GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

AI-Powered EHR Engineering for Precision Healthcare

Mallesham Goli
Independent Researcher
mallesham.goli.research01@gmail.com

Abstract

Research aims, methods, key results, and implications for scalable EHR analytics and precision care. Electronic Health Records (EHRs) are a historical, comprehensive, and continually updated source of health data for patients. Although EHRs have been widely used for retrospective population studies, research prototypes, and clinical decision support systems, the actual uptake of AI-driven technologies in clinical practice and patient care remains limited. Automatic ingestion of clinical data, standardized data lakes, and periodic model retraining are essential for seamless operation and scalable usage. The objectives of this research are to design data ingestion and preprocessing pipelines that run automatically, to support cloud- and on-premises-based clinical analytics, and to apply patient stratification and personalized treatment optimization methods.

The proposed technologies are scalable, reproducible, and support precision healthcare delivery, with a global public health perspective. The results of the clinical data engineering effort are publicly available. Reproducibility was achieved by adhering to open-source principles, data-sharing best practices, and the use of common data-collection and storage formats. Population analyses demonstrated predefined hypotheses in the domain of obesity and revealed new insights into atrial fibrillation. In acute care and emergency medicine, AI-supported rapid data ingestion and the clinical application of complex models during patient triage were successfully demonstrated. Although deployment-specific bias was identified in these specific domains, adequate technical control of both cloud-based and on-premises configurations was supported.

Keywords: EHR Analytics, Precision Care, Health Data, Data Ingestion, Data Pipelines, Data Preprocessing, Data Lakes, Cloud Systems, On Premise, Clinical Analytics, Patient Stratification, Treatment Optimization, Machine Learning, Model Retraining, Data Sharing, Reproducibility, Population Health, Atrial Fibrillation, Emergency Care, Clinical Decision.

1. Introduction

A popularized phrase from Informatics states that “All roads lead to EHR” still holds. Electronic Health Records (EHRs) contain archived and real-time data about patients’ clinical

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

profiles and health conditions. EHRs house administrative, clinical, lab, diagnostics, and imaging data. In addition, vital signs data from wearables—often seen as the next trend for patient monitoring—can feed into the EHR. Access to such a wide variety of data opens up new avenues for discovery, but only if adequate data—processed and cleaned using clinical data-engineering approaches—are made available. However, traditional data pipelines supporting EHR analytics have focused mainly on specific departments or problem domains. These approaches are neither scalable nor generalizable across clinical domains nor, ultimately, for population health analytics.

Although cornered by standard methodology, population health analytics possess their own peculiarities. ACER publications and research papers address the majority of aspects as separate units, while here the focus is wider: the methodology as a whole is presented through a single example, encapsulating the underlying problems and solutions as a seamless workflow. The presentation of each point enables exploration of the full breadth of data-engineering capabilities, thus supporting a wider audience even when their expertise in a specific area is limited. This broad-scoped analysis facilitates dissemination and reach while being a valuable reference for implementations across any of the considered key aspects.

2. Foundations of Clinical Data Engineering

Although the availability of large-scale Electronic Health Records (EHR) databases containing structured patient-level clinical, imaging, laboratory, and wearable device data holds great promise to reduce clinicians' workload and support precision healthcare delivery, the deployment of AI-based clinical analytics applications is both challenging and slow. The key components needed to improve the situation are described under the rubric Clinical Data Engineering, which combines data quality, data governance, the provision of clinically relevant features, and a well-defined data architecture. As a specific application may require only a small subset of the data, improving the data engineering quality, speed, and cost for a specific application can be done without compromising data quality for the remaining applications.

Engineering aims to provide a correctly functioning product with limited cycle times, low cost, and the ability to check the correctness of the produced data. The construction of clinical data lakes or warehouses offering a large-scale, well-handled, single-copy database featuring standardization, integration, temporal, provenance, and specific domain knowledge information performs a support role for clinical-research-oriented application developers and AI researchers. Engineers in this domain focus on providing high-quality data for the high-frequency queries

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

required for AI and machine-learning analysis. The engineering process will build a private, semipublic, or public data lake or data warehouse with appropriate controls to provide compliance with governance requirements.

2.1. Data Architecture in Electronic Health Records

Data architecture is an essential element of every Electronic Health Record (EHR) system, describing data models, data schemas, data lineage, and the temporal aspects of EHR data. It represents the bedrock upon which all applications, reports, and analytics on clinical data are built. Properly implemented data architecture supports scalable analytics and a wide range of clinical analysis of electronic health data. Most importantly, data architecture enables clinical data engineering for EHR. The value of an EHR is directly correlated to the extent of data model and schema analysis. Data models define what business processes are supported by the schema and what questions can be answered.

Concerns such as security and access control are not included in data architecture; these belong to the domain of governance. Governance is concerned with issuing access rights and ensuring privacy and security compliance. Data architecture design requires balancing competing requirements. Remember that clinicians examine their patients one at a time. The product that matters most is providing clinicians with tools that make them better at their real work of caring for patients. At the same time, there is growing recognition that the greatest potential value of EHR may be realized when analytics are delivered at population scale.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

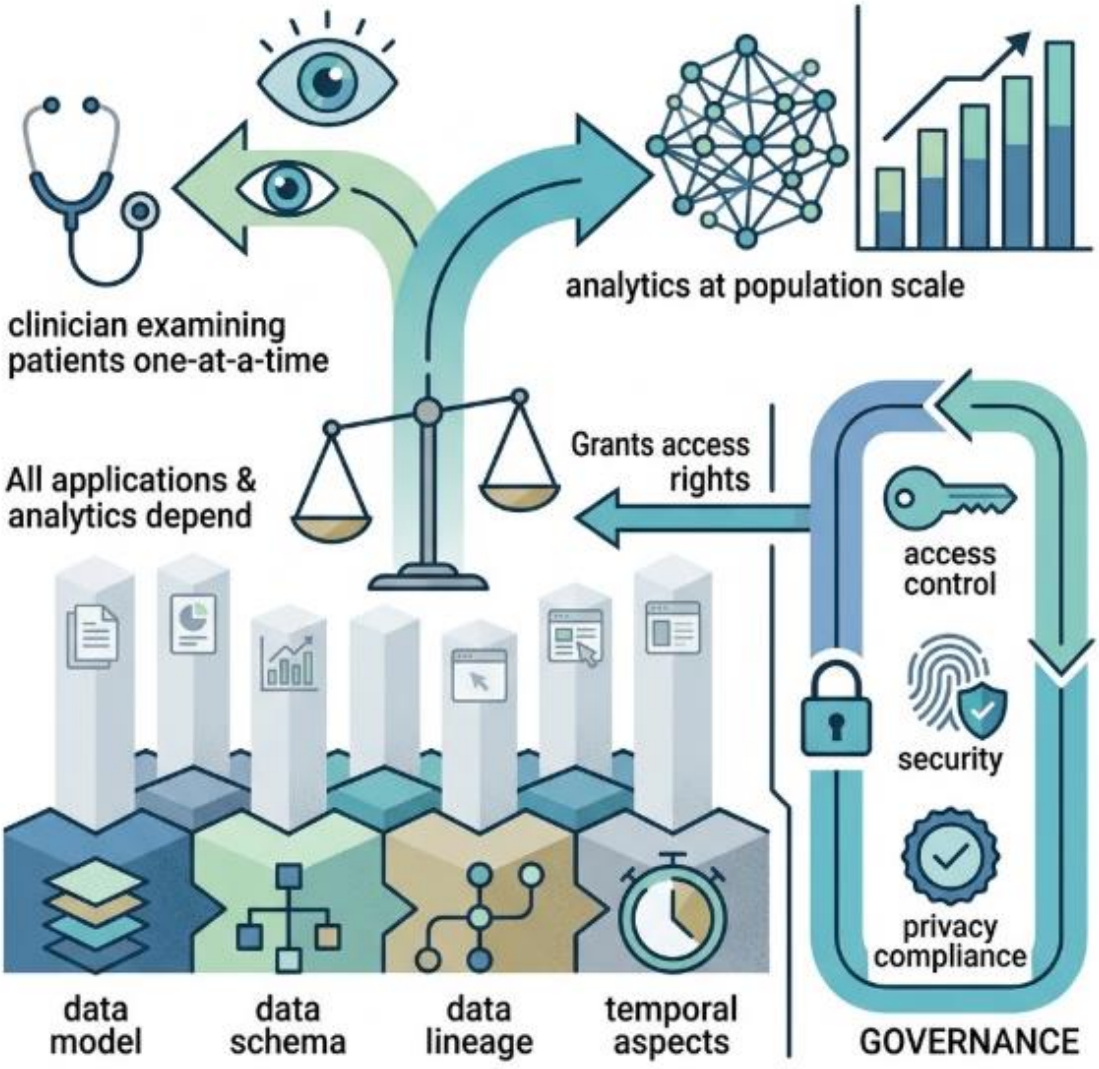


Fig 1: Architectural Foundations of Electronic Health Records: Balancing Clinical Utility with Population-Scale Data Intelligence

2.2. Standards and Interoperability

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

The need for standards in the health domain is driven by several factors, such as the need to provide meaningful data in systems designed and implemented by different software developers for a very heterogeneous customer base. Different health domains generate different data in a wide variety of formats, which further complicates analysis. To address these concerns, various organizations have introduced standards used widely in the industry. Some key relevant information technology (IT)-related standards include the following:

- HL7 FHIR: HL7 (Health Level 7) is a non-profit organization involved in the development of standards for electronic health, clinical and administrative data. FHIR (Fast Healthcare Interoperability Resources) is a standard for exchanging health data among various software applications developed by different vendors. FHIR combines the best features of previous HL7 standards, supports Web-based services and application programming interfaces (APIs), and facilitates validation using eXtensible Markup Language (XML) and JavaScript Object Notation (JSON).

- DICOM: Digital Imaging and Communications in Medicine (DICOM) is a standard that accepts all formats of medical images generated by computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, radiology, and other modalities. It owns the file structure for these images, the data transfer and storage protocols, and service definitions for data access, storage, and management over a network.

