



AI-Engineered EHR Intelligence for Precision Care

1st Murali Malempati , 2nd Robert Martinez

Murali Malempati, Senior Software Engineer , mmuralimalempati@gmail.com

Robert Martinez Research Assistant

Abstract

The growing availability of clinical data across Electronic Health Records (EHR) systems has created an urgent need for AI-driven data engineering solutions that enable scalable EHR analytics and precision healthcare decision-making. To address this need, this chapter presents a scalable, AI-driven clinical data engineering framework that automatically curates and prepares clinical data for a wide range of downstream applications.

The framework is built around a clinical data engine that uses deep learning models for natural language processing (NLP) of clinical text and Clinical Concept Extraction (CCE) to extract clinical concepts from notes and structured tables. This enables clinical data preparation for large-scale research that is efficient, accurate, and easy to implement. The engine integrates four recursive deep learning components, each addressing a distinct stage of the data preparation pipeline:

- Clinical Text Recommending System (CTRS) – supports small-data EHR analytics based on clinical text
- Clinical Data Preparation Engine (CDPE) – curates multi-source, multi-type clinical data for both large-scale analysis and exploratory research
- Automated AI-driven Chatbot for Answering Clinical Questions (AICACQ) – answers clinical questions efficiently and accurately
- Clinical Concept Recommendation System (CCRS) – supports CCE-related recommendation tasks

The framework's novelty lies in two areas: the deep lifecycle relationships linking these components (and their extensibility to new applications), and its ability to drive clinical-data preparation according to analytic demand. By supporting a wide range of tasks across the clinical research pipeline, the engine scales flexibly with both the complexity of analytical demands and the diversity of clinical data sources.

Ultimately, by enabling task-recommendation-driven, data-engine-supported AI application development, this framework empowers both front-end designers and back-end developers of EHR analytics systems — bridging the gap between clinical analytics and the infrastructure that supports it.

keywords: Clinical Data Engineering; Clinical Data Lake and Data Warehouse; EHR Analytics; EHR as Data Source; Data Compression in EHR; Precision Healthcare Decision Systems; Dynamic Scalable Architecture; Markov Decision Process; Reinforcement Learning.



1. Introduction

Clinical analytics and decision-making at scale across institutions and geographies demand integrated access to electronic health record (EHR) patient data without patient identifiers. Overcoming the institutional and practical challenges of building shared, de-identified, heterogeneous, real-world data sets, moreover, requires automated data cleaning, firm-deidentifier and restructuring, AI-scaling of data engineering, and quality control of the latter using synthetic data generation. Using LSTM neural networks for imputing missing data; and the generation of diverse synthetic patients for quality assurance is demonstrated on the MIMIC-III EHR data set.

Collaborations on global COVID medical analytics illustrate the realisation of population-level decision-support systems. A structured survey of the population-level COVID response is analysed using graph theory-based clustered aggregation for discovering the dominant combinations of pandemic containment actions across 181 political jurisdictions. A hybrid supervised-unsupervised framework integrates these findings as external knowledge, generating predictions for case load and infection fatality rate across over 200,000 locations worldwide. A novel explainable deep-learning system designed for EHR analogs supports clinical information retrieval and decision-support systems. COVID-19 research is used for empirical validation, demonstrating explainable, hierarchical, patient-stratified modelling for an abundant selection of clinical endpoints.

1.1. EHR-Driven Phenotyping

The increasing availability of Electronic Health Record (EHR) data provides new opportunities for precision medicine and drug development. The transition from a symptom-based drug development strategy to a disease-target strategy began about a decade ago, and many pharmaceutical companies are focusing their efforts on therapies for genetically defined strains of diseases characterized by heterogeneous patient populations defined by different biological pathways. The ability to identify well-defined patient populations for clinical trials is critical for improving the chance of success.

EHR data have not yet been used in a significant way for drug-target disease-characterization, although for an earlier purpose, namely the identification of drug side effects and adverse drug reactions. Phenotype stratification of disease cohorts based on a systematic approach to the medical ontological vocabulary and disease markers is important for supporting drug repositioning and associated clinical trials. EHR data can be considered to provide an ontology-based triggering of disease-targeted phenotypes, providing stratified patient cohorts for various drugs and diseases.

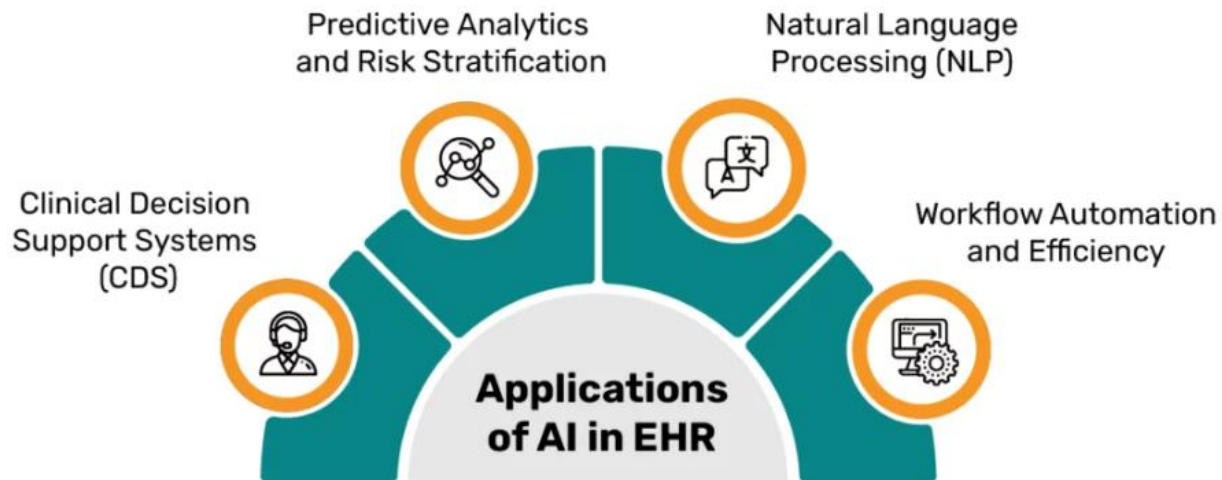


Fig 1: Clinical Data Engineering for Scalable EHR Analytics

1.2. Precision Oncology and Genomic Integration

Another promising area for precision medicine applications of EHR and clinical imaging data methods is precision oncology, which aims to optimize treatment selection for cancer patients. For several tumor types, a growing set of genomic data has been made available as part of the publicly available Genomic Data Commons (GDC) repository from The Cancer Genome Atlas and Genomic Data Commons initiatives. However, the power of these novel datasets derived from discrete genomic measures and analyses remains limited. Approaches that combine EHR, clinical imaging, and genomic data have been proposed to leverage all possible clinical data sources to select clinical therapy for the right patient at the right time, with the goal of increasing response to therapy, improving prognosis and reducing the risk of adverse drug effects. Data engineering and new methods that extend existing work in these areas are needed to facilitate the development of integrated EHR-genomics foundations for precision cancer therapy selection.

These integrated EHR-genome methods facilitate the automatic learning of therapy outcome phenotype trees, which can be further validated using the genomic data. Machine learning is used to directly model the mapping from the complex high-dimensional and sparse genomic data to the therapy outcome tree parameters. Dynamic templating based on the learned induction policies allows the unified modeling of different therapy types, as well as the combination of different data sources, such as a patient's complete genomic profile and the cumulative genomic similarities with the other patients in the dataset. The learned therapy selection therapies are positioned in the context of an additional tumor-free overall survival requirement.

1.3. Population Health Analytics and Risk Stratification

Increasing stratification of population health allows healthcare research to expand toward timelier and more relevant preventive and therapeutic interventions, which include Community Network Homelessness and Health Services Data, Health System Readmission and Discharge Services Data, Security Data, Campus Crime Reports, and Climate and Weather Data.



Recent works offer concise and robust prediction indicators as well as tailored recommendation services to minimize winter season health risks. Built on long-term historical electronic health records, Health System Administrative Data and other publicly available data collected between 2004 and 2016, a 4-year experimental time period allows a sufficiently stable temporal distribution for training. A neighborhood-level temporal clustering approach further shapes health risk predictions with resident migration in consideration. Specifically, supervised winter season chronic and acute illness risk predictions reveal a balanced performance among different classifiers, while the ensemble model peaks for the serious-risk level. Risk stratification for elderly and homeless groups attracts dedicated exploration. Neighborhood-level homelessness risk and immigrant avoidance recommendations achieve favorable completeness.

Recent research illustrates a novel, two-level fault-tolerant, ten-layer predictive engine for community-problem and service-demand forecasting. Real-time, software-as-a-service web-based interactive delivery caters to interactive decision-making needs. An ensemble prediction strategy adopts binary classification for the upper level, while regression for the lower level. Ten-fold cross-validation demonstrates that the stratified K-fold technique enhances accuracy and reduces predictive uncertainty compared to the baseline.

