American Journal of Life Sciences
2026, Vol. 14, No. 2, pp. 49–59
https://doi.org/10.11648/j.ajls.20261402.13

SciencePG
Science Publishing Group

Research Article

AI-driven Natural Language Processing and Stochastic Modeling for Transforming Unstructured Electronic Health Records into Structured Clinical Data

Philip de Melo*, Marie St. Rose, Mildred Jackson, London Mohammed

Department of Nursing and Allied Health, Norfolk State University, Norfolk, United States

Abstract

Electronic health records (EHRs) contain large amounts of valuable clinical information, but a substantial portion of this information exists in unstructured form, including physician notes, discharge summaries, and narrative clinical reports. Because these data are recorded as free text, they are difficult to aggregate, standardize, and analyze using conventional statistical or database methods. As a result, a significant amount of clinically relevant information remains underutilized in healthcare analytics and decision support systems. This study proposes a hybrid framework that combines artificial intelligence–based natural language processing (NLP) with stochastic modeling to transform unstructured EHR narratives into structured clinical datasets. The approach first applies AI-driven NLP techniques to identify and extract clinically meaningful entities such as diagnoses, symptoms, medications, laboratory values, and procedures from free-text clinical notes. The extracted information is then organized into relational tables suitable for large-scale analytics. To account for patient heterogeneity and uncertainty in clinical observations, stochastic analytical methods are applied to reconstruct latent health trajectories and estimate time-dependent risk patterns. The proposed framework enables the integration of narrative clinical information into structured data environments, facilitating more comprehensive analysis of patient health dynamics. By combining AI-based text extraction with stochastic modeling, the method improves the ability to analyze complex clinical datasets and supports more realistic representation of disease progression and patient variability. This approach has the potential to enhance population health analytics, clinical research, and predictive modeling using electronic health record data.

Keywords

Electronic Health Records (EHRs), Natural Language Processing (NLP), Unstructured Clinical Text, Clinical Entity Extraction, Stochastic Modeling

1. Introduction

Electronic health records (EHRs) have become a central component of modern healthcare systems and represent one of the most important sources of data for clinical research, population health analytics, and decision support. EHR systems capture a wide variety of patient information, including demographic data, laboratory results, diagnoses, medications, and clinical procedures. However, a substantial portion of clinically meaningful information is recorded in the form

*Correspondence: Philip de Melo (pdmelo@nsu.edu)

Received: 24 March 2026; **Accepted:** 7 April 2026; **Published:** 23 April 2026

of unstructured text such as physician notes, discharge summaries, radiology reports, and nursing narratives. These narrative documents often contain detailed descriptions of patient conditions, symptoms, treatment responses, and clinical reasoning that are not fully captured in structured database fields.

Despite their richness, unstructured clinical narratives are difficult to aggregate, standardize, and analyze using traditional data management and statistical approaches. Studies have estimated that a large fraction of information stored in EHR systems remains underutilized because it exists in free-text form [1]. The variability of medical language, the use of abbreviations, incomplete sentences, and context-dependent expressions make automated interpretation of clinical notes particularly challenging. As a result, much of the valuable information embedded in EHR narratives cannot be directly incorporated into large-scale analytics, predictive modeling, or clinical decision support systems.

Recent advances in artificial intelligence and natural language processing (NLP) provide promising approaches for extracting structured information from unstructured clinical text. AI-based NLP techniques can identify clinical entities such as diseases, medications, laboratory results, and procedures within narrative documents and convert them into structured variables suitable for computational analysis. These methods have significantly improved the ability to transform narrative clinical information into structured datasets that can be stored in relational databases and used for epidemiological studies, machine learning models, and healthcare analytics.

However, simply converting clinical narratives into structured variables does not fully capture the complexity and variability of patient health dynamics. Patients with similar observed characteristics may follow very different clinical trajectories due to biological variability, treatment responses, and environmental factors. Deterministic analytical methods often assume that patients with similar observed data share the same risk profile, which may not adequately represent the heterogeneity present in real-world clinical populations. Stochastic analytical frameworks offer an alternative approach by modeling the probabilistic evolution of patient health states and incorporating uncertainty into the analysis of disease progression and risk [2].

This study proposes a hybrid framework that integrates AI-based NLP methods with stochastic modeling to transform unstructured EHR narratives into structured clinical data and to analyze the resulting information using probabilistic approaches. The proposed methodology first applies AI-driven NLP techniques to extract clinically relevant entities from narrative clinical documents and organize them into structured relational tables. Stochastic analytical methods are then applied to model the dynamic behavior of the underlying health state and to capture variability among patients. By combining automated information extraction with probabilistic modeling, the framework enables more comprehensive utilization of EHR data and supports more realistic analysis of disease progression, population health dynamics, and clinical risk prediction.

The framework combining AI-based NLP and stochastic methods, such as Stochastic Artificial Intelligence for Hazard Analytics (SAIHA) for structuring electronic health records has many applications in healthcare analytics, clinical research, and health system management [3]. Once unstructured clinical narratives are transformed into structured datasets and latent health states are reconstructed, the information can support a wide range of analytical and clinical tasks.