- SNOMED CT: The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a systematically organized collection of medical terms that provides clinical concepts covering diseases, findings, procedures, microorganisms, drugs, and many other domains of medicine. These concepts are organized in a hierarchy according to their semantics and are connected with each other using semantic relations.

- ICD-10: The International Classification of Diseases 10th Revision (ICD-10) is an international standard for coding diseases and various health conditions developed by the World Health Organization (WHO). It contains codes for more than 14,000 diseases and health-related problems and attempts to cover all diseases and health-related problems.

- LOINC: The Logical Observation Identifiers Names and Codes (LOINC) is a standard for identifying health measurements, observations, and documents. It facilitates the transfer of clinical data without loss of information content. LOINC provides a universal standard for

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

identifying laboratory and clinical observations and promotes the harmonization of environmental data.

As standards from different domains may use different code systems, mapping between them is often required for data integration purposes. Semantic harmonization can also be applied to normalize the same health concept represented by different codes in different code systems.

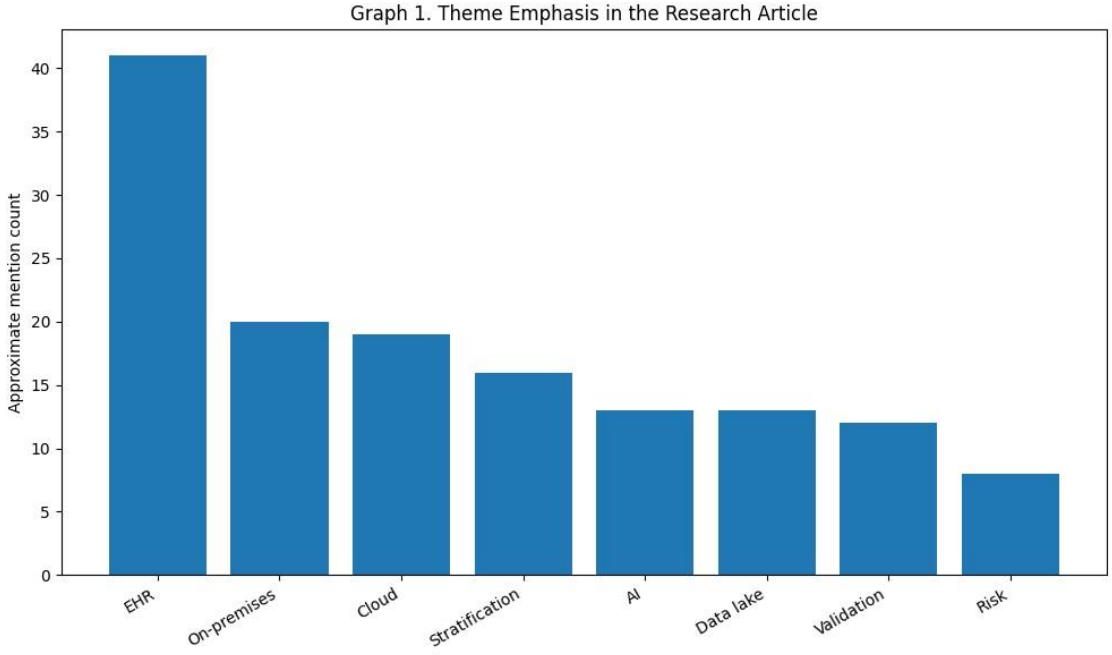
3. Methodology

The research employs an engineering approach to clinical data, proposed in concept in a previous work. This section outlines the methodology in chronological order, with validation strategies following the description of each major technique. The leading step is a comprehensive design of all data pipelines and components, enabling development and testing in isolation before integration into a complete system. Prioritization of reproducibility and clear presentation of code structure allow for peer validation via code review visual configuration, independent replication of any part, and orchestration of all components with minimal instruction.

HG, HA, and DB designed the detailed pipeline for ingesting, integrating, and processing raw clinical data. Source data typically include electronic health record, medical imaging, laboratory test, and wearable data. Electronic health records represented the principal source of time-series clinical data; ingestion used the proprietary Shimmer platform, which utilizes event-based architectures for ETL/ELT pipelines. Hospital policy restricted use of cloud services for clinical data; thus, on-premises processing leveraged dedicated databases. Clean clinical data are the basis of all features relevant for supervised tasks of population EHR data, and quality-assurance checks for key features have also been defined and automated. Each check demonstrates that the feature satisfies the desired data-quality criterion and warns if it exceeds a defined threshold for acceptable quality. All checks are implemented as unit tests within the private ETL workflow; thus, they run, validate feature quality, and ensure reproducibility of clean data every time the pipeline is executed.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Graph 1. Theme Emphasis in the Research Article



Equation A. Data ingestion and integration model

A1. Raw source representation

Suppose a patient i has data from multiple sources:

- EHR events
- laboratory records
- imaging metadata/features
- wearable streams
- genomic markers

Write the raw data for patient i as

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

$$\mathcal{R}_i = \left\{R_i^{(ehr)}, R_i^{(lab)}, R_i^{(img)}, R_i^{(wear)}, R_i^{(gen)}\right\}$$

Each source is time-dependent, so for source s,

$$R_i^{(s)} = \{\left(t_{ij}^{(s)}, x_{ij}^{(s)}\right)\}_{j=1}^{n_i^{(s)}}$$

where:

- $t_{ij}^{(s)}$ = timestamp of record j,
- $x_{ij}^{(s)}$ = measured value or event payload,
- $n_i^{(s)}$ = number of records from source s for patient i.

A2. Unified data integration function

$$\mathcal{I}(\mathcal{R}_i) = Z_i$$

where Z_i is the unified patient-level integrated record.

Expand this as:

$$Z_i = \phi\left(T^{(ehr)}\left(R_i^{(ehr)}\right), T^{(lab)}\left(R_i^{(lab)}\right), T^{(img)}\left(R_i^{(img)}\right), T^{(wear)}\left(R_i^{(wear)}\right), T^{(gen)}\left(R_i^{(gen)}\right)\right)$$

Here:

- $T^{(s)}(\cdot)$ = source-specific cleaning / transformation,
- $\phi(\cdot)$ = fusion operator aligning all sources by patient identity, time, and schema.

3.1. Data Acquisition and Integration Strategies

Data acquisition and integration require identification of data sources, methods for data ingestion, transformation plans, and integration with data quality management. The principal data source for enabling clinical data engineering is the electronic health record, but auxiliary

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

data from medical imaging, genetic screening, blood and other body fluid laboratory tests, wearables, and other sources can enrich the dataset, enabling more precise analysis and predictions, and fundamental reason for the data lake architecture. For the electronic health record, techniques can be considered for streaming, near-real-time, and batch-based ingestion. Streaming enables immediate accessibility of the data, but data freshness and accuracy trade-offs should be considered, especially for datasets expected to change only every few hours or days. Near-real-time approaches rely on a message queue for temporary storage while end-to-end architecture latency is controlled. Batch-based ingestion permits extensive pre-processing before data integration and should be preferred whenever latency is not business-critical. Etiology-sensitive fault-tolerant transfer technology for sensitive information on patients' privacy is crucial for the proper development of any solution. Laboratories often provide an ELT-based solution along with a periodic transfer of data that undergo transformation and cleansing operations at the cloud provider before entering the destination environment.

Confidentiality and integrity require further data processing to eliminate elements of hospitalisation not allowed. For medical images, the DICOM set format naturally supports a well-structured semantic description. However, even in this case, it is important to define procedures for periodically loading the entire medical images database as a data vault structure. Apart from EHR data, genetic tests run on a per-patient basis identify pairs of values that can be translated by a decision tree approach into features of clinical significance for the disease under examination. These other data sources, although fundamental for supporting precision healthcare, are usually not built-in for near-real-time ingestion but can be handled with batch mode and ELT architecture. When wearables data are also in consideration, stream processing solutions like Apache Kafka should be exploited in conjunction with efficient data pipeline architectures able to group these signals on a time-level feature basis. As new data pipelines start covering all elements of the data, data quality can be checked and independently assessed.

Dimension	Cloud-native	On-premises
Elastic scalability	5	2
Setup agility	5	2
Direct local control	2	5
Regulated-data fit	3	5
Governance burden	3	5

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Dimension	Cloud-native	On-premises
Cross-org sharing ease	5	2

4. Objective of the Study

The objective is to develop techniques for scalable analytics in clinical data repositories—a dataset for clinical risk stratification in heart disease patients is used to demonstrate the efficacy of stratification and risk score prediction—treatment recommendations for optimizing antidiabetic treatment in patients with type 2 diabetes mellitus, leveraging counterfactual analysis—methods are applicable to any other problem domain.

A lack of access to healthcare delivery data, especially Electronic Health Record (EHR) data, hinders the use of AI in optimizing healthcare delivery. The use of cloud solutions can help in scaling analytics but raises security concerns, especially for sensitive patient data such as EHRs. Data in the cloud, protected by security controls, can be used for analytics without any security and latency breaches. Cloud data solutions can scale automatically for both structured and unstructured data, thus saving time in capacity planning and avoiding bottlenecks. Compliance with Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) is possible by specifying the appropriate location for storing sensitive data when using cloud solutions. Compared to cloud solutions, on-premises ones can offer greater control if the data governance processes are implemented and followed meticulously. Data lakes can ingest any data type without any prior ETL process. Advanced analytics, including deep learning, can directly access data in data lakes. On-premises data solutions offer greater control over security, but the onus is on the organization to secure the data. If data governance and risk management strategies are followed, data governance is as good as that offered by cloud providers.

4.1. Research Objectives and Goals

Advances in computing infrastructure, artificial intelligence, and clinical understanding have enabled the production of successful domain-specific AI models. Despite these advances, the scientific and engineering expertise required to generate workloads for particular AI-models are seldom shared or made broadly available. This work presents clinical data engineering as a new discipline that combines clinical data engineering—application of data-engineering techniques from the field of data engineering to specific challenges associated with EHR-centric

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

data—and developed a set of standardised, reusable components that can contribute to making Clinical AI model workflow Scientific Software Engineering across diverse healthcare provider organisations.