Equation 1: Disease-risk prediction with logistic regression

$$p(y = 1 | x) = \sigma(z) = \frac{1}{1 + e^{-z}}, z = w^T x + b$$

Step-by-step derivation

Start with the linear score:

$$z = w^T x + b$$

For a patient with features

$$x = [x_1, x_2, \dots, x_n]^T$$

and weights

$$w = [w_1, w_2, \dots, w_n]^T,$$

the dot product is

$$w^T x = w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

So:

$$z = w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b$$

Now convert this real-valued score into a probability using the sigmoid:



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Substitute $z = w^T x + b$:

$$p(y = 1 | x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Expand if needed:

$$p(y = 1 | x) = \frac{1}{1 + e^{-(w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b)}}$$

2. Foundations of Clinical Data Engineering

A clinical data engineering (CDE) approach is foundational for the capabilities described. CDE exploits the convenience of big data while accounting for the particular challenges imposed by clinical data: their multimodal, heterogeneous, and information-rich nature combined with the requirements imposed for patient safety in healthcare environments. In particular, it provides the engineering framework for the graphical models required in recent research on explainable AI; the identification of clinically relevant features in the data; and the required linking to domain knowledge. CDE delivers a scalable infrastructure for the development and realisation of precision decision systems in the clinical domain.

Most of the large-scale clinical studies result in a model or algorithm that is subsequently deployed in the healthcare environment. However, the patient groups identified and employed in these studies are usually relatively small (hundreds of patients) when compared with the patient flows of treating hospitals or healthcare regions (several millions). Despite this fact, the models often do not scale for daily practice since they are not able to capture the complexity of clinical practice nor cover all possible cases at the same time. Domain knowledge is either omitted, underused or not accessible in the final decision-making system. Recently developed precision healthcare approaches allow personalised recommendations for different ODEU participants by addressing the possession of a clinical decision support system with explanatory capabilities in AI-supported environments. The additional urban data sources pose new questions that require an integration between the CDE infrastructure and the new data available.



Fig 2: Clinical Data Engineering

2.1. Data Provenance and Governance

The availability and accessibility of large amounts of heterogeneous data created new opportunities for a variety of applications. However, these new methods pose data and methodological challenges that have not been adequately addressed. First, provenance is often lost in exploratory analyses and aggregation, resulting in a lack of transparency and, more important, generalizability of the results. With loss of recording, transparency for interpreting and understanding results is lost and the decision rules applied to the data during exploration, cleaning, aggregation, and interpretation are not retained. Therefore, methods are needed to consistently record, store, and visualize analytical provenance at a level of detail that can aid interpretation and generalization of the analytical results. Second, data provenance is needed to facilitate data governance when EHR data are shared internally and externally.

The increasing complexity of modern data analysis, composition of other experts has raised awareness of the necessity to not only curate and document the data but also to manage the data flows through building trusted data provenance frameworks. The proposed solutions facilitate trusted sharing of EHR data for analyses such as identifying severity scores for tracking of patients infected with SARS-COV-2 during the global pandemic. Such works provide a basis for transparency of analytical workflows for higher levels of accountability by record-keeping during the course of the analysis and simplify the ready adaptation of methods when applied to other temporal or longitudinal data. Furthermore, a complementary methods framework concept is highlighted to address the loss of recording during exploratory EDA.



Domain	Key focus in the article	Main contribution to EHR analytics	Expected healthcare impact
Clinical Data Engineering (CDE)	Automated curation, preparation, and structuring of heterogeneous clinical data	Builds a scalable framework for turning raw EHR data into analytics-ready datasets	Enables large-scale clinical research and operational decision support
EHR-driven phenotyping	Identifying disease-targeted phenotypes and stratified patient cohorts	Supports cohort discovery from routine EHR data	Improves trial recruitment, disease characterization, and precision medicine
Precision oncology and genomic integration	Combining EHR, imaging, and genomic data	Learns therapy-outcome relationships from multimodal sources	Improves therapy selection and prognosis estimation
Population health analytics	Risk stratification using EHR plus social/community/environmental data	Extends clinical analytics beyond hospital-only data	Enables preventive interventions and community-level health planning

Table 1: Core domains and contributions of the proposed AI-driven clinical data engineering framework

2.2. Semantic Interoperability and Standardization

As pointed out by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, health IT providers were required to ensure EHR interoperability or health information exchange (HIE) across health care organizations by satisfying International Organization for Standardization (ISO), the American National Standards Institute (ANSI), LOINC, and other standards relevant to the ability to share information among providers, labs, and patients. Within the scope of research and public health, Semantic Web technologies have in mind achieving a cost-effective way of creating polypathic, analogical, scaling, and cooperating systems that would massively share, use, acquire, and learn from distributed, heterogeneous, and possibly noisy data sources within a privacy-preserving framework. The expressiveness of Semantic Web languages, such as RDF(S), OWL, RDFS, SPARQL, SWRL, and others, focuses on Shareable, Easily-Interoperable, and Analyzable Data.

Community development of semantically annotated Web resources and organization of the heterogeneous data sources into distributed knowledge bases along with a plethora of Semantic Web enabling tools and technologies have proven that Semantic Web technologies enable achieving the above requirements at low costs while camouflaging the technical difficulties of resource sharing and usage. The Semantic Web data description formalism of OWL2 DL language contributes to data shareability, data interoperability, and data analysis feasibility, thus supporting clinical and medical HIE in the framework of clinical and biomedical research and public health. Building an HIE infrastructure requires SKOS-based vocabularies (or ontologies) that define the semantics of the relevant RDF graphs and a layer of specialization ontologies that exploits the expressiveness of OWL2 EL and OWL2 QL languages.



2.3. Data Quality and Curation

Poor data quality constitutes a hurdle for clinical analytics and applications based on electronic health records (EHRs). To aggregate longitudinal data from multiple sites—including data from up to 20 cancer centers in the National Cancer Institute’s Cancer Biomedical Informatics Grid (caBIG)—an appropriate data-cleansing strategy was designed, as described elsewhere in this volume. It consisted of three modules for removal of low-quality data, logically inconsistent data, and unrealistic or impossible data, respectively (see Section 2.3.1). Because the development of high-throughput clinical data pipelines requires the execution of data extraction, transformation, and loading (ETL) scripts into public cloud systems, the ETL scripts target and curate one database at a time. Cloud-based ETL execution is achieved using Amazon Web Services and the cloud-based deployment of the OpenStack JupyterHub service (<http://jupyter.org/>) using NumPy, SciPy, and Pandas. The bene-fit of this approach is that it can easily scale to multiple sites—each with its own JupyterHub implementation—by executing the ETL scripts in parallel for each site database.

Data-engineering methods have been implemented that are descending from traditional data cleaning for consistency to state-of-the-art natural-language-processing techniques that parse clinical free-text narratives into modeled knowledge representation. The pipeline for clinical data engineering uses several levels of curation, beginning with the more structured time-series data and climaxing with the clinical narrative text sources (see Section 2.3.2). The different levels are grouped into clean, curated, and structured levels of curation. The clean data are generated by using three different data-cleansing modules that are designed to remove low-quality data, logically inconsistent data, and unrealistic or impossible data, respectively. Data at the curated level have been subjected to different expensive curation techniques—developed through applied AI software engineering—that are not run routinely because of their high computational resource demand. The clinical free-text narratives are parsed into modeled knowledge representation using natural-language processing and natural-language understanding techniques. Passage detection using neural networks, author assignment, and high-level semantic content analysis of the source narratives are among the available NLP methods.

Equation 2: Binary cross-entropy loss for clinical outcome prediction

$$\mathcal{L} = -[y\log(\hat{y}) + (1 - y)\log(1 - \hat{y})]$$

Step-by-step derivation

Let:

- $y \in \{0,1\}$ be the true label
- $\hat{y} = p(y = 1 | x)$ be the predicted probability

For a Bernoulli outcome, the likelihood is:

$$P(y | x) = \hat{y}^y (1 - \hat{y})^{1-y}$$

Take the log:

$$\log P(y | x) = \log(\hat{y}^y (1 - \hat{y})^{1-y})$$



Using $\log(ab) = \log a + \log b$:

$$\log P(y | x) = \log(\hat{y}^y) + \log((1 - \hat{y})^{1-y})$$

Using $\log(a^c) = c \log a$:

$$\log P(y | x) = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})$$

Loss is negative log-likelihood:

$$\mathcal{L} = -\log P(y | x)$$

Therefore:

$$\mathcal{L} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

3. Architecture of Scalable EHR Analytics Platforms

A prerequisite for producing personalized medicine-related recommendations is integrating the analytical capabilities for the predictive models with electronic health records. In order to produce a health ecosystem that scales to both patient numbers and complexity, system architecture must be designed and adapted. A RAKING framework has been proposed, which is composed of three well-defined steps under a new architecture model. The first step is Record linkage and Knowledge integration, which forges the foundation of a Predictive model development database.

