patients by providing efficient healthcare analysis operations.

In order to ameliorate patients' healthcare condition and life expectancy at the end, a specific part of healthcare practitioners, medication research and analysis organizations are engaged in predicting the possibility of the disease in a specific person. Such prediction depends on various internal and external factors. Each and every predictive healthcare analysis models are still stick on achieving the healthcare analysis goal. To provide an effective predictive healthcare analysis support for the clinical experiments, the output from predictive model playing a main role. According to it, these predictive models are examined with different document containing the output result. In this context, the performance of numerous healthcare activity problems is look into. Different set of parameters are employed in the majority of considered healthcare models to measure the output efficiency of that model.

<H3> Federated and Privacy-Preserving Learning

Compliance with healthcare regulations governing the use of sensitive or personal data for research purposes can significantly reduce the datasets available for the development of predictive analytics models. Given the necessity of large, diverse, and representative datasets for this task, the possibility of overcoming data-sharing restrictions by employing federated and privacy-preserving learning techniques has become a subject of active research in predictive analytics. The implementation of these techniques enables the training of machine learning or deep learning models without requiring the transfer of the original datasets from their private repositories. Consequently, healthcare providers can freely participate in research projects without the need to share their internal patient data. Moreover, potential patient data breaches or abuses are avoided.

In federated learning, a reference model is trained collaboratively by several nodes, each responsible for communicating locally learned model parameters instead of sharing their data samples. Recently developed privacy-preserving model training techniques further enhance privacy by adding a noise component derived from information theory to the locally learned parameters before their exchange.

Table 3: Evaluation Metrics Used in Predictive Healthcare Analytics

Metric	Meaning
Accuracy	Correct predictions / total
Precision	True positives / predicted positives
Recall (Sensitivity)	True positives / actual positives
F1-score	Harmonic mean of precision & recall
AUC-ROC	Class discrimination ability
AUC-PR	Performance on imbalanced data
Calibration	Reliability of probability estimates
Robustness	Stability across datasets

<H1> Architecture of an End-to-End Predictive Analytics Pipeline

The present stage of the proposed research can be characterized by the implementation of an end-to-end predictive analytics pipeline capable of handling a multi-cloud data infrastructure for scientific research governance. From a design standpoint, it integrates data ingestion, storage, and transformation subsystems featuring support from a cloud-agnostic feature store, automated feature engineering, and a cloud-native training pipeline for arbitrary machine-learning models. Orchestration of predictive model development enables an experiment tracker that keeps logs of the entire training process on the model level, allowing reproduction of model results and

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 01 Issue: 01
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 08 ACCEPTED: NOVEMBER 28 PUBLISHED: DECEMBER 12

management of ML workflows. Aspects of feature engineering (cleansing of features), model training, validation, and hyper-parameter optimization are wrapped into pipeline design patterns.

The multi-cloud setup allows multi-organization data sharing and enables sensitive data to remain in its organization by leveraging federated learning techniques. Privacy-preserving machine-learning methods make sure that, even if the data is not federated but comes from a single source, its sharing does not lead to breaches in patient confidentiality. These specialized frameworks are crucial for an efficient organization-level-level sharing of data between institutions because legal constrains appear when dealing with patients' information. Research teams conducting project-related work with several different institutions want to share data with the other institutions when the research context requires state-of-the-art ML-model performance on patient-related problems. Nevertheless, even for these institutional collaborations, some limits exist, which are closely related to the health system that governs that partnership.

<H2>. Data Ingestion and Feature Stores

An overview of data ingestion processes defines databases, data lake sets and data warehouses. A description of the data lineage system exhibits how training and testing dataset creation can be automated. Three supporting feature stores are presented as reusable assets for predictive health monitoring models.

Data ingestion consists of the means to insert data into a database, data lake or data warehouse. Data processes take raw data and transform that into data formats suitable for analytical reports. For databases and data warehouses, the suitable data format is mainly relational. For data lakes however, the raw data can also stay in its original format and can be batch loaded or stream loaded. When raw data is persistently stored in a data lake, data quality and consistency are ensured by routinely scheduled data processes. Once exploratory analysis and in-depth analytics are performed from a data lake, those data selections or explorations can also be persistently stored in a data warehouse.

Automating the creation of training and testing datasets for experiments with models from machine learning and deep learning is essential to ensure reproducible and comparable results. Such a system can take a metadata overview of the labelled dataset and the models for which predictions should be generated and select folders and files for both the input and output storage. The predictions are written in the structure required by the experiment tracking solution.