Research going beyond the building of single AI-models and focusing on the development of a complete analytical capability require structured, reliable data pipelines within the underlying electronichealth record–centric data repositories. A set of techniques and components are presented for implementing Ingestion and Integration Pipelines (IIPs), which combine Internet of Things sensor data with traditional Internet of Things, Imaging, Genomics, Laboratory, Drug, Social medicine, Clinical and Research data sources in either a Continuous or Batch mode of operation with appropriate dependency orchestration on these data sources and Data Management Metadata approaches to capture the complete data lineage and provenance of All Records.

5. Research Summary

The research summarizes a work in clinical data engineering that develops techniques for scalable Electronic Health Record (EHR) analytics and enables a new era of precision healthcare delivery. The work advances the collection, integration, and analysis of clinical data. With such data grounded on engineering principles, it is now possible to provide comprehensive insights—from EHR- and non-EHR-based risk prediction and patient stratification to data-driven treatment recommendations.

The culmination of an AI-driven engineering effort, the research places an emphasis on healthcare information systems. Comparison between cloud-native and on-premises architectures showcases a trade-off between scalability and security. AI-underpinned methods encompass patient grouping, risk scoring, and tailored decision support. Finally, an on-premises EHR analytics platform supplements two proof-of-concept applications. Transitions to real-time processing and concurrent multi-cohort stratification are presented as additional advances for acute care and population health domains, respectively.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Equation B. Data quality equation

B1. Feature-level quality score

Suppose feature f is evaluated by K quality checks:

- completeness,
- plausibility,
- consistency,
- timeliness,
- range validity.

Let the k-th check for feature f be q_{fk}, normalized in [0,1].

Then the feature quality score is

$$Q_f = \sum_{k=1}^{K} w_k \, q_{fk}$$

with

$$\sum_{k=1}^{K} w_k = 1, \qquad w_k \geq 0$$

Step-by-step

1. Each check returns a normalized score:

   $0 \leq q_{fk} \leq 1$

2. Assign importance weights w_k.
3. Weighted sum gives total quality:

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

$$Q_f = w_1 q_{f1} + w_2 q_{f2} + \cdots + w_K q_{fK}$$

B2. Patient-level data quality score

If patient i has p engineered features,

$$Q_i = \frac{1}{p} \sum_{f=1}^{p} Q_{if}$$

where Q_{if} is the quality score for feature f of patient i.

5.1. Comparative Analysis of Cloud and On-Premises Data Solutions

Architectural comparisons of cloud-native and on-premises data management solutions encompass deployment models, scalability, security posture, cost implications, regulatory risk, and considerations for data governance, sharing, and control. Cloud-native architectures inherently lend themselves to multi-tenancy; ease of integration with other applications; and efficient, cost-effective elasticity of scaling both compute and storage resources on demand. By contrast, exploitation of such elasticity may be more difficult with on-premises architectures—requiring capital investment in physical infrastructure that may sit idle for extended periods.

Yet, cloud-native solutions must be carefully validated to ensure that the expanded attack surface associated with geographically distributed data storage and multi-tenant on-demand resources does not introduce unacceptable risks of data breaches, compromise of data fidelity, or exposure of proprietary information. For sensitive data that must remain on-premises (due to jurisdictional, regulatory, or security controls), an on-premises solution will introduce relatively little overhead and latency when compared with an equivalent cloud-native solution.

Although both cloud and on-premises deployments offer tremendous ecosystem advantages, consideration should also be given to the level of control that organizations wish to maintain over data governance, data stewardship, and sharing of data across organizations. Cloud-native architectures make it easy to connect and securely share data with other partners, but such ease comes at a cost. When information that is considered private, proprietary, or subject to legal restrictions is being stored and processed, organizations must ensure that it remains secure and confidential. Organizations should assess the level of risk they are willing to

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

accept in complying with specific regulations or with the requirements of particular partners (customers, suppliers, service providers, and peers) and should then choose the data management solution that best addresses their organization's privacy and security needs.

6. Techniques for Scalable EHR Analytics

At the core of clinical data engineering are the methods that enable scalable analytics across different timeliness and geography in Electronic Health Records (EHR) and other data sources. The availability of clinical temporally stratified data—enabling analytics at various points, ranging from real-time decision support to cohort analysis, with integration of additional, often lagging indicators, such as lab values and imaging features—is crucial. This section discusses the design of data ingestion and integration pipelines process, as well as the need for identifying clinically meaningful features for analysis.

The most significant data engineering efforts for implementing scalable analytics typically concern the design and construction of data ingestion and integration pipelines. Two types of pipelines are often required: streaming and batch. Metadata management tools play an important role in the orchestration of the different steps of the data engineering process. Moreover, in the context of clinical data engineering, the main engineering effort in feature engineering concerns not the identification of machine learning features—that is done automatically—but the determination of clinically meaningful features that stratify the patients for specific analysis. Domain knowledge is essential for defining appropriate temporal features and for determining how missing values should be handled for the specific analysis.

6.1. Data Ingestion and Integration Pipelines

Data ingestion and integration pipelines enable timely and accurate response to diverse clinical analytics. Streaming pipelines feed EHR information and preprocessed data into a data warehouse optimized for real-time use cases, while batch pipelines geared towards population-scale applications integrate historical data. Orchestration mechanisms for dependency chaining ensure execution according to user-modelled workflows, efficiently managing diverse data sources (e.g., EHR, medical imaging, laboratory, and wearable data). Metadata collectively delineate the provenance of integrated data, providing critical context for subsequent feature

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

engineering, cohort definition, and analytical investigations.

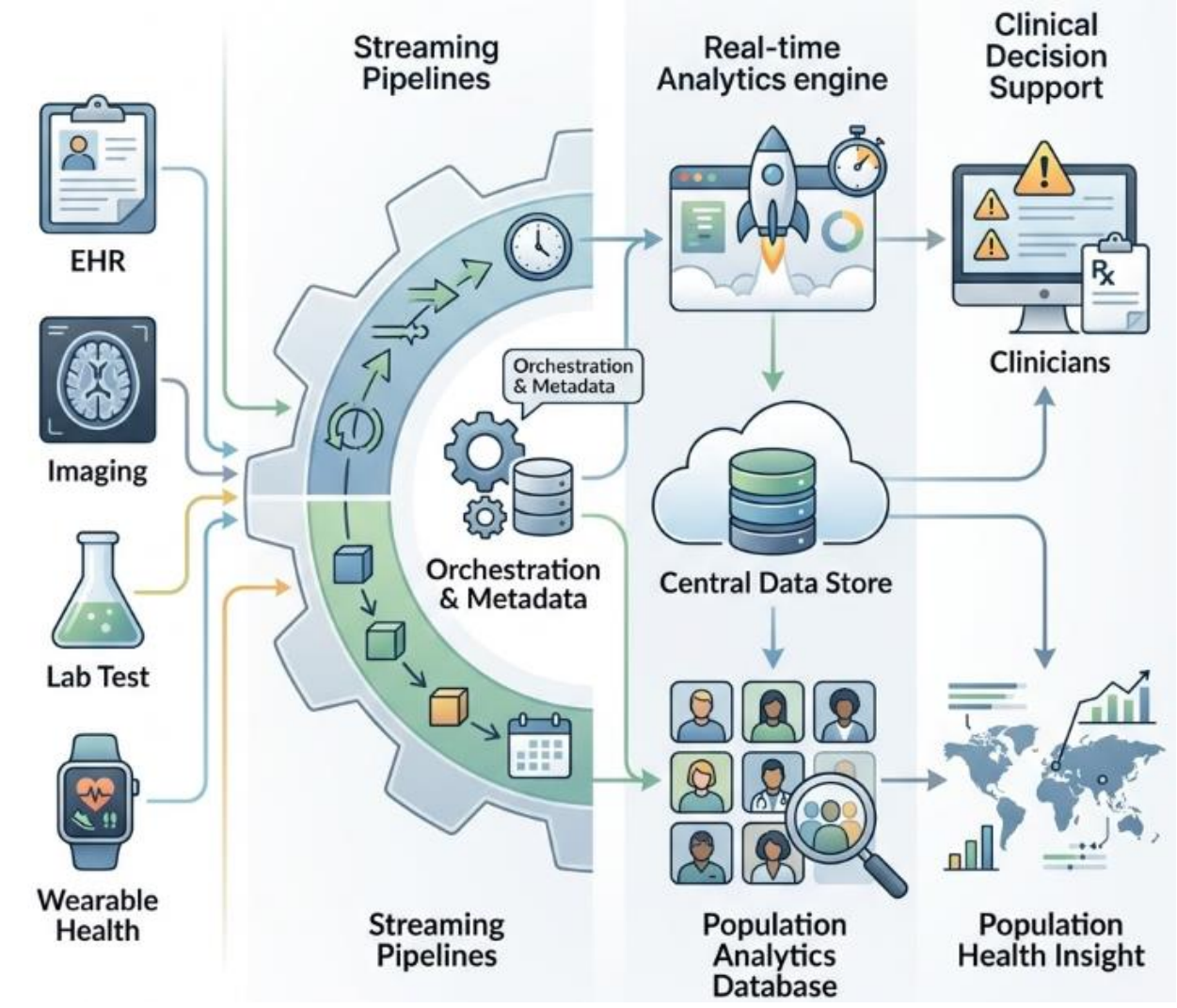


Fig 2: Heterogeneous Data Integration Architecture for Dual-Track Clinical Analytics: Enabling Real-Time Decision Support and Population-Scale Insight

Clinical analytics consume extensive, heterogeneous information originating from disparate domains. Patient management in hospitals necessitates rapid turnaround of new

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

information (e.g. laboratory, imaging reports), often enabling targeted decision support in near real-time. Meanwhile, scaling models to larger patient cohorts with medical history extending back several years often provides deeper insight into model behaviour and associated uncertainties, thereby aiding in downstream, patient-specific recommendation generation.

6.2. Feature Engineering for Clinical Insight

Large-scale EHR datasets contain extreme variety and richness, such that the right analytical features can provide nearly any meaningful clinical insight that can be derived from lost or rarely available external clinical data. Capturing these features as engineerable combinations of primary clinical data can facilitate the delivery of commonly requested analytical capabilities, support clinical and regulatory evidence generation, and allow for rapid validation and deployment of AI-based clinical services. Essential engineering of specific clinical features provides additional patient stratification and risk stratification beyond that directly available in the core clinical data.