The second step is Adaptation of the predictive model and the third step is a Knowledge-based health ecosystem for third parties. The third step integrates five analysis disciplines: Computer-aided disease detection, Disease prediction, Disease risk assessment, Treatment outcome prediction and Health risk assessment. The Predictive model development database enables heterogeneous predictive models, built for sources beyond the health domain, to link and adapt using patients' EHR data. As an illustration of its use, the key Predictive model development database for the Health Risk Assessment is outlined.

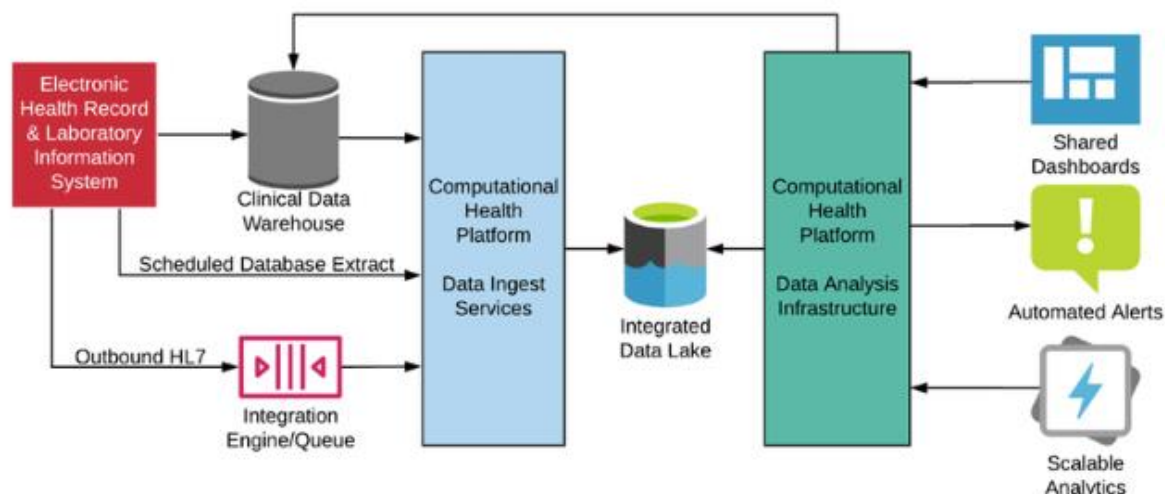


Fig 3: Architecture of Scalable EHR Analytics Platforms

3.1. Data Ingestion and Extraction

The first step is to transform raw EHR data into a data model that facilitates data analysis and the implementation of a decision support engine. The most voluminous and complex data typically reside in NoSQL document collections and are generated by message-oriented middleware systems, which decouple the actors of a use case. The messages store the activity recorded during the tasks performed on service request instances by all users in a workflow scenario. The data model maps two collections: task and concept, for encoding both the actions in the workflow and the domain participation in the application layer (e.g., disease, drug, laboratory test, therapy).

Further non-textual collections are dealt with from collections labelled image, as EHR images are typically stored in internal information repositories of the provider's organisation platform. Throughout the overall process, Natural Language Processing (NLP) techniques are adopted for detecting, filtering, tokenising, and lemmatising natural language texts in messages. The methodologies for extracting textual information from images, if any, and for what reasons, should also be in the scope of analysis. This could also cover similar extraction from pre-existing clinical documents: clinical findings, discharge letters, etc., supporting the residual feed to analysis. Doctors support knowledge creation explicitly in the EHR by entering findings in the free-text fields proposed by the application. Data in those fields are usually huge and seldom used in previous research, as the volume and possibly confusing content look daunting. NLP applied exclusively to those EHR fields should enhance extraction processes by detecting disease-break levels suitable for the explored domain, thus yielding many documents to be aligned with existing knowledge in explicit or implicit form.

3.2. Storage Architectures for Clinical Data

Storage architectures for clinical data evolve in tandem with data volume and the demands and resources of information users. As applications mature and data need to support analytics, a schema on write model, augmented data representations that enable



semi-structured and unstructured data analytics, and large-scale de-identification processes provide native support for near real-time business intelligence and precision healthcare decision systems.

EHR data volume continues to grow rapidly, spurred on both by increasing clinical data capture and by population studies that access and aggregate decade-long clinical histories on large cohorts. The predictive capabilities of AI methods and the use of clinical natural language processing open the door for the use of unstructured parts of clinical records or conversational datasets such as clinical portals and physician-notes, thereby driving the need for NLP-related storage layers. Private-public federated deployments behave similarly for these types of data because of the need to guarantee data privacy in the analytics of conversation transcripts.

Different needs, different architectures. Early EHR analytics relied largely on classical BI based on aggregate summaries produced by ETL processes running on a schedule. For these, a traditional relational warehouse with a schema on write model showed to be adequate. However, with the increasing demand for near real-time BI and with data-science-like use cases that require the data engineers to generate ad-hoc views on the data, biomedical observatory architectures have gained traction. These architectures have naturally expanded storage to include NoSQL or raw data lakes, avoiding the upfront data structure schema on write holds and widening the range of accessible data to become the information supply for university research cores.

Equation 3: Sensitivity and specificity

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{Specificity} = \frac{TN}{TN + FP}$$

Step-by-step derivation of sensitivity

Sensitivity means:

among all truly positive patients, how many did the model correctly identify?

All actual positives consist of:

- True positives: TP
- False negatives: FN

So total actual positives:

$$\text{Actual Positives} = TP + FN$$

Correctly detected positives are TP , hence:

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{Correctly predicted positives}}{\text{All actual positives}} \\ \text{Sensitivity} &= \frac{TP}{TP + FN} \end{aligned}$$



Step-by-step derivation of specificity

Specificity means:

among all truly negative patients, how many did the model correctly identify?

All actual negatives consist of:

- True negatives: TN
- False positives: FP

So total actual negatives:

$$\text{Actual Negatives} = TN + FP$$

Correctly detected negatives are TN , hence:

$$\text{Specificity} = \frac{\text{Correctly predicted negatives}}{\text{All actual negatives}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

3.3. Processing Frameworks for Real-Time and Batch Analytics

An overview of recommended frameworks to support real-time and batch analytics of clinical data, is discussed next. Lambda and kappa architectures are widely accepted process models for scalable data processing in EHR analytics and operational monitoring.

The lambda architecture defines two pipelines—one for real-time analytics and the other for batch analytics—alongside a serving layer that combines the results of the two pipelines to simplify user queries. The mechanism guarantees access to the latest updates using the real-time layer while also supporting long-term queries on the batch layer. The kappa architecture simplifies the lambda approach by having only a single pipeline for all data-processing jobs, thereby removing the associated maintenance overheads. Nevertheless, several researchers report queries with kappa pipelines running into performance bottlenecks owing to slow data-processing speeds of stream-processing systems like Apache Storm and Apache Flink. Moreover, distributed batch-processing systems like Spark SQL, Hive on Tez, and Impala often outperform stream-processing frameworks for large batches of data. These considerations have led to the proposal of a modified lambda architecture that substitutes the kappa stream-processing pipeline with a high-throughput batch-processing pipeline.

4. AI Methods for Precision Healthcare

The AI-assisted pipeline for precision healthcare is guided by the advanced AI cloud through an AI-driven approach supported by empirical data. The innovations in precision healthcare rely on a unique eDNA concept, described in detail in Chapter 3, that

relies on the integration of supervision, measurement/observation, assessment/evaluation, and knowledge generation with controlled corrections.

Within the framework of precision healthcare, the proposed message-passing model replicates the diffusive communication infrastructures of human organs and organs' systems. The clinical interpretation of biological samples, such as blood, urine, faeces, and others, can thus be simplified by focusing AI-assisted connections on only a small class of diseases, which is especially critical for Ai-assisted long-core COVID analysis. The smart eye networks comprise any number of nitrocellulose strips locally connected according to their measurements, which can be intelligently exploited by a real network of smart cameras as offered by Meta's avatar concept.

The second-phase precision healthcare system aims to receive all accessible organ-transmitted eDNA messages even in the case of very few samples analysed. The automatic labelling of analysed messages allows the adopted AI cloud to rescale the connections with the other organs' systems and the reconstruction quality before sending them to the ICO. To cope with the current extreme research acceleration, the launched AI cloud must fulfil a very important task: allow all research groups worldwide to propose their best working hubs even if they include only few analysed messages. In this case, a specific decision-making eDNA set defines all accessible communications with the deconcentration commands.