One important application is clinical outcome prediction. Structured data extracted from clinical notes can be used to model patient trajectories and estimate the probability of outcomes such as mortality, complications, or disease progression. Stochastic methods allow these predictions to account for patient variability rather than assuming a single deterministic trajectory, which leads to more clinically realistic risk estimation.

Another major application is early detection of disease deterioration and clinical risk stratification. By analyzing time-dependent information extracted from EHR narratives, the framework can detect subtle patterns indicating worsening patient conditions. For example, changes in symptoms, laboratory values, and treatment responses recorded in clinical notes can signal the early stages of sepsis, cancer progression, or cardiovascular complications. Identifying these patterns allows clinicians to intervene earlier and improve patient outcomes.

The AI-based NLP and stochastic methods also support population health analytics and epidemiological surveillance. Unstructured clinical notes often contain valuable information about symptoms, exposures, and clinical observations that are not present in structured fields. Extracting this information enables large-scale monitoring of disease prevalence, outbreak detection, and tracking of public health trends across healthcare systems [4].

A further application is clinical research and real-world evidence generation. Researchers can use structured datasets derived from clinical narratives to study treatment effectiveness, patient responses to therapies, and long-term disease trajectories (???). Because clinical notes often contain detailed contextual information, incorporating these data improves the quality of observational studies and comparative effectiveness research.

The framework can also enhance clinical decision support systems. Structured information extracted from narrative notes can be integrated into predictive models that assist clinicians in diagnosis, treatment planning, and monitoring of patient risk. Stochastic modeling further improves these systems by representing uncertainty and variability in patient health states.

Another important application is healthcare quality improvement and hospital operations analytics. By analyzing structured information derived from EHR narratives,

healthcare organizations can evaluate treatment outcomes, identify patterns of adverse events, and monitor adherence to clinical guidelines [5]. These insights can inform policy decisions and improve healthcare delivery.

Finally, The AI-based NLP and stochastic methods enable advanced AI-driven healthcare analytics, including the development of predictive models, digital twins of patients, and learning health systems. By converting previously inaccessible narrative data into structured and analyzable form, the framework expands the usable information contained within EHR systems and supports more comprehensive understanding of patient health dynamics [6].

In summary, integrating AI-based NLP with stochastic modeling allows healthcare systems to unlock the information contained in unstructured EHR narratives and apply it to prediction, surveillance, research, decision support, and health system optimization.

The rapid digitization of healthcare systems has led to the widespread adoption of Electronic Health Records (EHRs), which contain a mixture of structured and unstructured data. While structured fields such as laboratory values and billing codes are readily analyzable, it is estimated that up to 80–85% of clinically relevant information resides in unstructured clinical narratives, including progress notes, discharge summaries, and physician observations [7]. EHR data often exhibit challenging characteristics: they are typically not organized or filtered, low in quality (rarely subjected to systematic audits), high-dimensional (containing thousands of distinct medical events), sparse (with many missing or zero values), heterogeneous (collected from diverse sources), temporal (recorded over time), incomplete, largescale, and multimodal (capturing various data types such as images, notes, and lab results) [8].

To identify existing methodologies for transforming unstructured EHR data into structured clinical representations, a comprehensive literature search was conducted across major biomedical and computational databases, including PubMed, IEEE Xplore, Scopus, and Google Scholar [9].

Early approaches to clinical text processing relied heavily on rule-based systems and domain ontologies, such as the Unified Medical Language System (UMLS), which enabled mapping of free text to standardize medical concepts [10]. However, despite their success, these approaches primarily focus on syntactic and semantic structuring of text and often lack mechanisms to capture the underlying stochastic dynamics and latent health states that evolve over time [11].

To address these limitations, a growing body of research has explored probabilistic and stochastic modeling frameworks for healthcare data analysis. Methods based on hidden Markov models, Bayesian networks, and stochastic differential equations have been proposed to model disease progression and patient trajectories under uncertainty [12]. Recent literature indicates that applying natural language processing-based information extraction to unstructured clinical notes enhances the identification and representation of social determinants of health, significantly augmenting structured EHR data and improving the ability to address patient risk factors and care outcomes [13].

Current research indicates that integrating clinical notes with structured electronic health record data in deep learning models enhances the prediction of serious bloodstream infections in pediatric patients with central venous lines, improving predictive accuracy and enabling more timely identification of rare clinical events [14]. Existing research indicates that rule-based text-mining approaches applied to unstructured electronic health record data can effectively identify the severity of Alzheimer's disease and related dementias, though variations in documentation practices may introduce biases into clinical interpretation [15].

Current research indicates that integrating clinical notes with structured electronic health record data in deep learning models enhances the prediction of serious bloodstream infections in pediatric patients with central venous lines, improving predictive accuracy and enabling more timely identification of rare clinical events [16]. Existing literature suggests that integrating free-text clinical information extraction with structured electronic health record data enhances the accuracy and effectiveness of estimating personalized, multistage treatment strategies, particularly by improving data completeness and identifying critical patient characteristics for optimal decision-making [17].

Recent literature indicates that natural language processing (NLP) plays an important role in extracting information from unstructured clinical notes. Most valuable patient data in electronic health records is stored in a free text form; this causes difficulty for them to be used during analysis. NLP supports converting information that is discovered into efficient and usable formats, allowing a deeper understanding of patient data and supports improved clinical and public health outcomes [18].