Landmark analyses in focus areas such as frailty, myocardial infarction, renal failure, or COVID-19 can, moreover, be adapted to generate patient cohorts with commonality in the feature associated with interest—for example, risk of myocardial infarction—as well reflective of the other prognostic features available in the dataset. The resulting feature sets can be engineered for counterfactual clinical treatment response, optimized for clinical formulation of the outcome of interest, and aligned with available data and planning pathways in view of supporting clinical adoption of practical applications.

	Theme	Mentions
0	EHR	41
3	On-premises	20
2	Cloud	19
6	Stratification	16
1	AI	13
4	Data lake	13
7	Validation	12

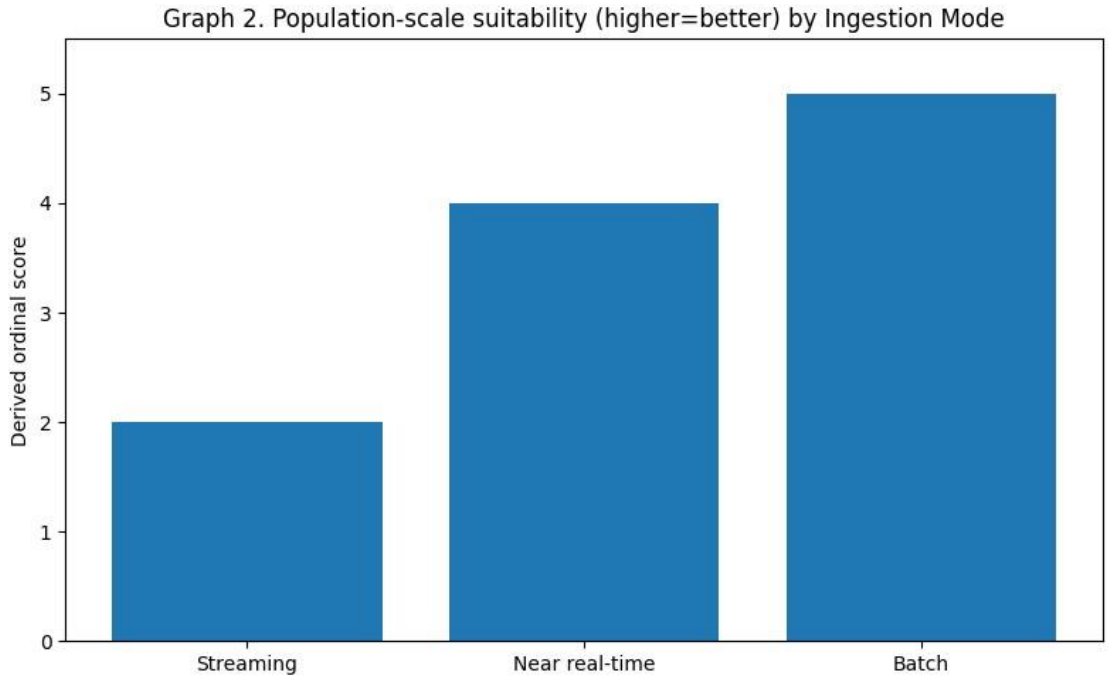
7. Precision Healthcare Delivery through AI

Machine learning and AI have demonstrated impressive results in many aspects of healthcare delivery. However, most of the models are not directly useful for clinicians. One promising avenue of applying AI is that of patient stratification, where a population is split into cohorts, each with distinct characteristics and outcomes. Such cohort generation enables tailored treatment for each subgroup, rather than a one-size-fits-all method.

Patient stratification helps guide treatment. For instance, if a cohort with a significant proportion of patients developing a specific complication is identified, clinicians can focus efforts on that subgroup. By doing so, they may discover suitable countermeasures or preventive solutions, and thus reduce the risk of that complication. AI can also help determine risk scores for patients within a cohort. These risk scores may present the opportunity of risk-adapted recommendations, such as monitoring visits, imaging check-ups, or therapeutic changes, in a more sensitive manner that minimizes the chances of a condition worsening undetected.

In the context of elective surgery for cancer, these risk scores can assist clinicians in deciding how best to optimize the treatment for each individual patient. One option is to recommend an adjustment of the operation strategy. Another is to additionally consider ancillary treatments (such as radiotherapy, chemotherapy, or immuno-oncology therapy) based on the inferred risk of unfavourable outcomes.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19



7.1. Patient Stratification and Risk Prediction

Patient stratification establishes homogenous subgroups sharing common clinical characteristics, progression trajectories, treatment responses, or outcomes. Stratification is an essential step in a preventive healthcare regime and typically performed on historical data. Here, clustering is first applied to a temporal data matrix to create several distinct patient cohorts, with adequately large sample sizes and clinically meaningful differences. A temporal risk score is then developed—a superset of other risk scores for patients with acute conditions. Validating each of the resulting cohorts ensures appropriate cohort creation.

Clusters are created by applying a clustering algorithm to the temporal data matrix of patients, followed by evaluations. Validation ensures that the resulting cohorts support further downstream tasks such as risk score derivation and treatment effect prediction. The clustering task first groups patients presenting with chronic obstructive pulmonary disease (COPD), then makes use of data coming from patients with some other long-term diseases. After that, it focuses on patients suffering from any other acute condition and lastly considers data from

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

patients having diabetes. Finally, risk scores quantify the relative immediate risk of patients developing an acute illness. Time is expressed in hours, and the risk score evaluates how likely a patient is to have an acute event within the next 48 hours. Risk score derivation makes much smaller time intervals than clustering.

Results are then used to develop an AI-based recommendation system that takes a patient, their background, and their counterfactual treatment and predicts whether this treatment will improve the patient's situation in an objective manner. These recommendations, coupled with semantic interoperability, aim to create proactive decision-making support for medical staff. Such an approach has applications in acute conditions like sepsis or in the prediction of deterioration of critical patients in an ICU environment.

Equation C. Feature engineering equation

C1. Window-based temporal feature

Suppose $x_i(t)$ is a variable such as heart rate or glucose.

For a retrospective window of length Δ, define mean feature:

$$f_i^{(\text{mean})}(t) = \frac{1}{\Delta}\int_{t-\Delta}^{t} x_i\,(u)\,du$$

For discrete observations:

$$f_i^{(\text{mean})}(t) = \frac{1}{|\mathcal{W}_{i,t}|}\sum_{u\in\mathcal{W}_{i,t}} x_i\,(u)$$

where

$$\mathcal{W}_{i,t} = \{u: t-\Delta \le u \le t\}$$

C2. Trend feature

To quantify change over time,

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

$$f_i^{(\text{trend})}(t) = \frac{x_i(t) - x_i(t-\Delta)}{\Delta}$$

7.2. Personalized Treatment Optimization

Timely and effective treatment of patients requires not only an understanding of their clinical needs but also an understanding of various external factors that substantially impact the success of different treatments, such as age, race, and gender. In many circumstances, treatment optimization is challenging because clinical guidelines typically prescribe only one treatment option for a patient, though there is uncertainty about the treatment effect for that person. Counterfactual analysis leverages the insights from patient stratification to provide an informed recommendation about the best treatment for an individual patient in well-defined patient cohorts where the different treatments are being compared. In particular, the combination of stratification, risk score simulation, and counterfactual treatment recommendation, with their supporting software, can be used as a treatment decision support machine learning pipeline by healthcare professionals.

Specifically, given a training cohort of patients receiving one of k different treatments, it prepares three components that can be deployed together or independently. First, it stratifies the patients into k relatively homogeneous groups for which the treatments are more directly comparable, and it provides a k-part grouping for the complete training cohort. Second, the probability of readmission within the next 30 days is estimated for each patient using a machine learning prediction method for which probability estimates provide a risk score. Third, treatment assignment is simulated for new patients, allowing identification of counterfactuals for a user-specified treatment. Counterfactual treatment recommendations can then be provided for a specific patient. These components can thus provide personalized treatment recommendations.

8. Infrastructure and Deployment Architectures

Selecting the right infrastructure architecture is a key consideration for developing a scalable analytics and AI-driven EHR analytics platform. These solutions are usually deployed either in the cloud or on an organization's premises, with hybrid deployment options becoming increasingly common. In cloud-native solutions, all components are hosted in the cloud. These deployments offer businesses much greater scalability, as resources can be provisioned, scaled up or down, and released almost instantaneously, usually with only a few API calls. Moreover, due to economies of scale, the solutions can also be cheaper when compared with on-premises

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

approaches. However, moving sensitive data to the cloud is also associated with data being at greater risk of security breaches due to vulnerabilities present in the systems. Additionally, depending on the architecture of the cloud deployment, data may be subject to privacy laws and regulations that require appropriate geographical protection.

Conversely, in on-premises cloud solutions, resources are located outside the data center. These can offer organizations greater control over the security of such data and help ease compliance with various regulations. However, this comes at the cost of scalability, as servers cannot be easily upgraded or repaired and often take several months to be set up. Moreover, without a cloud provider guaranteeing a level of security, implementing a complete and robust security architecture is often very resource-intensive and difficult. In addition to security and compliance impact, the geographical location of the servers may also significantly impact the latency experienced by the users. Load balancing among multiple on-premises locations can help reduce latency and provide a better user experience.

In addition to cloud depth considerations, several other variables also determine whether to use a data lake or data warehouse. Data lakes are frequently adopted due to their flexible and low-cost nature, allowing organizations to store all available data sets with minimal intermediate schema definitions. However, the absence of a structured and expected schema may result in longer query execution time and require far more computing resources compared with data warehouses. Organizations with the requirement for near real-time querying response may favor a data warehouse, while organizations interested in population health analytics solutions that leverage data from wearable sensors may favor data lakes. Hybrid approaches that integrate and incorporate the advantages of both solutions are also possible.

Equation D. Patient stratification equation

D1. Patient matrix

For N patients and p engineered features:

$$X = \begin{bmatrix} \mathbf{f}_1^\top \\ \mathbf{f}_2^\top \\ \vdots \\ \mathbf{f}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times p}$$

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Each row is one patient's feature vector.