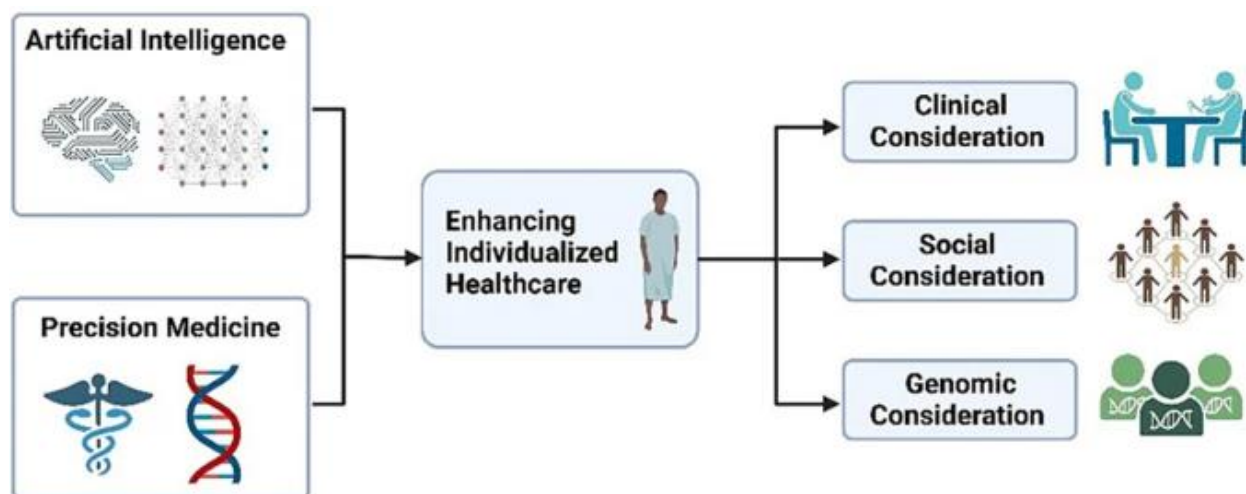


Fig 4: The integration of AI and precision medicine

4.1. Diagnostic and Prognostic Modeling

Accurate decision-making is an indispensable condition for a successful medical and therapeutic process. The correctness and speed of diagnosis, the planning of therapeutic measures, and the correct prediction of possible complications are key factors determining the success of therapy. At the present stage of medical science development, there is a system of formalized models that makes it possible to compare the application of specialized preparation by specialists of different qualifications and their specialists for decision-making regarding diagnosis, diagnosis, prognosis, and therapy of diseases. Statistical methods play a leading role in solving such tasks.



Models based on specialized preparation allow the physician to carry out the same operations as a mathematician. For the prediction of efficiency and the choice of the optimal type of therapy of a patient with appendicitis, simulation models predicting the rate of postoperative complications are built, on the basis of this model the greatest influence on the complication rate is determined and the more advantageous variant of antibiotic therapy is chosen. Models for pre-operative diagnosis of pancreatic carcinoma, differential diagnosis between pancreatic carcinoma and chronic pancreatitis and prognosis of possible complications of cholecystitis are constructed. These models are intended for a wide audience of physicians, without requiring them to possess special mathematical training, to help doctor with data collected preoperatively.

Equation 4: k-anonymity constraint for EHR de-identification

$$\forall q, |E(q)| \geq k$$

where $E(q)$ is the equivalence class of records sharing the same quasi-identifier pattern q .

Step-by-step derivation

Suppose each patient record contains quasi-identifiers such as:

- age group
- ZIP prefix
- sex

When several records share the same quasi-identifier combination, they form an equivalence class $E(q)$.

If a quasi-identifier pattern is

$$q = (\text{Age } 40 - 49, \text{ZIP } 123 **, \text{Female})$$

then $E(q)$ is the set of all records with that same pattern.

To guarantee k-anonymity, no patient should be distinguishable from fewer than k patients.

So the size of each equivalence class must be at least k :

$$|E(q)| \geq k$$

And this must hold for every quasi-identifier group q :

$$\forall q, |E(q)| \geq k$$

Example: if $k = 5$, then every quasi-identifier group must contain at least 5 records:

$$|E(q)| \geq 5$$



4.2. Treatment Personalization and Decision Support

Rapid developments of deep learning tools incite excitement in the entire healthcare community with a variety of methodological contributions and clinical applications. Their application to the precision medicine problem continues to follow a data-centric trend, moving towards models trained on large cohorts of medical images and clinical records. The main limitation in the quest for precision in these models comes from their generalization capability, highly influenced by population heterogeneity and the difficulty to disentangle spurious correlations appearing in non-typical populations.

Comparative diagnostic studies evaluate the performance of deep-learning models across populations. However, guidelines seldom recommend their direct use. In the absence of ground-truth, distinguishing between model performance and spurious correlations becomes difficult and, without principled adjustment or selection, decision support based on those models remains risky.

Generative models estimate the joint distribution of images and diagnoses from healthy controls. When tested on unseen subjects, likelihood ratios, however, indicate a generalization risk that compromises the reliability of predictions across population pairs. Despite enjoying successful use in other domains, national cohort stratification by generative likelihood does not completely eliminate generalization risk. Control-coordinated conditional distributions appear useful for adaptation; yet, simple stratification by sexes proves more effective.

Adaptation, selection, and end-to-end training of clinical deep-learning decision-support systems resting on single-condition generator-embedded diagnostic classifiers illustrate the underlying concept. A network pretrained over a large cohort components-free frontal chest radiograph image repository enables accurate pneumonia detection, despite drastic interpopulation characteristics mismatch, and approaches performance closeness to a model specially trained on the target test set after adaptation.

4.3. Causal Inference and Counterfactual Reasoning

Counterfactuals allow the reasoning about what would have happened, had the facts been different. It enables one to know the outcome of a treatment on an individual level when that individual has not been treated. A causal effect can be understood from two points of view: (1) the average causal effect on the population (the standard point of view) and (2) the causal effect on a particular individual (the counterfactual point of view).

In order to answer the first question, one needs random assignment of treatment and control groups. In the absence of such random assignment, one relies on causal models, particularly graph-based causal models based on Jaynes's principle of maximal entropy. These graph-based causal models allow development of a causal-inference engine. That engine can be used to estimate both the average total causal effect of a treatment and the average causal effect of the treatment for each individual. The average causal effect can then be estimated on the population using stratification on the covariates of the biased nodes. One can translate such a causal model into a probabilistic model and fit it to observed data.

Architectural layer	Main functions	Technologies / methods mentioned	Output
Data ingestion and extraction	Transform raw EHR records, messages, images, and free text into analyzable representations	NoSQL collections, NLP, tokenization, lemmatization, concept extraction	Structured clinical entities and workflow events



Architectural layer	Main functions	Technologies / methods mentioned	Output
Data storage layer	Store structured, semi-structured, and unstructured clinical data at scale	Relational warehouses, data lakes, NoSQL storage, NLP-oriented storage layers	Analytics-ready storage for BI and AI
Data processing layer	Support both real-time and batch analytics	Lambda architecture, kappa architecture, Spark SQL, Hive on Tez, Impala, Apache Storm, Apache Flink	Timely and large-scale computation over EHR data
Clinical data preparation engine	Curate multi-source and multi-type clinical data	AI-driven data preparation, recursive deep learning components	Prepared datasets for exploration and modeling
Clinical concept extraction / recommendation	Extract and recommend clinical concepts from notes and tables	NLP, CCE, recommendation systems	Enriched structured knowledge from unstructured records

Table 2: Scalable EHR analytics architecture derived from the article

5. Data Privacy, Security, and Compliance

Data privacy and compliance with related laws and regulations underpin user trust in health information technologies and services. Design and operation must meet jurisdictional mandates for the use, storage, and sharing of sensitive health, personal, and demographic information. General Data Protection Regulation (GDPR) is a landmark regulation in the European Union (EU) that sets a new standard for the global treatment of data privacy and security and provides enhanced data protection rights for citizens with respect to their data privacy in society. While it is focused on the protection of personal data within the EU, it also addresses the transfer of personal data to countries outside the EU that do not have similar privacy protections.

As health data is often subject to additional privacy, security, confidentiality, and ethical requirements that exceed those of other organization types, specific designs for these systems may also have to meet jurisdictions established by various health information protection laws and additional rules for health systems and related operators, institutions, and designated data repositories. The USA, for example, has developed the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which is the US government's first step in the development of broad-based protections for personal health information and was developed to implement provisions of the HIPAA. More specifically, the Privacy Rule establishes the first national standards for the protection of certain health information and set a national standard for the safeguarding of protected health information that is maintained or transmitted in electronic form. All these must be considered in the context of scaling and meeting the anticipated demand and subsequent trust and privacy requirements of the system.