Research explains the importance of combining unstructured clinical notes and structured EHR data. When used together in machine learning models, an improvement in predicting complex clinical outcomes is displayed. This highlights that unstructured data provides important context that structured data itself may not be able to obtain, leading to accurate predictions and quality patient care [19]. It is also important to consider that challenges arise during conflicts when working closely with unstructured data. Factors such as differences in documentation styles, inconsistencies in formatting, and system limitations can affect the reliability of extracted information. These challenges demonstrate the need for additional standardized approaches and improvements in system designs when processing clinical data [20].

Recent research conducted focuses on integrating free text data with structured electronics health records information. Combining both improves accuracy by capturing details that are missed. Additionally, it strengthens data quality and supports meaningful analysis, leading to better informed clinical decisions and efficient and personalized care [21]. This study builds upon these foundations by proposing an integrated

framework that combines AI-based NLP techniques with stochastic modeling to not only structure unstructured clinical narratives but also reconstruct latent patient states and infer dynamic disease progression patterns. By bridging the gap between text extraction and probabilistic modeling, the proposed approach aims to enhance the clinical interpretability and predictive power of EHR-based analytics.

2. Methods

The methodology of this study consists of a hybrid analytical framework designed to transform unstructured electronic health record (EHR) narratives into structured clinical datasets and to analyze the resulting information using stochastic modeling. The proposed approach integrates artificial intelligence–based natural language processing (NLP) techniques with stochastic analytical methods to extract clinically relevant information from narrative clinical notes and represent patient health dynamics in a structured analytical form.

The first stage of the methodology focuses on data acquisition and preprocessing. Unstructured clinical narratives, including physician progress notes, discharge summaries, and diagnostic reports, are obtained from the EHR system. These documents are preprocessed to prepare them for automated analysis. Preprocessing includes removal of formatting artifacts, normalization of abbreviations and medical terminology, sentence segmentation, and tokenization. The preprocessing stage ensures that the clinical text can be systematically analyzed by computational methods while preserving the contextual meaning of medical expressions. Finally, before the extracted data is used for analysis, de-identification procedures are often applied to remove protected health information such as names, addresses, and medical record numbers. The resulting dataset contains the raw clinical narratives and associated metadata necessary for subsequent preprocessing and natural language processing analysis.

The second stage applies artificial intelligence–based NLP techniques to extract clinically meaningful entities from the narrative text. Named entity recognition algorithms are used to identify key medical concepts such as diagnoses, symptoms, medications, laboratory measurements, procedures, and temporal information. Context analysis is performed to determine whether the identified concepts represent active conditions, historical events, or negated findings. For example, the system distinguishes between statements such as “patient has diabetes” and “no history of diabetes.” Extracted entities are subsequently standardized using common clinical vocabularies where possible. The extracted information is then organized into structured relational tables that include patient identifiers, clinical variables, and time stamps associated with each observation.

In the third stage, the structured clinical variables derived from the NLP process are analyzed using stochastic modeling methods. Instead of assuming deterministic clinical trajectories, the methodology treats the observed variables as noisy manifestations of an underlying latent health state that evolves over time. The stochastic analytical framework models the probabilistic dynamics of the latent state and accounts for heterogeneity among patients. This approach allows the estimation of time-dependent risk patterns and hazard functions associated with disease progression and clinical outcomes.

The integration of AI-based text extraction and stochastic modeling enables the transformation of narrative clinical information into a structured analytical representation while preserving the variability inherent in real-world clinical populations. By combining automated information extraction with probabilistic modeling, the proposed methodology provides a systematic approach for utilizing unstructured EHR data in population health analytics, clinical research, and predictive modeling applications.

2.1. AI-driven NLP

After the acquisition and preprocessing of clinical text, artificial intelligence–driven natural language processing (AI-NLP) methods are applied to identify and extract clinically meaningful information from unstructured electronic health record narratives. Clinical notes contain complex language patterns, abbreviations, and context-dependent expressions that are difficult to interpret using traditional rule-based methods alone. AI-based NLP techniques improve extraction accuracy by learning linguistic and semantic patterns from large collections of annotated medical text. These methods enable automated identification of key clinical entities such as diagnoses, symptoms, medications, laboratory measurements, procedures, and temporal information embedded within narrative clinical documents.

The AI-driven NLP process typically begins with clinical entity recognition, in which machine learning or deep learning models analyze the tokenized clinical text to detect medically relevant concepts. Modern approaches often employ neural language models such as transformer architectures that can capture contextual relationships between words in a sentence. These models can recognize that different expressions may refer to the same clinical concept. For example, phrases such as “high blood sugar,” “hyperglycemia,” and “poor glucose control” may all be interpreted as indicators of abnormal glucose metabolism. AI-based models also help address common challenges in clinical text, including ambiguous terminology and variability in physician documentation.

In addition to identifying medical entities, AI-NLP systems perform context analysis to determine the clinical meaning of extracted information. This step distinguishes between active conditions, historical diagnoses, family history references, and negated statements. For example, the phrases “patient has diabetes” and “no evidence of diabetes” contain the same keyword but represent opposite clinical interpretations. Advanced NLP models incorporate contextual cues to correctly classify these statements. Temporal expressions and treatment actions may also be identified to associate clinical events with specific

time points.