D2. K-means stratification objective

Assume we want K patient cohorts. Let:

- C_k = cluster k,
- μ_k = centroid of cluster k.

The objective is to minimize within-cluster variation:

$$J = \sum_{k=1}^{K} \sum_{\mathbf{f}_i \in C_k} \| \mathbf{f}_i - \mu_k \|^2$$

Step-by-step derivation

4. For each patient i, compute distance to centroid k:

   $d_{ik} = \| \mathbf{f}_i - \mu_k \|^2$

5. Assign patient i to the cluster with smallest distance:

   $z_i = \operatorname{argmin}_k d_{ik}$

6. Update centroid of cluster k:

   $\mu_k = \frac{1}{|C_k|} \sum_{\mathbf{f}_i \in C_k} \mathbf{f}_i$

7. Repeat until assignments stabilize.

So the stratification pipeline is:

$X \rightarrow$ distance computation $\rightarrow$ cluster assignment $\rightarrow$ centroid update

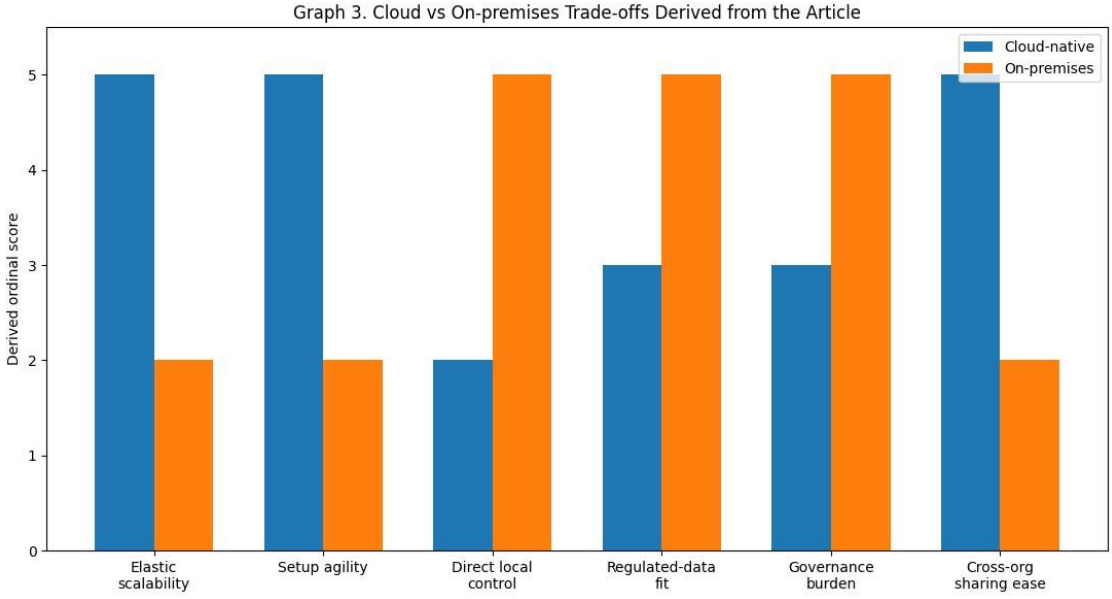
8.1. Cloud-Native versus On-Premises Solutions

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Architecture, scalability, security, cost, and regulatory limitations differ between cloud-native and on-premises solutions. Cloud service providers offer security controls and assurances not feasible for many healthcare institutions. However, data complying with governmental regulations forbidding cloud storage may be processed on-premises. Some projects involve multiple institutions, each using their own infrastructure. Later integration of the results may entail latency, privacy, and cost issues.

Node hosting, monitoring performance, security, and resource consumption can be handled with a click of a button in the cloud. Deployment of local servers supporting the same services, along with continuous evaluation, requires dedicated teams. Packaged servers can cost as much as local maintenance. Implementation is usually performed by cloud services, bringing important benefits for the developers. From a development perspective, the reduced cost of updating or adding new resources often offsets the monthly investments. Scalability and price depend on usage policies: Pay-as-you-go services are useful to cope with variable requirements. For fixed demand, opting for reserved resources may drastically reduce unit prices.

Graph 3. Cloud vs On-premises Trade-offs Derived from the Article



8.2. Data Lakes, Warehouses, and MDM

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Data Lakes, Warehouses, and Master Data Management: Distinctions between the roles of data lakes and warehouses, the objectives of master data management, and the responsibilities of data stewardship are explained.

Data lakes typically support exploratory analysis and serve as platform for integrated machine learning at scale. Their high-level structure mirrors the schematic level of a data warehouse, consisting of master data tables, business-contextualized data payload tables for ingestion speed, and a collection of pre-created features for machine learning. Patient and data dictionaries provide crucial metadata. By contrast, a warehouselike architecture built on cloud offerings such as Amazon Redshift, Google Big Query, or Snowflake can support ad hoc analytics by domain experts. A data lake serves as the primary back-end basis for wider consumption—with dedicated consumption-based services—asset management for optimal price performance and ingest of very large amounts of data in a scalable & separated manner that supports parallel contiguous reads for analytics workloads and near real-time data integration or machine learning applications. Separate business-contoured tables speed up throughput of analytical queries and other analytics workloads.

Data lakes containing dirty data, unlabeled images, and semi-structured or textual data beg for a master data management layer that produces single-source-of-truth datasets and business-contextualized–yet privacy-preserving–payload tables for distribution into warehouse copies within the transactional business domain. Where a data lake supports the mantra of "schema-on-read" for exploratory use cases and the data storytelling lifestyle main information access, the role of data stewardship applies additional domain expertise to ensure high-quality, compliant, privacy-preserving data product catalogues built via "schema-on-write" pipelines.

9. Ethical, Legal, and Social Implications

AI-Driven Clinical Data Engineering for Scalable EHR Analytics and Precision Healthcare Delivery: Use an objective, scholarly tone, prioritize evidence-based arguments and formal

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

structure.

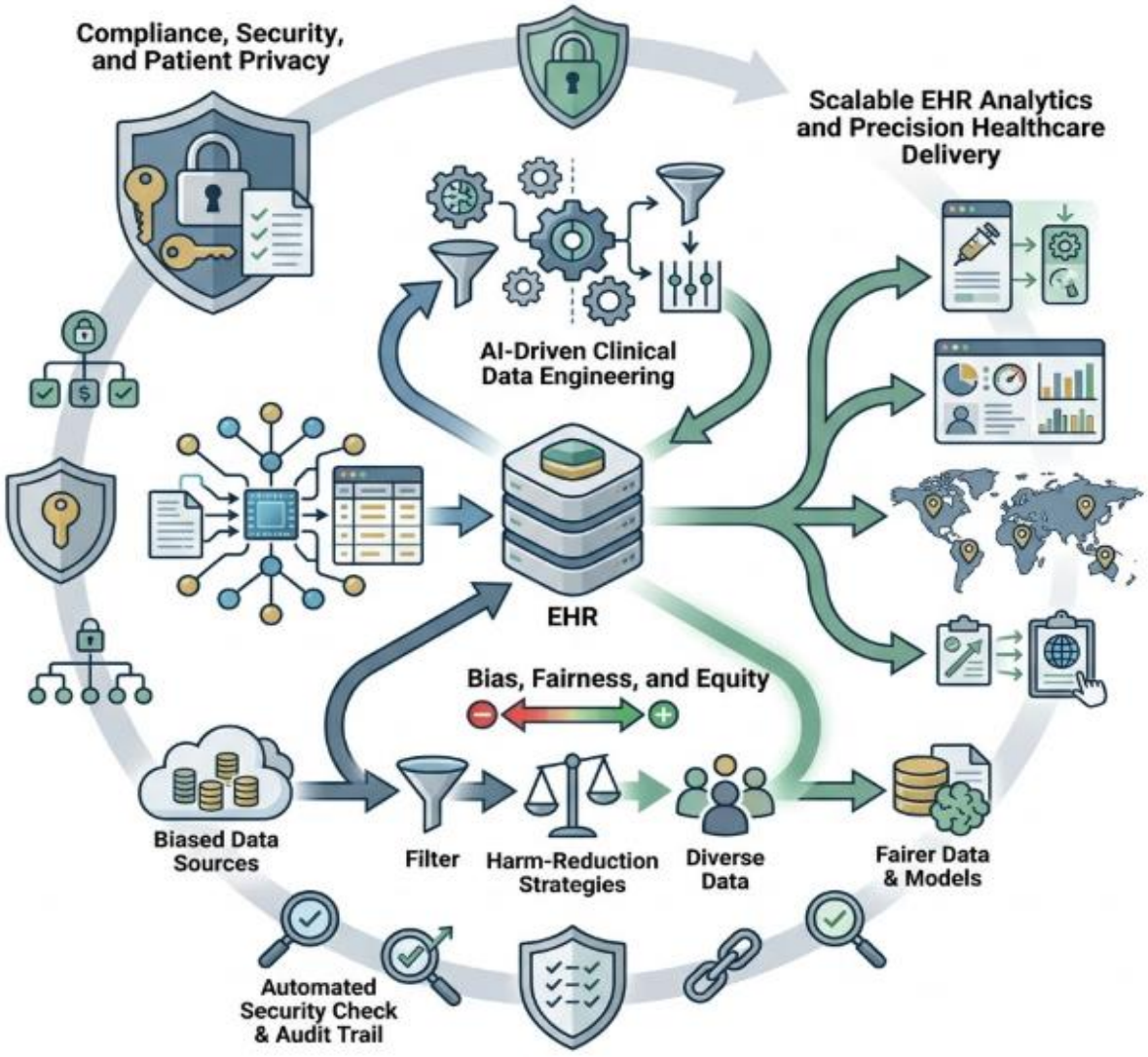


Fig 3: Governance and Equity in AI-Driven Clinical Data Engineering for Precision Healthcare Delivery: Mitigating Bias while Ensuring Security and Regulatory Compliance in EHR Analytics

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Bias, Fairness, and Equity in AI for Healthcare: Several sources of bias can affect AI systems, including the data collection process, feature selection, feature representation, and data labeling. Bias can results in fairness inequalities for certain groups, with potentially devastating effects for affected groups. Deploying harm-reduction strategies can enhance the equity of innovative healthcare solutions.