5.1. De-identification and Privacy-Preserving Computation

Securing personal data while preserving effective data utility is crucial for a realistic large-scale EHR analysis and longitudinal precision medicine initiative. A proper strategy for data protection in EHR information must satisfy the strictest privacy conditions, which include data anonymity, unlinkability, and resistance to reconstruction. The anonymization of records via

standard k-anonymity achieves similar protection against adversaries, but still relies on a limited set of attributes. Hence, group description techniques EHR de-identification methods are applied to eliminate sensitive properties. For instance, categories of values may be treated by suppressing some specific ones or changing specific values with a less frequent or unknown description using a coarsenization approach.

Research suggests that medical databases contain unique values with very localized usage and definition, which can attract risk of residual disclosure. Therefore, the time dimension may add another e-anonymization property, namely time-invariance, which assures that an individual does not change location in the database up to a defined time period. Future generation of useful data for mining purposes then employs data synthesizing methods via joint modeling or processing e-anonymized records with multiple Central Derived Request in medical center approach and grouping based on sensitive properties. Further approaches can verify the usability of the anonymized data with respect to groupwise severity level for specific diseases and techniques of providing on different servers those records participating to the exact same used groups, assuring distribution of disjoint k-anonymized sets for data analysts. Data loss under data transformation was also examined, testing the degree of difference compared to original data subsets with sufficiently large sample size.

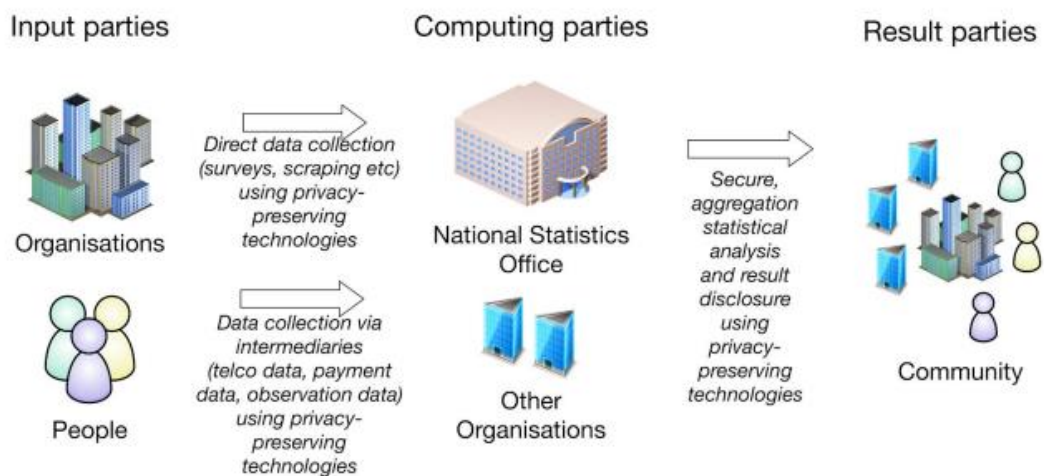


Fig 5: Privacy-Preserving Computation Techniques

5.2. Access Control and Auditability

Access control and auditability of the analytics process are crucial components in any regulatory-setting-enabled precise health decision system, particularly in precision EHR analysis systems utilized for various clinical decision-making. Healthcare organizations (HCOs) typically have stringent regulatory guidelines around auditability and access control. From an auditability perspective, as part of fulfilling HIPPA requirements, continuous logging of information will ensure clear identification of the users who accessed and/or altered sensitive information. Therefore, the framework is designed to have an improved auditability mechanism. Any alterations made to the data mini-ming by the authorized AS entities, as discussed above, are stored in the secure storage component. Whenever a read operation is performed for operational purposes, these logs are automatically attached to the output.



Equation 5: Bellman equation for Markov Decision Process / Reinforcement Learning

$$V^{\pi}(s) = \sum_a \pi(a | s) \sum_{s'} P(s' | s, a) [R(s, a, s') + \gamma V^{\pi}(s')]$$

Step-by-step derivation

Let:

- s = current state
- a = action
- s' = next state
- $\pi(a | s)$ = probability of choosing action a in state s
- $P(s' | s, a)$ = transition probability
- $R(s, a, s')$ = immediate reward
- $\gamma \in [0,1]$ = discount factor

The value of state under policy π is the expected discounted return:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t | S_t = s]$$

Define the return:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

Substitute into the value definition:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_{t+1} + \gamma G_{t+1} | S_t = s]$$

By linearity of expectation:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_{t+1} | S_t = s] + \gamma \mathbb{E}_{\pi}[G_{t+1} | S_t = s]$$

But the future return from next state s' is exactly $V^{\pi}(s')$, so:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R(s, a, s') + \gamma V^{\pi}(s') | s]$$



Now expand the expectation over actions and next states.

First over actions:

$$V^{\pi}(s) = \sum_a \pi(a | s) \mathbb{E} [R(s, a, s') + \gamma V^{\pi}(s') | s, a]$$

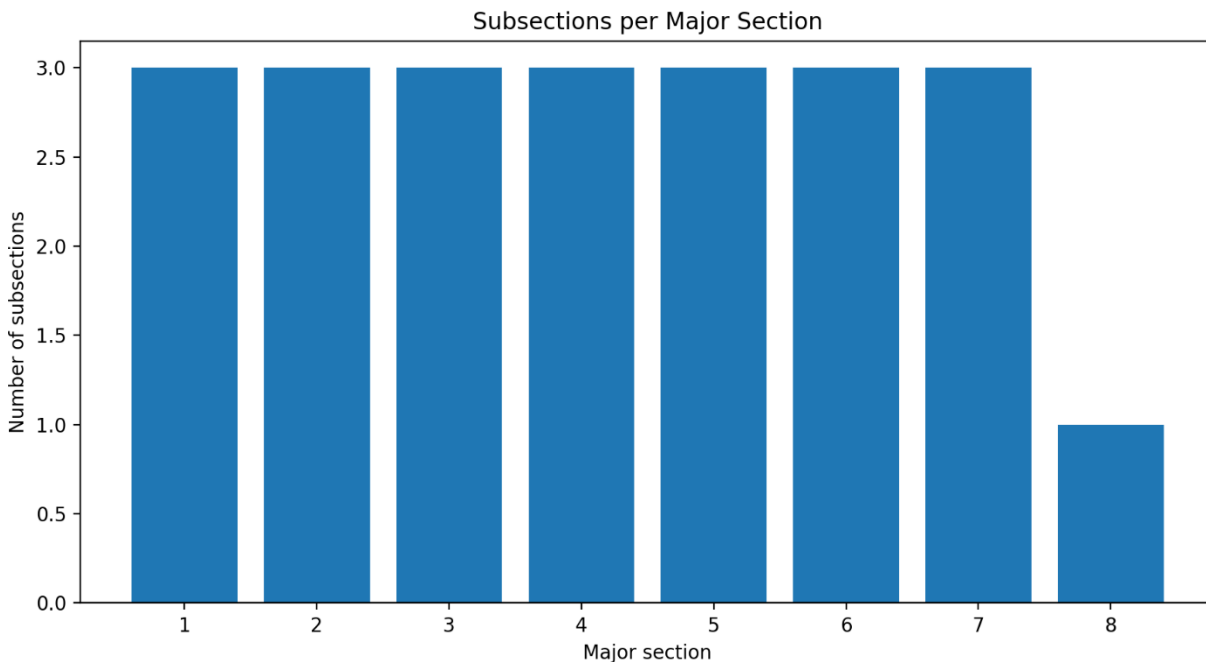
Then over next states:

$$V^{\pi}(s) = \sum_a \pi(a | s) \sum_{s'} P(s' | s, a) [R(s, a, s') + \gamma V^{\pi}(s')]$$

5.3. Regulatory Considerations and Ethics

Discussions on the ethics of AI and data privacy in health care have, to some extent, shifted from narrow concerns about over-hyped AI offering false hopes regarding trustworthiness and reliability toward broader questions concerning the responsible application of AI in scalable medical informatics. Yet pressing unresolved ethical and legal problems raised by the use of AI-nitigation in patients and analysis of clinicians remain. Many consider research on privacy-sensitive patient information especially controversial. Health information containing racial backgrounds, appointments with psychiatrists and obstetricians, or doctor's notes of very sensitive office visits for sexually transmitted diseases can help insurers assess individual health care risk more accurately. In contrast, some argue that access to such information should be restricted or even prohibited.

It has been suggested that causally revealing information is more sensitive than other information, making its use among individuals especially problematic. Studies further show that revealing specific aspects of identity leads to more nonconsensual sharing of personality profiles. In addition, an analysis of marketing messages from a leading insurance company shows that consumers' negative response to marketers or advertisers receiving sensitive information increases when parties were not involved in the interaction providing the information but rather received it through surreptitious means. In contrast, AI and data privacy in health care also promote decentralization of the health care economy and efforts to standardize the certification of drugs and procedures.