The extracted entities are subsequently mapped to standardized clinical terminologies when possible, enabling consistent representation of medical concepts across records. The resulting information is organized into structured relational tables that associate patients, clinical variables, and observation times. This transformation allows narrative clinical data to be integrated into computational workflows for statistical analysis, predictive modeling, and population health research.

For example, consider the following clinical note:

"Patient is a 67-year-old male with hypertension and chronic kidney disease. Blood pressure 172/96. Started on amlodipine 5 mg daily."

An AI-driven NLP system can extract the relevant clinical information and represent it in a structured form as follows:

Table 1. *NLP extraction of an unstructured clinical note.*

Patient ID	Age	Sex	Diagnosis	Blood Pressure	Medication	Dose	Frequency
P2054	67	Male	Hypertension Chronic Kidney Disease	172/96	Amlodipine	5 mg	Daily

The idea of AI-driven NLP comes mainly from linear algebra, probability theory, statistics, and optimization. These mathematical tools allow machines to represent language numerically, detect patterns in text, and learn from large corpora of documents such as clinical notes in EHR systems.

First, textual data must be converted into a numerical representation that algorithms can process. A common approach is the vector space model, where each document is represented as a vector of word frequencies. One widely used weighting scheme is Term Frequency–Inverse Document Frequency (TF-IDF), which measures the importance of a word within a document relative to a collection of documents.

$$TF\_IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

In (1), $TF(t,d)$ represents the number of times term t appears in document d, $DF(t)$ is the number of documents containing the term, and N is the total number of documents. This transformation converts text into numerical vectors that can be analyzed numerically.

Once text is represented numerically, machine learning models can be applied to identify patterns in language. For example, in classification tasks such as detecting medical conditions in clinical notes, models may estimate the probability that a document belongs to a particular class. A common probabilistic model used in NLP is logistic regression, which predicts the probability of a class label based on input features:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (2)$$

In this equation, x represents the vector of textual features extracted from a document, w is a vector of learned weights, and b is a bias term. The model learns these parameters during training by minimizing a loss function using optimization algorithms such as gradient descent.

NLP systems can rely on word embeddings, where each word is represented as a point in a high-dimensional vector space. These embeddings capture semantic relationships between words. For example, the words "diabetes," "hyperglycemia," and "glucose" may appear close to one another in the embedding space because they frequently occur in similar contexts in medical documents. These vector representations are typically learned using neural network models trained on large text corpora.

Modern AI-driven NLP systems frequently employ transformer neural networks, which model relationships between words using attention mechanisms. In these models, matrices representing word embeddings are transformed through layers of linear operations and nonlinear activation functions. The training process involves minimizing a loss function that measures the difference between predicted and actual outputs, allowing the model to learn complex linguistic structures.

In the context of EHR analytics, these numerical techniques enable AI-NLP systems to transform narrative clinical text into structured variables. The resulting structured representations can then be used for statistical analysis, predictive modeling, or stochastic frameworks that reconstruct latent health states and estimate clinical risk dynamics.

2.2. Stochastic Artificial Intelligence Hazard Analysis

SAIHA is designed to analyze complex health data when the true underlying disease state is not directly observable. Instead of relying only on deterministic statistical models, SAIHA reconstructs a latent health trajectory from noisy observations and estimates the associated risk (hazard) of clinical outcomes. At the current stage of development, SAIHA can perform several important analytical tasks.

First, SAIHA can reconstruct latent health states from incomplete or noisy clinical observations. In many healthcare datasets, recorded variables such as laboratory results, vital signs, or extracted clinical concepts from EHR notes represent

only partial indicators of the patient's underlying physiological condition. SAIHA models these observed measurements as manifestations of an evolving hidden health state. By applying stochastic modeling techniques, the framework estimates the most probable trajectory of the latent state over time. This allows researchers to infer disease progression even when direct measurements of the biological process are unavailable.

Second, SAIHA can estimate time-dependent hazard functions based on the reconstructed latent state. Traditional survival analysis methods such as Kaplan–Meier or Cox proportional hazards models typically rely on observed covariates and assume similar risk for patients with the same measured characteristics. SAIHA instead links the hazard rate to the latent health trajectory, enabling more dynamic estimation of risk. This approach allows the risk of events such as mortality, disease progression, or complications to change over time as the underlying health state evolves.

Third, the framework can capture heterogeneity among patients. Real-world clinical populations exhibit substantial variability in disease progression and treatment response. Deterministic models often represent an "average patient," which may not reflect individual trajectories. SAIHA incorporates stochastic variability, allowing multiple possible trajectories to emerge from similar observed conditions. This provides a more realistic representation of patient populations and supports probabilistic prediction of outcomes.

Fourth, SAIHA can integrate heterogeneous data sources, including structured clinical variables and information extracted from unstructured EHR narratives through AI-driven NLP methods. For example, diagnoses, laboratory values, symptoms, and medications extracted from clinical notes can be incorporated as observations that inform the latent health state. This capability allows the framework to utilize information that would otherwise remain embedded within narrative clinical documentation.