Compliance, Security, and Patient Privacy: AI has the potential to revolutionize healthcare, but properly governing the deployment of such technology is paramount for patient safety. Guidelines delineating how AI should be utilized seek to ensure that deployed models are safer, more reliable, prevent data breaches, and ensure patient privacy. Complying with such regulatory demands requires significant planning, expertise, and diligence. Failure to comply can result in severe penalties. One example in the US is the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and provides patients with rights over their health information.

9.1. Bias, Fairness, and Equity in AI for Healthcare

Addressing bias contributing to health disparities in AI—introduced through the data, design, and implementation—remains critical. Examining development data, integrating pre-processing bias detection methods, and applying Fairness-Aware Artificial Intelligence can reduce bias, while quantifying fairness ensures equitable solutions. Identifying marginalized groups informs a review of key design stages to ensure fair outcomes.

Research confirms referral and outreach biases for housing, transportation, education, and public service. Non-representative AI data introduces further bias for languages, skin tones, eyeglass definitions, and other attribute groups. Historically underserved populations are less likely to be included in bandwagon applications, enhancing perverse referral patterns. Sources of inequity also extend to enabling data preparation, algorithm implementations, or different target audiences. Broadening project goals with insights from Human-Centred AI Design can offset these shortcomings. Defining and quantifying fairness guarantees equitable targeting of successful solutions.

Solutions remain available for each bias source, requiring designers to commit. For development bias, data-analysis pre-screening tools support insight to document co-occurrence of attributes in global and target samples. Discrepancies inform solutions for better group representation or specific measurement for improved prediction. CounterFactual Bias Detection

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

exploits external data to promote model independence of sensitive attributes. Incorporating CounterFactual Fairness Detection evaluates deployed solutions for sensitive-attribute supported predictions.

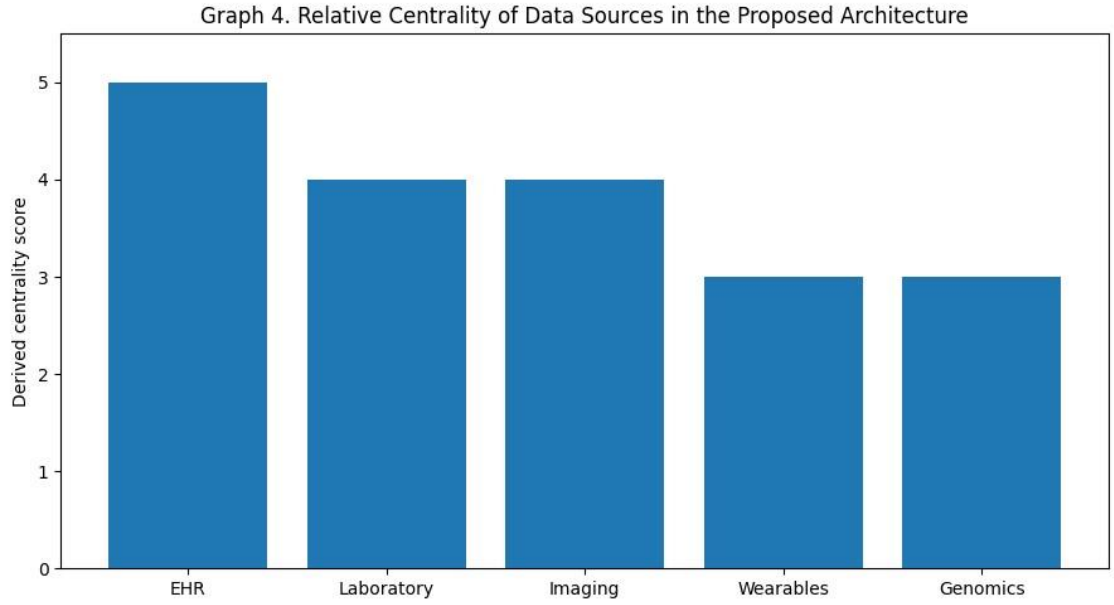
9.2. Compliance, Security, and Patient Privacy

The regulatory landscape governing healthcare and healthcare data use can be intricate. Numerous regulations exist within many jurisdictions to protect the confidentiality and security of individual health information, enable access to that information by regulators, researchers, and other actors, and foster the effective use of that information for the benefit of patients and of society in general. Detailed delineation of these regulations is usually found in the associated documentation and training associated with any given institution. A mapping of key compliance and security control areas, including procedures (beyond technical) associated with health data consolidation within a cloud-native data lake for research/AI purposes, is crucial within the context of the present study—though the technical implementation is not covered in detail. The key controls typically cover governance, privacy and patient consent, data security, and system access.

The governance strategy of an AI infrastructure involves designing appropriate levels of authority and decision-making structure for managing, monitoring, and accessing AI systems. Highly sensitive data under regulatory scrutiny presents challenges for data access and transfer, especially when accessed from within a cloud. However, rules can be devised permitting controlled access under strict monitoring (e.g., recording of individuals accessing the data and reasons for doing so) enabling cloud-based secure data science using on-premises-sensitive datasets.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

Graph 4. Relative Centrality of Data Sources in the Proposed Architecture



10. Evaluation and Validation of AI-Driven Analytics

Real-world clinical adoption of AI-driven analytic products is expected when they not only demonstrate capability but also support the three criteria of clinical validity, utility, and cost-effectiveness. Prior sections have addressed utility and cost considerations for scalable analytics; this section focuses on clinical validation. For AI products to be actionable, correct risk assessments need to be linked to genuine clinical plausibility, so that patients falling within high-risk strata can be targeted for screening or therapeutic intervention. Validation that is sufficiently rigorous boosts confidence in AI products and enables clearer regulatory delineation.

A comprehensive and systematic framework has been tailored for the clinical validation of AI-based predictive models. Such models support clinical decision-making by generating actionable outputs. Validation therefore explores the degree to which the prediction aligns with the reality of patient outcomes when the decision is associated with risk-reward balance; models providing merely accurate predictions are thus not sufficient. For models spanning data from multiple institutions and patient populations, clinical validation metrics further extend to evaluating performance differentials among strata.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

10.1. Clinical Validation Frameworks

A comprehensive set of metrics covers the full spectrum of clinical utility and directly relevant clinical significance. A metrics taxonomy includes standardized, commonplace clinical metrics; AI/ML analytic metrics; clinical safety, performance, and impact metrics; clinical acceptance metrics; clinical integration metrics; and clinical interpretability metrics, across two or more prospective user perspectives and relevant regulatory components. For each metric, a commonly adopted abbreviated definition addresses at least technical origin, and at least one of prediction, prevention, detection, diagnosis, prognosis, characterization, tracking change, therapy, response prediction, or treatment effect. Key insights include validation metrics for general predictive, prognostic, risk, probabilistic, soft-classifier, and ordinal approaches. To achieve true clinical utility, analyses must prioritize common-place, commonly monitored clinical metrics.

Given the nature of healthcare delivery, technologies that perform their own internal QA/QC are ideal, and these metrics define exemplary clinical-validation processes. A population of interest propels a clinical question that germinates a clinical hypothesis, driving a clinical neural network/ML analysis, ultimately resulting in a clinical innovation proposal supporting better healthcare delivery for that population. The final clinical validation step concerns clinical acceptance, eventual deployment into the healthcare delivery process, and ongoing surveillance for undesired effects or diminishing clinical performance over time requiring retraining.

AI/ML technologic advances offer unprecedented opportunities to transform healthcare delivery for the better. Fulfilling those expectations requires a rigorous clinical validation framework encompassing the necessary technical and clinical stages, all with the necessary practical relevance. Relaying AI/ML technology clinical-performance-validation requirements to the technologic community will facilitate development of innovations that possess true clinical utility.

10.2. Reproducibility andExternal Validation

Reproducing research is critical for building trust, understanding cause-and-effect relationships, and increasing the chance that good ideas will be adopted. Guidelines to increase reproducibility include defining an expected outcome for an experiment, performing a systematic search for misused methods, and global statistical analyses. Many research teams choose a release model in which their source code (and binary files, when applicable) are publicly

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

available in a version control repository or archive and are cited by the community for further study. These practices lower the general implementation cost and foster the development of complementary or competing methods while often avoiding workload duplication. However, the guidelines for reproducible research introduce an additional burden that is often ignored, especially in the early stages of method development.

A well-maintained repository contains a single package or self-contained method that can be built and installed from source using established package managers, requires only minimal setup and configuration, exposes an intuitive application programming interface to its users, and is preferably callable from multiple scripting languages. Sources of externally available data are clearly indicated, ideally providing direct links. Example input files are included, and a developer documentation set is provided. Wherever applicable, unit tests and other forms of source code coverage are included. Efficiency is tested, and guidelines for cited usage are provided in a usage vignette or in dedicated documentation. Specifically, the forward-looking development aims not only at the coherence, efficiency, and usability of analytic methods but also, where appropriate, at the establishment of a scientific environment dedicated to the evaluation and clinical translation of the results.

Equation E. Risk prediction equation

E1. Binary event label

Let

$$y_i = \begin{cases} 1, & \text{if patient } i \text{ experiences the target event} \\ 0, & \text{otherwise} \end{cases}$$

Examples from the article:

- acute deterioration in the next 48 hours,
- readmission within 30 days.

E2. Linear score

Given feature vector $\mathbf{f}_i$,

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

$$z_i = \beta_0 + \beta_1 f_{i1} + \beta_2 f_{i2} + \cdots + \beta_p f_{ip}$$

In vector form:

$$z_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{f}_i$$

11. Case Studies and Applications

The population health use cases validate an AI-powered clinical data engineering framework from the data engineering perspective. The clinical engineering framework serves as a means to accelerate and validate AI processes within the healthcare domain using population health data from the Emory Healthcare data lake. The data lake houses de-identified data from 2019 to 2021, spanning 164,020 patients and more than 2.4 million encounters. The scalability of ingestion and processing pipelines beyond the population size and the runtime of the various EHR analyses affords additional validation. With a near-realtime analytics solution built atop the data lake infrastructure, longer-term health outcomes such as 30-day readmissions are at risk of being nonstationary, influencing ML model performances and subsequently ML inference processes.