6. Validation, Evaluation, and Trustworthy AI

Validation and evaluation of any AI or machine learning (ML) solution is critical and essential for real-world deployment, impacting availability of decision-makers' confidence in harnessing the solution to make timely and accurate decisions or to achieve any underlying objective in practical applications.

Accurate and traceable historical clinical data is key to the reliability of insights and predictions produced by clinical data engines in precision healthcare decision. According to Artie E. Chen, Associate Professor of the Department of Health Information Management at National Yang Ming Chiao Tung University, the solutions using these engines should be "trustworthy or trustworthy AI". AI and ML systems are defined as trustworthy only when their quality is guaranteed and accountable, and when their objectives, activities, and implications are transparent, explainable, fair, and not harmful. A solution that does not satisfy any of the AI or ML quality assumption is questionable. Hence, validation and evaluation of trustworthiness are essential throughout the entire implementation pipeline, starting from trustworthy data.

Historical clinical data is frequently found to be inaccurate or inconsistent. Given that training data errors are bound to exist, Chen categorized these important parameters into three dimensions corresponding to data quality, labeled data accuracy, and label noise issues. Data quality refers to the requirements for the input data to be sufficiently accurate, complete, consistent, and up-to-date. It is essential to examine the quality of both the independent and target variables in a healthcare scenario. Clinical data engines with accurate label data world-wide can guarantee that their derived models are also accurate.

Hospital standard operating procedures must be defined for ML researchers to validate the label data according to reliable methods before the data is applied in training the ML model and to guide clinicians’ efficient data annotation. Labels can be correct for specific subpopulations but wrong for others. A straightforward solution is to check the label quality for different population segments using trustworthy control data.

6.1. Evaluation Metrics for Clinical AI

The quality evaluations for AI systems in healthcare can be divided by domain. For clinical efficacy, the focus is placed on identifying patient outcomes, with metrics accounting for sensitivity and specificity of patient-level predictions. The unique structure of electronic health records creates an opportunity for leveraging timeliness of predictions at the encounter level to inform prospective clinical decision-making. A set of evaluation metrics tailored to assess the practical value of clinical predictions is proposed based on directional scoring in a multi-label setting. The second key domain where the quality of AI applications is evaluated is in the user interface and AI-user interaction components. A framework correlating usability with predictive and clinical efficacy is supported by results from a deployment study. Usability assessments offer predictive power for clinical efficacy evaluations and make a compelling case for integration into multi-stage studies of AI quality in healthcare.

Evaluation of clinical AI systems can focus on clinical methods, which can be separated into task-driven methods. The most common use of clinical AI is systems that make predictions about patient outcomes, such as predicting the onset of death or disease. In this case, it is standard to evaluate the predicted labels with respect to their correspondence with true labels, using sampling-based testing procedures with sensitivity and specificity as the metrics of correct evaluation. Modeling has been extended beyond one-step-evaluated methods to trajectory prediction that aims for sequential accuracy, but in practical deployment, clinicians are typically interested in predictions to support immediate medical decisions. Systems providing a multi-label diagnosis on patients in the near future, chosen from a pool of multiple disease labels, are among the most common applications in the field. In light of this fact, the use of directional scoring as a complement to generalization testing is proposed in an encounter-level accuracy, and therefore predictive value, fulfilling a critical component of a practical evaluation framework suitable for the EHR context.

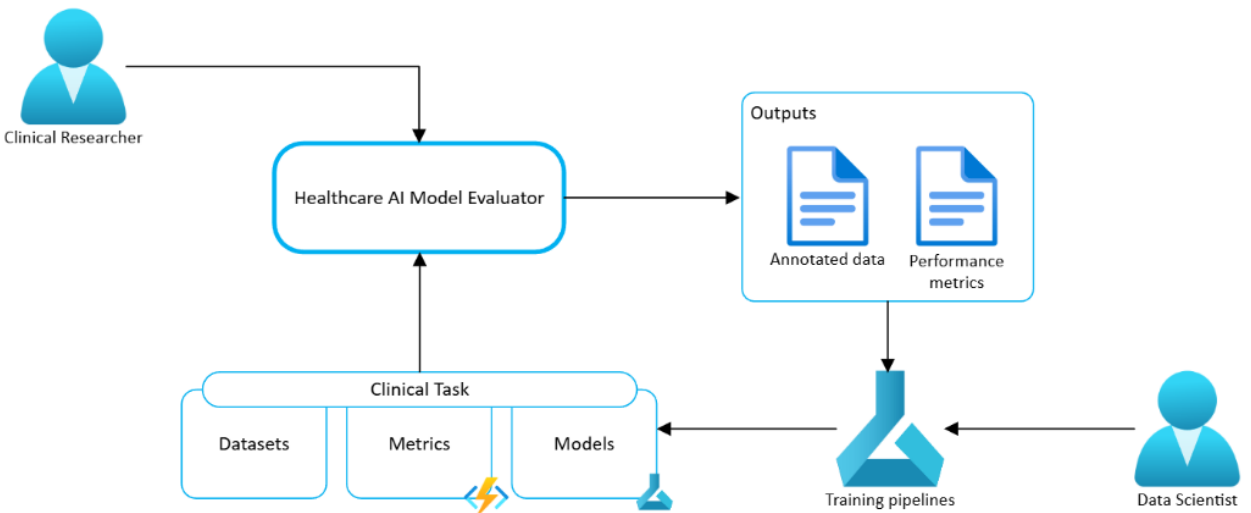




Fig 6: Healthcare AI Model Evaluator

6.2. Robustness, Generalizability, and Bias Mitigation

Most clinical decision support systems (CDSSs) are developed in specific environments (e.g. a particular healthcare institution) or domains (i.e. a specific disease), and yet their clinical performance has to satisfy specific requirements in terms of robustness and generalizability in order to be deployable in different settings and used in different populations of patients. Robustness and generalizability are commonly considered orthogonal properties. On one hand, the performance of a model can be sensitive to small perturbations of either the input data or the underlying distributions, causing large variations on unseen data. More formally, a classifier can be considered locally robust if small perturbations of the data do not lead to misclassifications for any of the samples in a given neighbourhood of the original input sample. On the other hand, models typically achieve good performance on the subset of the test data on which they were trained but fail on other data drawn from a different distribution. Generalizability refers to a model's ability to accurately predict the class labels of data outside of the training set.

Bias mitigation is closely related to generalizable and robust models. Although most CDSSs describe validated accuracy results either on in-house or publicly available datasets, the performance of predictive models is often merely a trivial statistical property, lacking any generalizable medical meaning. In fact, models trained on imbalanced datasets often perform better on the majority than on the minority class. Thus, the major concern when designing a model that will actually be deployed within a clinical setting must address fairness and, consequently, how to leverage dataset rebalancing strategies both for training and evaluation with the aim of developing practically usable and unbiased CDSSs.

Layer	Role Strength
Data Ingestion	High
Storage Architecture	High
Processing Framework	Very High
Analytics & Decision Systems	Very High

Table 3: System Architecture Layers

6.3. Explainability and Clinician Engagement

Explainability is closely tied to the notion of coherence through the meaning of, and relationships between, variables to support clinical inference. While tuning hyper-parameters may enhance explainability through visual, narrative, or other methods, too little attention to tuning may invalidate rendering explanations. History provides numerous instances where technically sound classification models have nevertheless performed poorly in practice, rendering generating accurate explanations moot. The progressive separation of deep learning model design from principal data engineers and those responsible for delivering explanation to the target audience is a necessary safeguard.