Finally, SAIHA can support predictive analytics and early detection of adverse clinical events. By continuously updating the estimated latent health trajectory as new observations become available, the system can detect emerging patterns that indicate increased risk. This capability may be particularly useful for monitoring disease progression, identifying patients at high risk of deterioration, and supporting clinical decision-making.

In summary, SAIHA currently provides a stochastic analytical framework capable of reconstructing hidden health dynamics from observed clinical data, estimating time-dependent hazard functions, accounting for patient heterogeneity, integrating heterogeneous EHR data sources, and supporting predictive modeling of disease progression and clinical outcomes.

SAIHA is based on stochastic processes, differential equations, and survival analysis. The key idea is that the true health condition of a patient is not directly observable. Instead, the patient's health is represented by a latent state that evolves over time, while the measurements recorded in the electronic health record are noisy observations of this hidden process.

Latent health state dynamics

SAIHA assumes that the underlying health state $x(t)$ evolves according to a stochastic differential equation (SDE) that combines deterministic disease dynamics and random variability among patients.

$$f[x(t),t]dt + g[x(t),t]dW_t \quad (3)$$

Here $x(t)$ is the latent health state at time t, $f[x(t),t]$ is the deterministic drift term describing systematic disease progression or treatment effects, $g[x(t),t]$ represents the stochastic diffusion term capturing biological variability, and W_t denotes a Wiener process representing random fluctuations.

This formulation allows the model to represent heterogeneous trajectories of disease evolution rather than a single deterministic path. The clinical variables observed in EHR data are treated as noisy measurements of the latent state.

$$y(t) = h[x(t)] + \epsilon(t)$$

Where $y(t)$ represents observed measurements such as laboratory values, symptoms, or extracted clinical entities, $h(\cdot)$ is a mapping between the latent health state and the observed variable, and $\epsilon(t)$ is measurement noise. This observation model connects the hidden health dynamics to real-world clinical data. The probability distribution of the latent state evolves over time according to the Fokker–Planck equation, which describes how uncertainty in the system propagates.

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}[f(x,t)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[g^2(x,t)p(x,t)] \quad (4)$$

Here $p(x,t)$ represents the probability density of the latent health state at time t. This equation captures both deterministic trends and stochastic dispersion in patient trajectories.

SAIHA links the reconstructed latent state to the hazard rate describing the instantaneous risk of a clinical event such as disease progression or death.

$$h(t) = h_0 e^{\beta x(t)} \quad (5)$$

Here $h(t)$ is the hazard function, h_0 is the baseline hazard, $x(t)$ is the latent health state, and β quantifies how changes in the latent state influence risk. This formulation allows the risk of adverse outcomes to evolve dynamically as the underlying health condition changes.

Numerically, SAIHA integrates several components:

1) stochastic differential equations describing hidden health dynamics
2) observation models connecting latent states to clinical measurements
3) the Fokker–Planck equation governing probability evolution
4) hazard functions linking latent states to clinical risk

Together, these mathematical elements allow SAIHA to reconstruct latent health trajectories from noisy EHR data and estimate time-dependent risk patterns, providing a probabilistic framework for analyzing disease progression and patient heterogeneity.

Table 2. An unstructured clinical note.

Patient ID: P1007
Encounter date: March 10, 2026

The patient is a 64-year-old male with a history of type 2 diabetes mellitus, hypertension, and chronic kidney disease stage 3. He presents today with increasing fatigue, shortness of breath on exertion, and bilateral ankle swelling for the past 2 weeks. He denies chest pain. Blood pressure is 172/98 mmHg, heart rate 96 bpm, temperature 98.4 F, and oxygen saturation 94% on room air. Laboratory results show HbA1c 9.1%, serum creatinine 2.1 mg/dL, BNP 640 pg/mL, and haemoglobin 10.2 g/dL. Chest x-ray suggests mild pulmonary vascular congestion. Assessment includes poorly controlled diabetes, uncontrolled hypertension, worsening renal function, and possible early congestive heart failure. The patient was started on furosemide 20 mg daily and lisinopril 10 mg daily and advised continuing metformin 500 mg twice daily.

Norfolk EHR Processing Package consists of the following steps:
Step 1: AI/NLP application transforms the clinical note to the following Table (5 first rows):

Table 3. AI-driven NLP extraction of categorical and numerical values.

Age	64
Sex	Male
Diagnosis	Diabetes mellitus
Diagnosis	Hypertension
Diagnosis	Chronic kidney disease stage 3

Step 2: Splitting the initial table into specialized tables.