<H3>. Model Training Orchestration

Automated orchestration of AI-model training is key to making the analytics framework cloud-agnostic, reproducible, and innovative. Based on the Open Neural Network Exchange (ONNX) standard, football-like format that supports interoperability, visual tracking and comparison of model-training experiments is achieved through the use of the MLflow tool. An MLflow tracking server running in a K8s cluster can track experiments irrespective of the cloud environment in which different models are trained. Having a persistent storage mounted to the Ubuntu node running OpenShift enables saving artifacts, which could be any executables, logistic model graphics, or embeddings needed for prediction or inference. There is no need to set up ONNX Model Zoo, a centralized repository of models hosted by ONNX, since models can be downloaded automatically within the pipeline during the actual usage without requiring extra storage. Cloud-native design of MLflow allows deploying the wrapped models for REST predictions either on a local-hosting site or on a different cloud provider that is selected to make prediction. Any model compatible with the MLflow framework can easily plug into the pipeline for the predictions.

All cloud-native components are designed to support the K8s environment. In addition, deep Honorary Federated Learning performs centralized Differentiated Privacy and

decentralized PPO-CT model personalization with local Differential Privacy requirement satisfied at each hospital, allowing AI models to be privacy-preserving when divulging the prediction-table results to a different environment where no local data can be stored for AI-driven prediction or inference. The whole decentralized framework achieves privacy protection during the training of local models while allowing model sharing across institutions for performance improvement without violating the local data-sharing restriction.

<H1> Evaluation Metrics and Validation Frameworks

Predictive performance is typically quantified in terms of accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve, and area under the precision-recall curve. Calibration and discrimination are also crucial, as predictive systems must provide well-calibrated probabilities for secondary clinical decision-making. For example, a model predicting a 5% risk for a disease should be correct in 5 out of 100 cases. A further dimension of clinical utility is robustness against sampling variability and shifts in the data distribution. It should therefore be possible to confirm model performance through repeated cross-validation and, ideally, to validate with an independent external cohort.

These validation requirements motivate an architecture that facilitates rapid training of a wide range of predictive models under consistent conditions. A multi-cloud strategy widens the pool of federated learning participants while permitting any participant to deviate from the standard without affecting the whole ecosystem. Empirical tests can take advantage of multiplexed external cohorts to evaluate performance, calibration, and discrimination while ensuring sustained patient privacy and compliance with local regulations.

<H2> Predictive Performance Metrics

Predictive Healthcare Analytics Using Multi-Cloud Data Engineering and Intelligent AI Models

Evaluating the quality of a predictive model may encompass several quantitative facets intended to measure its performance in accurately forecasting outcomes, unlabeled events, and still-to-happen conditions that are crucial in supporting complex decision-making processes such as those applied in health-related problems. A subset of such metrics, predictive performance metrics (PPMs), aims at accounting for the predictive performance of a model on data, i.e., how good the model's predictions are. However, their use is usually limited to supervised learning models since they require a known ground-truth target feature in the validation data. They can be single or multi-dimensional in nature.

Moreover, modeling involves choosing the model type and corresponding hyperparameters iron out a so-called hypermodel to minimize a loss of desired characteristics while operating over a particular dataset. As such, benchmarking involves quantifying the quality of the minimization procedure, reporting on the degree to which a model achieves the desired operating characteristics such as accuracy over the dataset, and in so doing comparing two or more models using the same benchmark. Such richness of PPMs makes them the ideal candidates for benchmarking the predictive healthcare analytics application, where predictive performance constitutes a necessary and sufficient condition for health practice adoption.

Predictive calibration tests the degree to which a classifier estimates the probability of an event. For instance, when the estimated probability of survival, or of being in one outcome or another, equals the real observed frequency of survival, or of being in one outcome or another, the model is properly calibrated. Predictive discrimination quantifies how well the data split into two or more groups relating to a binary or multi-class outcome respectively. Predictive robustness tests the resiliency of a model, under small but important changes in the data, by measuring the effect of repeating the hold-out

data split on the model accuracy. Predictive reliability makes use of other independent datasets, acquired using different settings, to study how a model lingers or not with datasets that differ in origin. Cross-validation schemes and external advantages efficiently sharing data across healthcare entities while keeping patient privacy protected strenghten overall model validation.

<H1> Conclusion

Healthcare delivery and policy worldwide could be improved by predictive healthcare analytics that integrates multi-cloud data engineering and cloud-based artificial intelligence models. Data engineering follows a cloud-agnostic approach that supports data ingestion, storage, integration, and preparation in several publicly available cloud providers. Cloud-native orchestration copes seamlessly with heterogeneous data and the complexities of scaling resources according to demand, while data governance ensures a high level of quality. Predictive performance, calibration, and discrimination metrics, along with cross-validation schemes, offer quantitative guarantees of the ML models' reliability.

Numerous gaps in the predictive healthcare analytics literature remain. AI models used throughout the dataset possess no definitive ground truth and exhibit considerable variability because the data of any pair of hospitals is often nonoverlapping. Healthcare providers are typically apprehensive about sharing such sensitive information, even in centralized data-sharing environments. Nevertheless, facilitating data-sharing structures remains vital for predictive healthcare analytics. Multi-cloud social platforms as described here represent a promising direction for future research.