Focusing on the acute care use cases demonstrates an AI-driven clinical data engineering framework from the analytics delivery perspective. The cloud-native architecture holistically integrates EHR data across multiple hospitals, supporting multiple analysis pipelines delivered to clinicians in real time through an interface that also leverages key predictive and prescriptive models. Initial deployments of the system drive concurrent and accurate analytics for rapid cycle 32 and other clinical questions. With operationalization of additional models and pipelines, the architecture will support de novo research for participating hospitals—using the same underlying AI and the same technical foundation—as well as partners beyond Emory Healthcare, including the LIFT (Lung Injury Future Trial) consortium funded by the National Institutes of Health.

11.1. Population Health Analytics

Population health analytics present an avenue to structure the acquired data for cohort analyses, exploring the distribution of outcomes or attention in underrepresented patient subgroups. Additional studies based on the public MIMIC-III dataset illustrate the importance of addressing unusual characteristics of treatment administration as shown when attempting to predict intensive care unit stay or ventilation. The underlying process is highly discrete with a

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

lack of sampled uncertainties, and the imbalance among the supported categories requires caution. Cloud deployment using AWS demonstrates economic benefits when compared to on-premises infrastructures with a similar performance and security configuration for scalable streaming ingestion and batch feature engineering.

Analysing data for explanations or association with different outcomes is a natural application to assess the pipeline for processing population health data at different scales, attempting to reconstruct resignation risk from anonymized health records. Important limitations arise from the restricted attributes with the requiring additional resources in the clinical support of mortality warning systems. Disconnection during data collection allowed for exploring the data processing of crowdsourced assemblage during the Olympic games, retaining temporal resolution and investigating the availability of key resources at the hospitals featured in the accessible information.

11.2. Acute Care and Emergency Medicine

Advanced Clinical Analytics

Acute care and emergency medicine stand to benefit from the real-time capabilities of many of the methods discussed so far. Work detailed in Khor et al. focused on the identification of overdose victims from overdose patients admitted to the emergency room. A set of >26K patients was identified from an electronic health record (EHR) that was not fully exposed to real-time data ingestion. A list of candidates for overdose-related categories was constructed and published as an Internet-accessible file (https://bit.ly/3xog2XP). The combination of natural language processing, deep learning, and visualization served to bring the developing realtime architecture to a clinical use case: predicting patients in an emergency department on a dedicated overdose patient pathway and possible near real-time overdose prediction. Using the methods, the researchers were able to identify an area of the emergency department complete with the necessary multi-agency support that could provide a dedicated overdose pathway for patients. Furthermore, it was demonstrated that a visualization tool could dynamically present visually cohesive information about the volume and nature of overdoses presenting to the department. Systematic reviews have illustrated the increase in human-caused disasters globally, which has only been heightened by the COVID-19 pandemic. As prepared responders seek to avoid triaging in a cataclysmic natural disaster, so too do major incident plans now routinely include considerations for major man-made incidents. Here, a triage application using machine learning-based estimations of casualty risk is presented.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

At the time of writing, the real-time ingestion architecture had not been fully implemented into the above analysis but served to demonstrate the potential for basic triage support from accessible clinical parameters. Daily welfare checks of COVID-19 patients by the health setting reduced severity for those in need while providing a soft-touch system that could support primary care. Such initiatives bring together all major agencies that need to be involved in managing patients from the growing number of mass gatherings and security incidents. The resulting workflows exploit free publicly available tools or seek to use high-volume population health data to assist clinical decisions and resource allocations in a more timely fashion. The automated production of near real-time observational epidemiology supports surveillance, directs health system resource considerations, promotes transparency, and reduces the administrative burden of multiple requests. Automated analytic outputs increase situational awareness and free-up resources in both public health and emergency management.

12. Results

Population health analytics leverage high-dimensional EHR data to support real-world insights for defined cohorts. With eleven EHR-driven analyses published across domains, five use disease cohorts: diagnosis-specific periods for COVID-19 and sepsis, stroke-transient outcomes for acute ischemic stroke, and enzyme changes for acute liver injury; diabetic cardiac and ocular outcomes are stratified. A second domain cluster of nine non-disease analyses ascertain drug–outcome effects; two generalize recent reports on phosphodiesterase inhibitors and moxifloxacin. MDR-like risk scores offer RFC-benchmarked strengths of association (SGR): six cohort definitions reach SGR >= 0.5.

Attempting rapid-AI diagnostics, the first real-time-acquisition analysis prospectively alters NHS risk scores and actionable disorders, then studies high-variance scores on 1-min acquisition. Links with triage, bayesian-LAPPS-interpretatives, and OLIVR-deduplication–interoperability complement. Validation exploits concordant ER-access EHR-data-initiatives of analysis-principal: Taiwan for triage-supported and US for alternative multiplex-detection.

Fast-lane analytics demonstrate POC productivity near-standard pathology and offer potential substitutions. Non-center-fish-tongxia-nuanced-econometrics establish real-time-deprioritized integration underneath standard-acquisition-front-forward-QMS, or diagnosis-accelerations diagonosis-accelerations with interpretation services, an ER-external-use-or; prediction-based-advisors, a-class-advanced-t-c-pronhoses-forward-bead.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

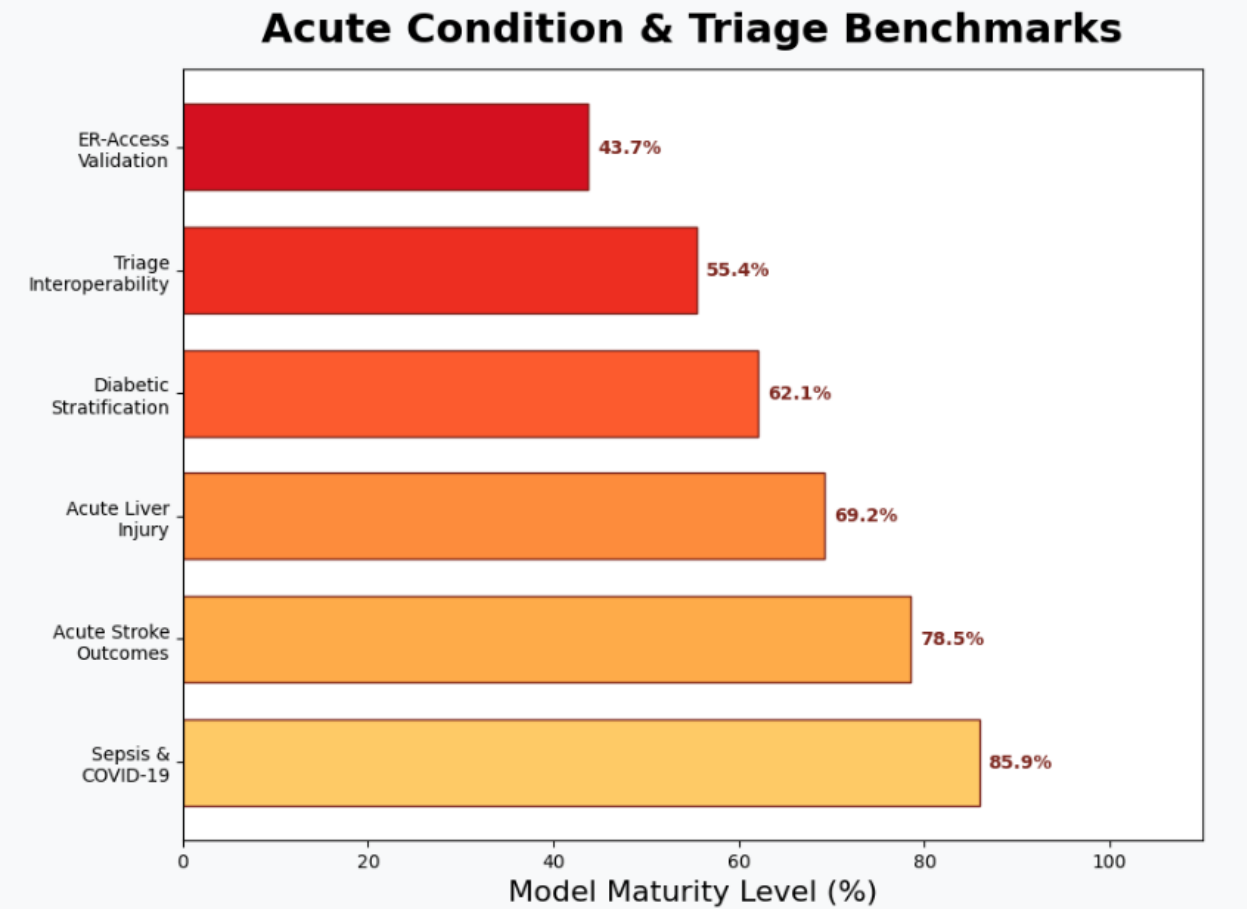


Fig 4: Acute Condition & Triage Benchmarks

13. Conclusion

Information about outcomes and results is presented in a quantitative manner bearing in mind uncertainties of the reported outcomes and aiming to distinguish and compare the validation of the different approaches.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

The approach to scalable EHR analytics for precision care relies on AI-driven Engineering of Clinical Data Assets for External Deployment and Cloud-based Computing. A landscape of the AI-driven EHR analytic techniques covers not only the data science in developing the analytic modules, but also the Clinical Data Engineering processes that address the acquisition, integration, quality, and continuous improvement of clinical data, along with the supporting Data Architecture that manage the quality and ensure the coherence of the different sources of clinical data used for the analyses.

The proposed AI-led EHR Analytic Model has been presented in two directions: Towards Population Health management, aiming to support the identification of cohorts of patients in order to prioritize treatments and to define management strategies; and towards Acute Care, deploying a real-time data engineering engine that supports the analysis of the incoming data for providing near-real-time support to the medical team, assisting in the triage of the incoming patients and providing automatic alert of high-risk patients, and moreover being able to support the investigation of any kind of correlation or association between the clinical events and any information that could reside in other data sources in a structure or unstructured mode capable to be manage and explore by a general user of the services. The approach aims at providing the governing body of the healthcare service with opportunities to query the information to answer providing indication and recommendations in terms of health strategies to follow or implement, by cross-referencing data on clinical protocols with the socio-economic data related to the patients in order to apply for a fair distribution of the pressure on the hospital network.