Table 4. Patient table extracted from the clinical note.

patient_id	age	sex
P1007	64	Male

Table 5. Diagnoses table extracted from the clinical note.

patient_id	encounter_date	diagnosis
P1007	2026-03-10	Type 2 diabetes mellitus
P1007	2026-03-10	Hypertension
P1007	2026-03-10	Chronic kidney disease stage 3
P1007	2026-03-10	Possible early congestive heart failure

Table 6. Symptoms table extracted from the clinical note.

patient_id	encounter_date	symptom	status
P1007	2026-03-10	Fatigue	Present
P1007	2026-03-10	Shortness of breath on exertion	Present
P1007	2026-03-10	Bilateral ankle swelling	Present
P1007	2026-03-10	Chest pain	Negated

Table 7. Vitals table extracted from the clinical note.

patient_id	encounter_date	systolic_ bp	diastolic_bp	heart_rate	temp_f	spo2
P1007	2026-03-10	172	98	96	98.4	94

Table 8. Laboratory table extracted from the clinical note.

patient_id	encounter_date	lab_name	value	unit
P1007	2026-03-10	HbA1c	9.1	%
P1007	2026-03-10	Creatinine	2.1	mg/dL
P1007	2026-03-10	BNP	640	pg/mL
P1007	2026-03-10	Haemoglobin	10.2	g/dL

Table 9. Medications table extracted from the clinical note.

patient_id	encounter_date	medication	dose	frequency	action
P1007	2026-03-10	Furosemide	20 mg	daily	started
P1007	2026-03-10	Lisinopril	10 mg	daily	started
P1007	2026-03-10	Metformin	500 mg	twice daily	continued

Table 10. Latent state reconstruction from the clinical note.

Date	Evidence from AI/NLP-structured EHR	Latent state $x(t)$
2025-12-01	Diabetes + HTN, mild abnormalities	0.85
2026-01-15	Worse BP, rising creatinine	1.20
2026-02-20	Dyspnoea begins, HbA1c worsens	1.65
2026-03-10	BNP 640 pg/ml	2.30

Figure 1 is the reconstructed latent health trajectory estimated using the SAIHA framework across four clinical visits. The latent health state $x(t)$ is inferred from structured clinical observations extracted from electronic health

record narratives using AI-driven NLP. The progressive increase in the latent state values (from 0.85 to 2.30) indicates worsening underlying disease burden over time. This trajectory reflects the probabilistic integration of multiple clinical signals, including symptoms, laboratory measurements, and diagnoses, and provides the basis for estimating time-dependent hazard and risk of adverse clinical outcomes.

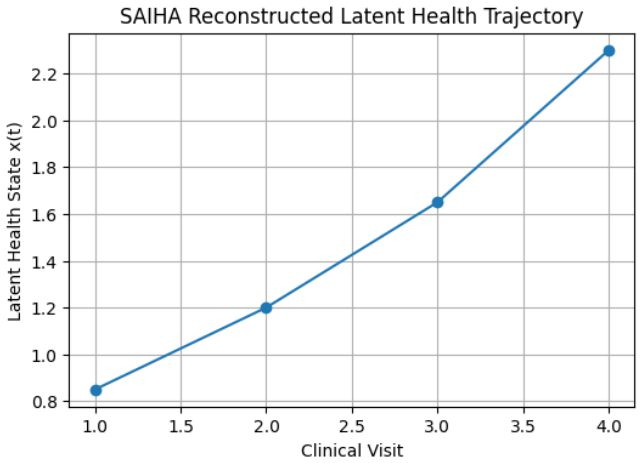

Figure 1. *Reconstructed latent health trajectory.*

The hazard model assumes that the risk depends exponentially on the latent state. The likelihood of observing the survival times and event indicators is written as

$$L(\beta) = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)$$

where

t_i is the observed follow-up time, δ_i indicates whether the event occurred (1) or the observation was censored (0), and $S(t_i)$ is the survival probability. Substituting the hazard function:

$$h(t|x) = h_0 e^{\beta x(t)}$$

the likelihood becomes a function of $L(\beta)$. The value of β is then obtained by maximizing the likelihood:

$$\hat{\beta} = \arg\max_{\beta} L(\beta)$$

This is solved numerically using optimization algorithms. Let us calculate hazards using the EHR data. Hazard is computed from the latent health state reconstructed from EHR observations. The hazard represents the instantaneous risk that an event occurs at a particular moment in time, given that the patient has survived or remained event-free up to that time. In other words, hazard answers the question: "At this moment, how likely is the event to occur for a patient who has not yet experienced it?" We use formula (5) to compute hazards.

Hazard trajectory estimated from the latent health state across four clinical visits. The hazard values $h(t)$ were computed using (5), where $x(t)$ represents the reconstructed latent health state inferred from clinical observations extracted from electronic health record narratives using AI-driven NLP. The increasing hazard values (0.043, 0.059, 0.088, and 0.158) indicate a progressive rise in instantaneous clinical risk as the latent disease burden worsens over time. Unlike traditional survival analysis approaches that estimate hazard primarily from structured time-to-event datasets, the SAIHA framework derives the hazard from a reconstructed latent physiological trajectory obtained from heterogeneous EHR signals, including symptoms, laboratory measurements, diagnoses, and treatment information. The upward trend in the hazard curve therefore reflects the increasing probability of adverse clinical outcomes implied by the evolving latent health state and demonstrates how stochastic latent-state modeling can translate complex EHR observations into dynamic estimates of patient risk.