User engagement is critically vital in leveraging the full potential of an enterprise-scale EHR analytics system for clinical decision support. Notably, Propensity Basis Clinical Data Engineering leverages continuous feedback from every decision to

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 03

RECEIVED: JULY 14

REVISED: JULY 28

ACCEPTED: AUGUST 02

PUBLISHED: AUGUST 18

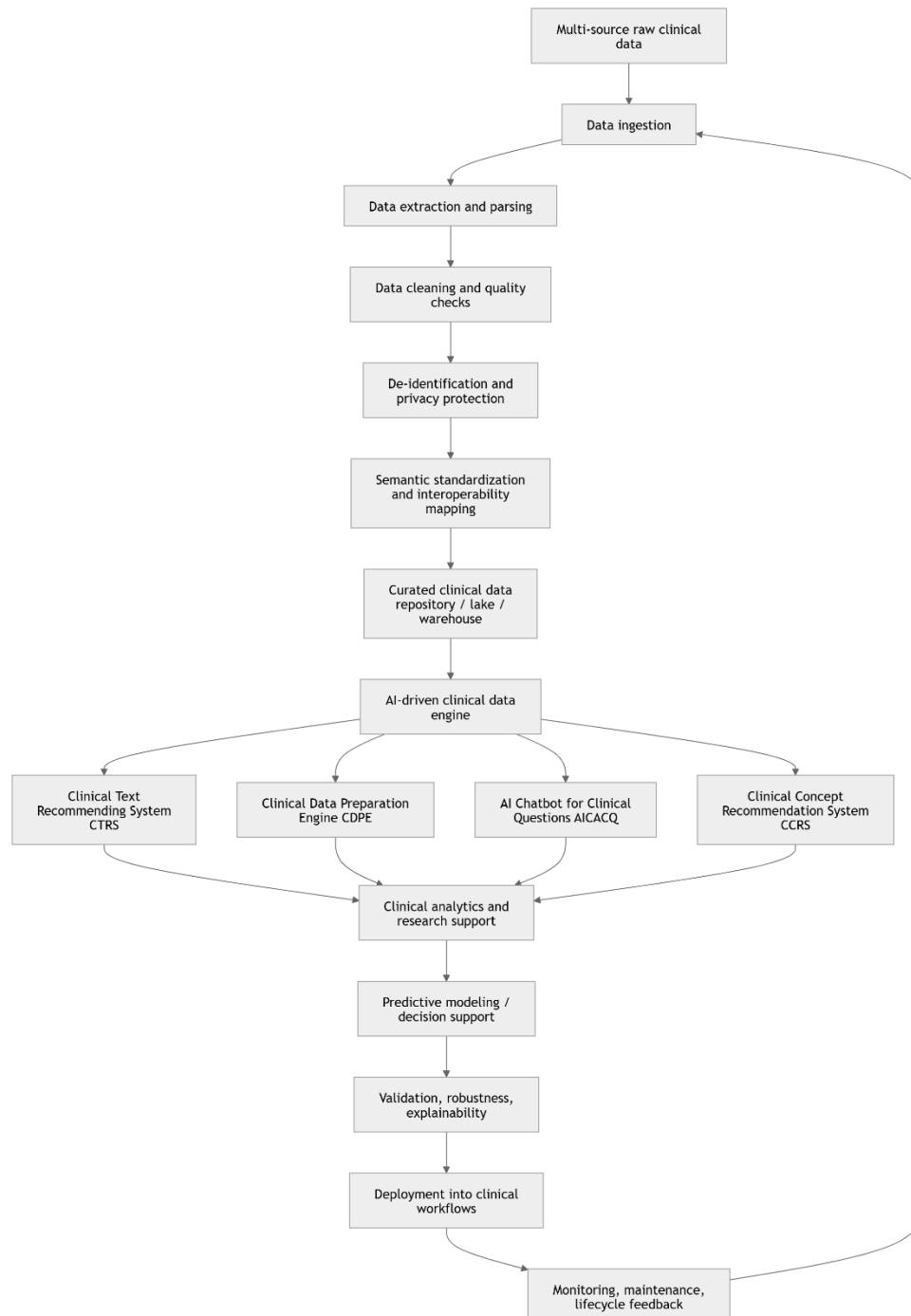


enhance clinical decision support. Probabilities from trained models can either be registered with the EHR for EHR-based decision support or informal reports employed to steer clinical recommendations and insights beyond conventional EHR reporting. Decision support must therefore remain a primary objective of clinical data engineering, for failure to do so risks poor adoption, ultimately undermining the entire enterprise-scale system.

Integrating AI-based predictive or prescriptive decision support in the workflow of busy clinicians presents significant challenges. Generating accurate predictions and clinically pertinent recommendations is difficult when such insights are not demanded, yet secondary decisions often occur with grave consequences. Much of this tension relates to the function of guidelines and protocols supplied by a clinical team's management structure. Generating AI-based predictions and recommendations in such contexts hinges on providing judgemental guidance through informal queries while respecting and working within the existing protocols.

This crucial aspect of Propensity Basis Clinical Data Engineering is underpinned by the existence of a stable foundation of accurate predictions and recommendations, allowing support queries to be answered with clinical coherence and practical significance. To date, most such queries have been exploratory and risk-driven, aiming to identify patients most at risk of a major adverse outcome over a clinically relevant time frame. However, when articulated by an experienced and astute clinician, even an informal query can receive a precise answer from the existing model cluster. The continued capture of such queries remains essential to fulfilling this step in decision support and should remain a priority.

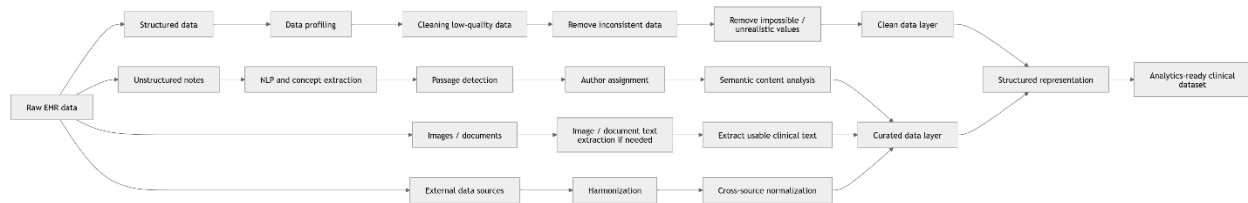
Transfer learning has opened new avenues of model development in recent years. However, across domains, training set and test set distribution shift remains a challenge. A monitoring, detection, quantification, and mitigation technique for data distribution shift in an EHR-based model is developed. A change point detection method MEWMA is utilized for detection while Kernel two-sample test is used for quantification. Deep canonical correlation analysis can simulate the samples from the lost domain in the target domain. Data shuffling based on the samples generated from deep canonical correlation analysis help to reduce the distribution shift on training and test sets.



7. Deployment and Operationalization

The development of deep learning models requires extensive training and evaluation. Training requires large workloads making development cycles again several weeks to months. In production deployment and operationalization must be handled to become beneficial for end-users. Firstly, end-users mostly interact through data visualizations. Therefore techniques like gradient weighted class activation maps, local interpretable model-agnostic explanations, and Shapley additive explanations can be utilized to convey messages to end-users. Secondly, a central aspect for operationalization is model performance monitoring. A thorough deployment of model performance monitoring keeps track of various model performance metrics. A custom framework monitors the metrics, unit tests, and explains the reasons why the unit tests failed if any set. Unit tests are regularly updated based on the features of deep learning models like distribution shifts, etc.

Flow chart:



8. Conclusion

This dissertation describes a number of innovative methods using model-based AI analytics and clinical data engineering, easily scalable from a single EHR system in academia to national-wide EHR analytics and patient decision-support systems for precision healthcare. In the proposed AI clinical data engineering, predictive analytics utilizing clinical information have been integrated into de-identification, publishing, and model-driven data evaluation methods, extending EHR analytics to health- and disease-related areas, such as:

1. Building a publicly available self-adaptive EHR de-identification tool for protecting patients' privacy without loss of clinical information
2. Developing a novel EHR anonymization technique that augments the K-anonymization by distributing the K-anonymized data on several servers
3. Enabling data publication by building zero-knowledge storage accessible to ordinary users without cryptographic background
4. Constructing an innovative EHR publication and sharing system as an extended version of the zero-knowledge storage



5. Mapping data from free-text clinical notes into a structured table for clinical text mining, by implementing a word representation method adapted for EHR analysis
6. Predicting missing features of EHR, including ICD codes and prescription drugs, using an ensemble learning approach
7. Translating diabetes EHR into a graphical query generation system for precise health-related message searching
8. Conducting a comparative study on the nuances of multiple EHR embeddings based on clinical information knowledge
9. Developing an EHR-based model-driven or intelligent agent tool for patients' health or disease (diabetes) management
10. Conducting EHR scientific research applicable to any health-related aspect of society

8.1. Emerging Trends

Five trends indicate ongoing transformation of the information systems underpinning healthcare decision support: cloud-based EHR systems are now emerging; clinical metadata management systems are being implemented; multimodal clinical databases are being deployed; EHR & other clinical analytics solutions are being supported by diverse clinical AI methods; and precision healthcare is rising with the COVID-19 pandemic. The ongoing transformation of the IA platform for large-scale clinical analytics is described and EHR-driven precision healthcare decision systems are outlined by examining the clinical AI methods used in EHR-driven decision support. This trends analysis highlights a new wave in the digital transformation of community and hospital healthcare.

Six years of R&D in AI-driven clinical data engineering have yielded emergent changes, also evident in emerging cloud-based EHR System, clinical metadata management, multimodal clinical databases, EHR & other clinical analytics supported by diverse clinical AI methods, and precision healthcare boosted by the COVID-19 pandemic. The ongoing transformation of the IA platform for large-scale clinical analytics is described and EHR-driven precision healthcare decision systems are outlined by examining the clinical AI methods used in EHR-driven decision support. This trends analysis highlights a new wave in the digital transformation of community and hospital healthcare.