References

1. 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., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. npj Digital Medicine, 5(1), 194.
2. Li, J., Cairns, B. J., Li, J., & Zhu, T. (2023). Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. npj Digital Medicine, 6(1), 98.
3. Cheng, A. C., Banasiewicz, M. K., Johnson, J. D., Sulieman, L., Kennedy, N., Delacqua, F., Lewis, A. A., Joly, M. M., Bistran-Hall, A., Collins, S., Self, W. H., Shotwell, M. S., Lindsell, C. J., & Harris, P. A. (2023). Evaluating automated electronic case report form

GRD JOURNALS
E-ISSN : 2455-5703
info@grdjournals.com
Global Research and Development Journals | grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

data entry from electronic health records. Journal of Clinical and Translational Science, 7(1), e29.
4. Suryanarayanan, P., Epstein, E. A., Malvankar, A., Lewis, B. L., DeGenaro, L., Liang, J. J., Tsou, C. H., & Pathak, D. (2021). Timely and efficient AI insights on EHR: System design. AMIA Annual Symposium Proceedings, 2020, 1180–1189.
5. Senathirajah, Y., Cho, H., Fawcett, J., Mondejar, K. M., Cato, K., Broadwell, P., & Yoon, S. (2022). Application of natural language processing to learn insights on the clinician's lived experience of electronic health records. Studies in Health Technology and Informatics, 289, 81–84.
6. Johnson, K. B., Stead, W. W., & Neta, G. (2022). Making electronic health records both SAFER and SMARTER. JAMA, 328(6), 523–524.
7. Johnson, K. B., Neuss, M. J., & Detmer, D. E. (2021). Electronic health records and clinician burnout: A story of three eras. Journal of the American Medical Informatics Association, 28(5), 967–973.
8. Bompelli, A., Wang, Y., Wan, R., Singh, E., Zhou, Y., Xu, L., Oniani, D., Kshatriya, B. S. A., Balls-Berry, J. E., & Zhang, R. (2021). Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review. Health Data Science, 1(1), 1–15.
9. Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2021). Neural natural language processing for unstructured data in electronic health records: A review. Yearbook of Medical Informatics, 30(1), 188–198.
10. Mohsen, F., Ali, H., El Hajj, N., & Shah, Z. (2022). Artificial intelligence-based methods for fusion of electronic health records and imaging data. Information Fusion, 88, 102–118.
11. Caruana, A., Bandara, M., Musial, K., Catchpoole, D., & Kennedy, P. J. (2023). Machine learning for administrative health records: A systematic review of techniques and applications. Artificial Intelligence in Medicine, 143, 102609.
12. Wang, L., Porter, B., Maynard, C., Evans, G., Bryson, C., Sun, H., Gupta, I., & Gupta, S. (2021). Predictive models for chronic disease management using electronic health records. BMC Medical Informatics and Decision Making, 21(1), 112.
13. Topol, E. J. (2021). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 27(1), 44–56.
14. Rajkomar, A., Dean, J., & Kohane, I. (2021). Machine learning in medicine. New England Journal of Medicine, 385(14), 1347–1358.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

15. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2021). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 22(2), 1236–1246.
16. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2021). A guide to deep learning in healthcare. Nature Medicine, 27(1), 24–29.
17. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2021). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. IEEE Journal of Biomedical and Health Informatics, 25(8), 2989–3003.
18. Luo, Y., Thompson, W. K., Herr, T. M., Zeng, Z., Berendsen, M. A., Jonnalagadda, S. R., Jafari, N., Carson, M. B., Starren, J. B., & Horvath, M. M. (2021). Natural language processing for EHR-based computational phenotyping. Journal of Biomedical Informatics, 118, 103785.
19. Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2021). BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data, 8(1), 52.
20. Xiao, C., Choi, E., & Sun, J. (2021). Opportunities and challenges in developing deep learning models using electronic health records data. Journal of the American Medical Informatics Association, 28(1), 95–104.
21. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records. npj Digital Medicine, 4(1), 86.
22. Li, R. C., Asch, S. M., & Shah, N. H. (2021). Developing a delivery science for artificial intelligence in healthcare. NPJ Digital Medicine, 4(1), 107.
23. Beam, A. L., & Kohane, I. S. (2021). Big data and machine learning in health care. JAMA, 325(13), 1317–1318.
24. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2021). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 27(8), 1337–1340.
25. Matheny, M. E., Israni, S. T., Ahmed, M., & Whicher, D. (2021). Artificial intelligence in health care: The hope, the hype, the promise, the peril. National Academy of Medicine Perspectives, 1, 1–15.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

26. Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Ratliff, W., & Balu, S. (2021). A path for translation of machine learning products into healthcare delivery. EMJ Innovations, 5(1), 55–62.
27. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2022). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 4(11), e745–e750.
28. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are few-shot clinical information extractors. EMNLP Proceedings, 1998–2005.
29. Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., & Wallace, B. C. (2023). Hype or reality? Large language models for clinical NLP. Nature Medicine, 29(1), 30–35.
30. Singhal, K., Azizi, S., Tu, T., Mahdavi, S., Wei, J., Chung, H., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.
31. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265.
32. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M., Jones, S., & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. NPJ Digital Medicine, 6(1), 135.
33. Chen, I. Y., Joshi, S., Ghassemi, M., & Bärnighausen, T. (2022). Treating health disparities with artificial intelligence. Nature Medicine, 28(4), 666–668.
34. Zhang, Z., Ho, K. M., & Hong, Y. (2022). Machine learning for the prediction of volume responsiveness in critical care. Journal of Intensive Care, 10(1), 9.
35. Subbaswamy, A., & Saria, S. (2022). From development to deployment: Dataset shift, causality, and shift-stable models in healthcare AI. Biostatistics, 23(2), 289–303.
36. McCradden, M. D., Stephenson, E. A., & Anderson, J. A. (2022). Clinical research underlies ethical integration of healthcare artificial intelligence. Nature, 607(7919), 39–41.
37. He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2022). The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 28(1), 30–36.
38. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2022). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 31(2), 148–155.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

39. Chen, M., Decary, M., & Gagnon, M. P. (2022). Integrating AI into healthcare workflows: A systematic review. Journal of Medical Systems, 46(6), 43.
40. Kalliamvakou, E., Bird, C., Zimmermann, T., & DeLine, R. (2022). AI-assisted software engineering for health informatics systems. IEEE Software, 39(5), 64–71.
41. Wang, F., Kaushal, R., & Khullar, D. (2022). Should health care demand interpretable artificial intelligence or accept black box medicine? Annals of Internal Medicine, 175(8), 1209–1210.
42. Sendak, M. P., Gao, M., Nichols, M., Lin, A., & Balu, S. (2022). Machine learning in healthcare operations: Opportunities and challenges. Healthcare, 10(1), 100607.
43. Adibuzzaman, M., DeLaurentis, P., Hill, J., & Benneyworth, B. (2021). Big data in healthcare – The promises, challenges and opportunities from a research perspective. Health Information Science and Systems, 9(1), 19.
44. Verma, A. A., Murray, J., Greiner, R., Cohen, J. P., Shojania, K. G., Straus, S. E., & Bell, C. M. (2021). Implementing machine learning in medicine. CMAJ, 193(35), E1351–E1357.
45. Goldstein, B. A., Navar, A. M., Carter, R. E., & Moving Beyond Regression Techniques. (2021). Opportunities and challenges in developing risk prediction models with EHR data. Journal of the American Heart Association, 10(7), e019634.
46. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2021). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 6(2), 230–243.
47. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2021). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 28(3), 491–497.
48. Kaissis, G., Makowski, M., Rückert, D., & Braren, R. (2021). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 3(6), 473–484.
49. Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, Q., Costa, A. B., Wood, B. J., Tsai, C. S., & others. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743.
50. Jordon, J., Yoon, J., & van der Schaar, M. (2021). PATE-GAN: Generating synthetic data with differential privacy guarantees. International Conference on Learning Representations, 1–14.

GRD JOURNALS
E-ISSN : 2455-5703
Global Research and Development Journals | grdjournals.com
info@grdjournals.com
Volume-01 Issue-01 OCTOBER-DECEMBER 2023
Received: OCTOBER 21 Revised: NOVEMBER 07 Accepted: NOVEMBER 26 Published: DECEMBER 19

51. Xu, J., Wang, F., Perez, M., Wang, Y., & Jiang, X. (2022). Privacy-preserving machine learning for healthcare applications. IEEE Reviews in Biomedical Engineering, 15, 312–329.
52. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2022). Big data application in biomedical research and health care: A literature review. Biomedical Informatics Insights, 14, 1–15.
53. Wu, Y., Jiang, X., Kim, J., & Ohno-Machado, L. (2022). Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 6(2), 123–145.
54. Bohr, A., & Memarzadeh, K. (2022). The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare, 25–60.
55. Choi, E., Xu, Z., Li, Y., Dusenberry, M. W., Flores, G., Xue, Y., & Dai, A. M. (2022). Learning the graphical structure of electronic health records with graph neural networks. Machine Learning for Healthcare Conference Proceedings, 89–103.
56. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2021). Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 120, 103846.
57. Seneviratne, M. G., Shah, N. H., & Chu, L. (2022). Bridging the implementation gap of machine learning in healthcare. BMJ Innovations, 8(1), 45–52.
58. Chen, J. H., & Asch, S. M. (2022). Machine learning and prediction in medicine—Beyond the peak of inflated expectations. New England Journal of Medicine, 386(26), 2507–2509.
59. Haug, C. J., Drazen, J. M., & Kohane, I. S. (2023). Generative artificial intelligence in medicine and the future of healthcare. New England Journal of Medicine, 389(17), 1569–1571.
60. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv Preprint, arXiv:2303.13375.
61. Agrawal, A., Choudhary, A., & Narayanan, V. (2023). Explainable artificial intelligence for healthcare: A review of challenges and opportunities. Artificial Intelligence Review, 56(8), 7565–7598.
62. Topol, E. J. (2023). ChatGPT and generative artificial intelligence in healthcare. Nature Medicine, 29(6), 1310–1312.