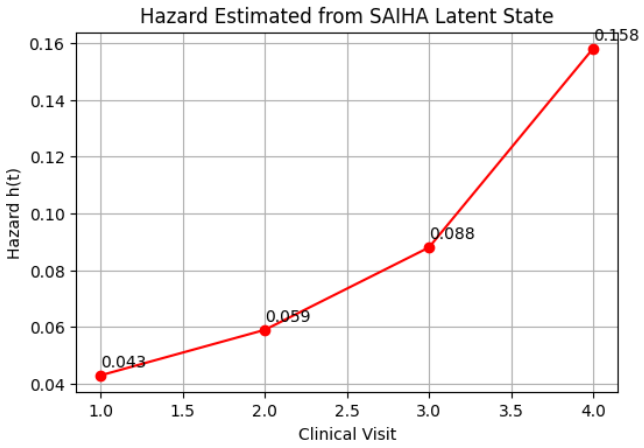

Figure 2. *Hazard is computed from the latent health state reconstructed from EHR observations.*

Figure 3 shows the continuous wavelet transform (CWT) of the smoothed hazard trajectory derived from the SAIHA latent health state. The hazard function, estimated from the reconstructed latent disease burden across clinical visits, was first smoothed using spline interpolation to obtain a continuous signal. CWT was then applied using the Morlet wavelet to analyze variations in the hazard signal across multiple scales. The horizontal axis represents clinical visits (time), while the vertical axis represents the wavelet scale corresponding to different frequency components of the signal. Color intensity indicates the magnitude of the wavelet coefficients, with brighter regions reflecting stronger local changes in the hazard dynamics. The increasing intensity toward later visits reflects the progressive rise in instantaneous clinical risk associated with the worsening latent health state inferred from EHR-derived clinical observations.

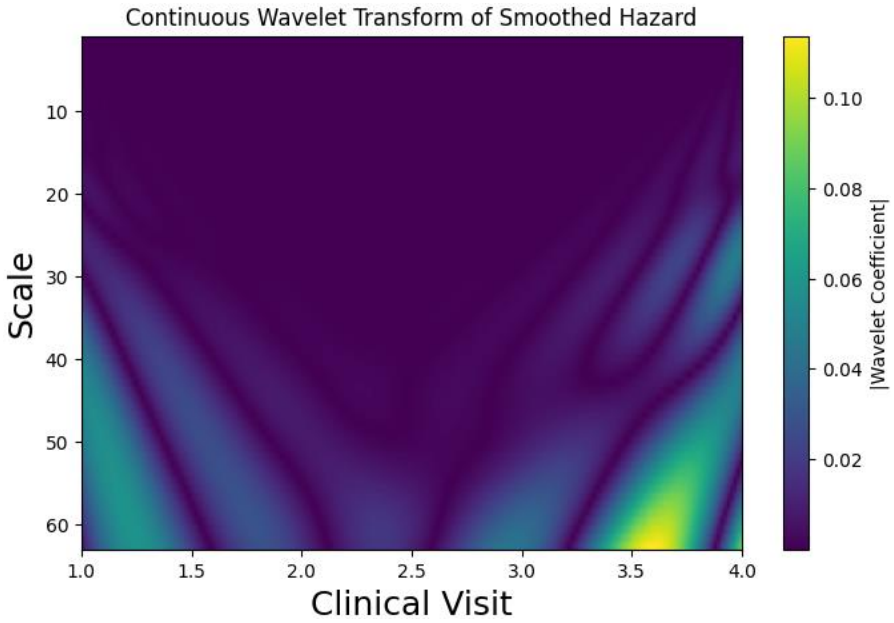

Figure 3. *The continuous wavelet transforms (CWT) of the smoothed hazard trajectory derived from the SAIHA latent health state. Between clinical visits 3 and 4, the wavelet scalogram shows a marked increase in the magnitude of the wavelet coefficients, particularly at larger scales. This indicates a strong local change in the hazard trajectory, which corresponds to a rapid increase in the estimated instantaneous risk derived from the latent state.*

3. Conclusion

Electronic health records contain a large volume of clinically meaningful information, much of which is recorded in unstructured narrative form. Because free-text clinical documentation cannot be directly analyzed using conventional database methods, a substantial portion of valuable clinical knowledge remains underutilized in healthcare analytics. This study presented a hybrid framework that integrates artificial intelligence–driven natural language processing with stochastic analytical modeling to transform narrative clinical records into structured datasets suitable for advanced analysis.

AI-based NLP techniques were used to extract clinically relevant entities such as diagnoses, symptoms, laboratory values, and medications from unstructured clinical notes. These extracted elements were organized into structured relational tables that represent observable clinical variables associated with individual patients and encounters. By converting narrative documentation into structured data, the proposed approach enables systematic integration of clinical narratives into computational workflows used in health analytics and clinical research.

The structured observations were subsequently analyzed using the Stochastic Artificial Intelligence Hazard Analysis framework, which models clinical measurements as noisy manifestations of an underlying latent health state. Unlike traditional deterministic survival models that assume homogeneous patient populations, SAIHA reconstructs stochastic trajectories of disease progression and links these trajectories to time-dependent hazard functions. This probabilistic representation captures the heterogeneity and uncertainty inherent in real-world clinical populations and allows more realistic modeling of disease dynamics.

The integration of AI-driven NLP and stochastic modeling therefore provides a unified analytical pipeline that connects narrative clinical documentation with probabilistic representations of patient health dynamics. By enabling the extraction of information from unstructured EHR narratives and linking these observations to latent disease trajectories, the proposed framework expands the analytical value of electronic health records. This approach has the potential to improve population health analytics, support predictive modeling of disease progression, and enhance data-driven clinical decision support systems.