9. References

1. Ardic, N., & Dinc, R. (2025). Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review. *Health Informatics Journal*, 31(3).
2. Caruana, A., et al. (2023). Machine learning for administrative health records: A systematic review of techniques and applications. *Artificial Intelligence in Medicine*, 144, Article 102642.
3. Sharma, C. K., Gorle, R., Mangala, N., Jindal, A., J. A., & Sundram, S. (2026). Blockchain-Based Human Resource Management System for Transparent Workforce Governance. In 2026 7th International Conference On Computational Vision and Bio Inspired Computing (ICCVBIC) (pp. 1741–1745). IEEE. 2026 7th International Conference On Computational Vision and Bio Inspired Computing (ICCVBIC). <https://doi.org/10.1109/iccvbic71195.2026.11689175>



4. Caterson, J., Lewin, A., & Williamson, E. (2024). The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review. *DIGITAL HEALTH*, 10.
5. Chaleplioglou, A., & Tsolakidis, T. (2025). Disruptive computational technologies in electronic medical records management. *Journal of Integrated Information Management*, 10(1), 64–75.
6. De Mello, B. H., Rigo, S. J., da Costa, C. A., et al. (2022). Semantic interoperability in health records standards: A systematic literature review. *Health and Technology*, 12(2), 255–272.
7. Movva, V. R., & Pareek, P. K. (2026). Multi-Agent AI Architecture for Autonomous Retail Operations: Coordinating Intelligent Agents Across Cloud, DevOps, and Enterprise Platform. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(23s), 925–932.
8. Guo, M., Wang, Y., Yang, Q., et al. (2023). Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interactive Journal of Medical Research*, 12, Article e44310.
9. Hama, T., Alsaleh, M. M., Allery, F., Choi, J. W., Tomlinson, C., Wu, H., Lai, A., Pontikos, N., & Thygesen, J. H. (2025). Enhancing patient outcome prediction through deep learning with sequential diagnosis codes from structured electronic health record data: Systematic review. *Journal of Medical Internet Research*, 27, Article e57358.
10. Hovenga, E. J. S., & Atalag, K. (2024). Optimizing electronic health records to support artificial intelligence. *AIH*, 1(3), 10–25.
11. Krumholz, H. M., et al. (2021). Harnessing the power of data to transform health care. *Circulation: Cardiovascular Quality and Outcomes*.
12. Arunakumari, M., Mattaparthi, R., Kothuri, S., Prakash, V. D., & Prasad, E. S. (2026, July). Data-Driven Framework for Real-Time Carbon Accounting and Sustainability Monitoring in Off-Grid Urban Residential Microgrids. In *2026 14th International Conference on Smart Grid (icSmartGrid)* (pp. 493–500). IEEE.
13. Lee, S. A., Jain, S., Chen, A., Ono, K., Biswas, A., Rudas, Á., Fang, J., et al. (2025). Clinical decision support using pseudo-notes from multiple streams of EHR data. *npj Digital Medicine*, 8, Article 394.
14. Morjaria, L., Gandhi, B., Haider, N., Mellon, M., & Sibbald, M. (2025). Applications of generative artificial intelligence in electronic medical records: A scoping review. *Information*, 16(4), Article 284.
15. Nasarudin, N. A., Al Jasmi, F., Sinnott, R. O., et al. (2024). A review of deep learning models and online healthcare databases for electronic health records and their use for health prediction. *Artificial Intelligence Review*, 57, Article 249.
16. Vakkalagadda, T., & Pareek, P. K. (2026). Self-Learning Artificial Intelligence Platforms for Autonomous Financial Risk Management. *International Journal of Special Education*, 41(15s), 238–254.
17. Ouanes, K., & Farhah, N. (2024). Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. *Journal of Medical Systems*, 48, Article 74.
18. Pinho, X., Meijer, W., & de Graaf, A. (2024). Deriving treatment decision support from Dutch electronic health records by exploring the applicability of a precision cohort–based procedure for patients with type 2 diabetes mellitus: Precision cohort study. *Online Journal of Public Health Informatics*, 16.
19. Penaka, B. (2026). AUTONOMOUS MULTI-AGENT AI FRAMEWORK FOR END-TO-END CLINICAL DECISION SUPPORT AND CARE COORDINATION. *Genetics and Molecular Research*.
20. Ramakrishnaiah, Y., Macesic, N., Webb, G. I., Peleg, A. Y., & Tyagi, S. (2025). EHR-ML: A data-driven framework for designing machine learning applications with electronic health records. *International Journal of Medical Informatics*, 196, Article 105816.
21. Saban, M., Alon, Y., Luxenburg, O., Singer, C., Hierath, M., Karoussou Schreiner, A., Brkljačić, B., et al. (2025). Comparison of CT referral justification using clinical decision support and large language models in a large European cohort. *European Radiology*, 35, 6150–6159.



22. Stöger, K., Schneeberger, D., Kieseberg, P., & Holzinger, A. (2021). Legal aspects of data cleansing in medical AI. *Computer Law & Security Review*, 42, Article 105587.
23. Lakhanpal, S., Tunjungsari, H. K., Raj, S., Chukka, A., Dawadi, D., & Mangalampalli, B. M. (2026, April). The Digital Transformation of Healthcare: Balancing Opportunities, Risks, and Ethical Considerations. In *2026 International Conference on Multidisciplinary Innovations For Smart & Sustainable Future (MISSF)* (pp. 1-6). IEEE.
24. Steinkamp, J., Kantrowitz, J. J., & Airan-Javia, S. (2022). Prevalence and sources of duplicate information in the electronic medical record. *JAMA Network Open*, 5(9), Article e2233348.
25. Sulieman, L., McCoy, A. B., Sama, L., & Peterson, J. F. (2024). The use of precision medicine to support the precision of clinical decisions in care delivery. *Yearbook of Medical Informatics*, 33(1), 168–174.
26. Wang, E. C., & Wright, A. (2020). Characterizing outpatient problem list completeness and duplications in the electronic health record. *Journal of the American Medical Informatics Association*, 27(8), 1190–1197.
27. Zhang, J., Symons, J., Agapow, P., et al. (2022). Best practices in the real-world data life cycle. *PLOS Digital Health*, 1(1), Article e0000003.
28. Potturi, P. (2026). Agentic Business Intelligence Architecture for Autonomous Data Exploration, Insight Generation, and Strategic Decision Orchestration. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(22s), 536-542.
29. Zhang, J., Morley, J., Gallifant, J., et al. (2023). Mapping and evaluating national data flows: Transparency, privacy, and guiding infrastructural transformation. *The Lancet Digital Health*, 5(10), e737–e748.
30. Cresswell, K., Domínguez Hernández, A., Williams, R., & Sheikh, A. (2022). Key challenges and opportunities for cloud technology in health care: Semistructured interview study. *JMIR Human Factors*, 9(1), Article e31246.
31. Hama, T., et al. (2025). Enhancing patient outcome prediction through deep learning with sequential diagnosis codes from structured electronic health record data: Systematic review. *Journal of Medical Internet Research*, 27, Article e57358.
32. Bandi, V. D. V. K. Autonomous Data Platforms: Converging AI, MLOps, and Cloud Engineering for Digital.
33. Yuan, H. (2024). Human-in-the-loop machine learning for healthcare: Current progress and future opportunities in electronic health records. *Medicine Advances*.
34. Hovenga, E. J. S., & Atalag, K. (2024). Optimizing electronic health records to support artificial intelligence. *AIH*, 1(3), 10–25.
35. Ouanes, K., & Farhah, N. (2024). Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. *Journal of Medical Systems*, 48, Article 74.
36. Sukumar, S., Kumar, S., Yandamuri, U. S., & KV, M. N. (2026). *Business Ethics and Corporate Sustainability*. BR Publications.
37. Nasarudin, N. A., Al Jasmi, F., Sinnott, R. O., et al. (2024). A review of deep learning models and online healthcare databases for electronic health records and their use for health prediction. *Artificial Intelligence Review*, 57, Article 249.
38. Lee, S. A., et al. (2025). Clinical decision support using pseudo-notes from multiple streams of EHR data. *npj Digital Medicine*, 8, Article 394.
39. Mahadevan, S. (2026). Governed Serverless Automation Ecosystem for AI-Driven Enterprise Integration and Sustainable Cloud Operations. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(16s), 1561-1570.
40. Morjaria, L., Gandhi, B., Haider, N., Mellon, M., & Sibbald, M. (2025). Applications of generative artificial intelligence in electronic medical records: A scoping review. *Information*, 16(4), Article 284.