<H1> References

1. Artzi, N. S., Shilo, S., Hadar, E., Rossman, H., Barbash-Hazan, S., Ben-Haroush, A., Balicer, R. D., Feldman, B., Wiznitzer, A., & Segal, E. (2020). Prediction of gestational diabetes based on nationwide electronic health records. Nature Medicine, 26, 71–76.
2. Brom, H., Brooks Carthon, J. M., Ikeaba, U., & Chittams, J. (2020). Leveraging electronic health records and machine learning to tailor nursing care for patients at high risk for readmissions. Journal of Nursing Care Quality, 35(1), 27–33.
3. Mandair, D., Tiwari, P., Simon, S., Colborn, K. L., & Rosenberg, M. A. (2020). Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data. BMC Medical Informatics and Decision Making, 20, 252.
4. Martinez, D. A., Levin, S. R., Klein, E. Y., Parikh, C. R., Menez, S., Taylor, R. A., & Hinson, J. S. (2020). Early prediction of acute kidney injury in the emergency department with machine-learning methods applied to electronic health record data. Annals of Emergency Medicine, 76(4), 501–514.

5. Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Scientific Reports, 11, 1980.
6. Su, C., Aseltine, R., Doshi, R., Chen, K., Rogers, S. C., & Wang, F. (2020). Machine learning for suicide risk prediction in children and adolescents with electronic health records. Translational Psychiatry, 10, 413.
7. Tjandra, D., Ritchie, M. D., H. M., & colleagues. (2020). Cohort discovery and risk stratification for Alzheimer's disease: An electronic health record-based approach. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 6(1), e12035.
8. Ye, C., Li, J., Hao, S., Liu, M., Jin, H., Zheng, L., Xia, M., Jin, B., Zhu, C., Alfreds, S. T., Stearns, F., Kanov, L., Sylvester, K. G., Widen, E., McElhinney, D., & Ling, X. B. (2020). Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. International Journal of Medical Informatics, 137, 104105.
9. Steinberg, E., Jung, K., Fries, J. A., Corbin, C. K., Pfohl, S. R., & Shah, N. H. (2021). Language models are an effective representation learning technique for electronic health record data. Journal of Biomedical Informatics, 113, 103637.
10. Tomašev, N., Harris, N., Baur, S., Mottram, A., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Magliulo, V., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C. R., Osborne, T. F., Reeves, R., Hassabis, D., King, D., Suleyman, M., Back, T., Nielson, C., Seneviratne, M. G., Ledsam, J. R., & Mohamed, S. (2021). Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nature Protocols, 16, 2765–2787.
11. Feng, A., et al. (2021). A machine learning pipeline for accurate COVID-19 health outcome prediction using longitudinal electronic health records. Proceedings of the 2021 AMIA Annual Symposium, 448–456.
12. Yang, X., Bian, J., Alwhaibi, M., et al. (2021). Predicting patient outcomes using electronic health records and machine learning: A clinical data engineering perspective. Journal of Biomedical Informatics, 114, 103657.

13. Xie, F., Chakraborty, B., Ong, M. E. H., Goldstein, B. A., Liu, N., & others. (2022). Benchmarking emergency department prediction models with machine learning and public electronic health records. Scientific Data, 9, 658.
14. Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. Nature Medicine, 28, 1240–1248.
15. Sadasivuni, S., Saha, M., Bhatia, N., Banerjee, I., & Sanyal, A. (2022). Fusion of fully integrated analog machine learning classifier with electronic medical records for real-time prediction of sepsis onset. Scientific Reports, 12, 5711.
16. Zou, Y., Pesaranghader, A., Song, Z., Verma, A., Buckeridge, D. L., & Li, Y. (2022). Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. Scientific Reports, 12, 17868.
17. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., & Wu, Y. (2022). A large language model for electronic health records. npj Digital Medicine, 5, 194.
18. Krishnan, R., Rajpurkar, P., & Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering, 6, 1346–1352.
19. Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., & Liu, N. (2022). Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. Journal of Biomedical Informatics, 126, 103980.
20. Abraham, A., et al. (2022). Dense phenotyping from electronic health records enables machine learning-based prediction of preterm birth. BMC Medicine, 20, 355.
21. Cui, S., Wang, J., Gui, X., & Wang, T. (2022). AutoMed: Automated medical risk predictive modeling on electronic health records. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 948–953.
22. Hurst, W., Tekinerdogan, B., Alskaif, T., Boddy, A., & Shone, N. (2022). Securing electronic health records against insider-threats: A supervised machine learning approach. Smart Health, 26, 100354.
23. Xie, F., Zhou, J., Lee, J. W., et al. (2022). Benchmarking emergency department prediction models with machine learning

and public electronic health records. Scientific Data, 9, 658.

24. Barda, N., Riesel, D., Finkelstein, Y., et al. (2020). Developing a COVID-19 mortality risk prediction model when individual-level data are limited. Nature Medicine, 26, 1753–1759.
25. Khera, R., Haimovich, J., Hurley, N. C., et al. (2020). Use of machine learning models to identify patients at risk of adverse cardiovascular outcomes. JAMA Network Open, 3(10), e2026062.