Abbreviations

AI	Artificial Intelligence
HER	Electronic Health Records
NLP	Natural Language Processing
SAIHA	Stochastic Artificial Intelligence Hazard Analysis

Author Contributions

Philip de Melo: Conceptualization, Data curation, Methodology

Marie St. Rose: Project administration

Mildred Jackson: Resources

London Mohammed: Validation

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Chen, S., Mao, X., Huang, X., Jiang, Y. Integrating unstructured and structured EHR data for readmission prediction in patients with colorectal cancer. *Studies in Health Technology and Informatics*. 2025, 329, 1828–1829. https://doi.org/10.3233/SHTI251235

[2] De Melo, P. Public health informatics and technology. Library of Congress, Washington, DC. 2024.

[3] De Melo, P., St. Rose, M. Accurate classification of diabetes via PM generative AI. *Advances in Bioscience and Biotechnology*. 2025, 16, 379–409. https://doi.org/10.4236/abb.2025.169025

[4] Hatef, E., Kitchen, C., Gray, G. M., Zirikly, A., Richards, T., Ahumada, L. M., Weiner, J. P. Enhancement of a social risk score in the electronic health record to identify social needs among medically underserved patients: using structured data and free-text provider notes. *JAMIA Open*. 2024, 7(4), ooae117. https://doi.org/10.1093/jamiaopen/ooae117

[5] Hatef, E., Rouhizadeh, M., Tia, I., Lasser, E., Hill-Briggs, F., Marsteller, J., Kharrazi, H. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Medical Informatics*. 2019, 7(3), e13802. https://doi.org/10.2196/13802

[6] Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M. Z., Paul, R., Smriti, K., Naik, N., Somani, B. K. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science*. 2022, 191(4), 1473–1483. https://doi.org/10.1007/s11845-021-02730-z

[7] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*. 2016, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[8] De Melo, P. Prediction of diabetes from electronic health records. *International Journal of Artificial Intelligence and Applications*. 2025, 16(4), 21–37.

[9] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. 2019, 4171–4186.

[10] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., McDermott, M. Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, 72–78.

[11] Jensen, P. B., Jensen, L. J., Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*. 2012, 13(6), 395–405. https://doi.org/10.1038/nrg3208

[12] Liu, Y., Chen, P. H. C., Krause, J., Peng, L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019, 322(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489

[13] Lybarger, K., Dobbins, N. J., Long, R., Singh, A., Wedgeworth, P., Uzuner, Ö., Yetisgen, M. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*. 2023, 30(8), 1389–1397. https://doi.org/10.1093/jamia/ocad073

[14] Neubig, L., Larsen, D., Kunduk, M., Kist, A. M. Unstructured electronic health records of dysphagic patients analyzed by large language models. *IEEE Journal of Translational Engineering in Health and Medicine*. 2025, 13, 237–245. https://doi.org/10.1109/jtehm.2025.3571255

[15] Prakash, R., Dupre, M. E., Østbye, T., Xu, H. Extracting critical information from unstructured clinicians' notes data to identify dementia severity using a rule-based approach: feasibility study. *JMIR Aging*. 2024, 7. https://doi.org/10.2196/57926

[16] Tabaie, A., Orenstein, E. W., Kandaswamy, S., Kamaleswaran, R. Integrating structured and unstructured data for timely prediction of bloodstream infection among children. *Pediatric Research*. 2022, 93(4), 969–975. https://doi.org/10.1038/s41390-022-02116-6

[17] Zhou, N., Brook, R. D., Dinov, I. D., Wang, L. Optimal dynamic treatment regime estimation using information extraction from unstructured clinical text. *Biometrical Journal*. 2022, 64(4), 805–817. https://doi.org/10.1002/bimj.202000388

[18] Au Yeung J, Shek A, Searle T, Kraljevic Z, Dinu V, Ratas M, Al-Agil M, Foy A, Rafferty B, Oliynyk V, Teo JT. Natural language processing data services for healthcare providers. BMC Med Inform Decis Mak. 2024 Nov 26; 24(1): 356. https://doi.org/10.1186/s12911-024-02713-x

[19] Chiu, C. C., Wu, C. M., Chien, T. N., Kao, L. J., Li, C., Chu, C. M. Integrating structured and unstructured EHR data for predicting mortality by machine learning and latent Dirichlet allocation method. *International Journal of Environmental Research and Public Health*. 2023, 20(5), 4340. https://doi.org/10.3390/ijerph20054340

[20] Tang AS, Woldemariam SR, Miramontes S, Norgeot B, Oskotsky TT, Sirota M. Harnessing EHR data for health research. Nat Med. 2024 Jul; 30(7): 1847-1855. https://doi.org/10.1038/s41591-024-03074-8

[21] Liu Y, Zhang Z, Mi J, Pan S, Chen T, Guo Y, He X, Bian J. GatorCLR: Personalized predictions of patient outcomes on electronic health records using self-supervised contrastive graph representation. J Biomed Inform. 2025 Aug; 168: 104851. https://doi.org/10.1016/j.jbi.2025.104851