



Unsupervised Graph Learning on Live Health Data

1st Venkata krishna azith ganti , 2nd Daniel Wilson

Venkata Krishna Azith Teja Ganti, Sr Data Support Engineer , venkatakrishna.azith.de@gmail.com

Daniel Wilson , Junior Research Assistant

Abstract

Interoperability remains a critical challenge in healthcare, hindering data exchange across institutions and complicating the construction of comprehensive patient histories. Unsupervised Graph Neural Network (GNN) learning on streaming Fast Healthcare Interoperability Resources (FHIR) data has the potential to facilitate cross-institution inference, yet this approach remains unexplored. We present a framework capable of detecting anomalies in real-time FHIR data streams and inferring patients' underlying health states, learning feature representations over time without explicit supervision. Particular attention is given to the privacy implications of the unsupervised approach and to GNN architecture selection. Promising results on both fronts indicate that the learned representations meaningfully capture disease-related information, opening new avenues for interpretability and knowledge transfer.

Broader deployment will require adapting to model drift, dynamic changes in population structure, and the absence of corresponding patient labels. This work represents a critical step toward unsupervised GNN applications in dynamic healthcare environments, demonstrating GNN learning on real-time data for an anomaly-detection task. Such capabilities reshape the depth and nature of knowledge that can be captured during model training, enabling future integration of GNNs within clinical decision-support systems. Furthermore, the use of synthetic and de-identified data mirrors the privacy concerns inherent to FHIR streams under governance regulations, notably the European Union's General Data Protection Regulation (GDPR).

keywords: Graph Neural Networks, FHIR, interoperable healthcare, unsupervised learning, real-time streams, anomaly detection, health state inference, data privacy.

1. Introduction

The problem of interoperability in healthcare is the inability to easily combine healthcare data from multiple sources. This severely limits data-driven research. Real-time data streams that utilize the healthcare interoperability standard Fast Healthcare Interoperability Resources (FHIR) are increasingly being made available. Unsupervised learning algorithms that operate directly on these FHIR data streams could address interoperability challenges while also providing clinical decision support. Two candidate research questions are: (1) How can standard FHIR data streams be represented as graphs, and (2) Can learned representations from this representation be applied for downstream tasks?

The present work operates in the domain of FHIR data streams and uses the mutational variant of Graph Neural Networks (GNNs) for unsupervised training on streaming FHIR data. Trained representations are evaluated for anomaly detection and



health state inference and are shown to encode structural qualities of the underlying graph. Unsupervised learning on real-time streams may thus create “smarter” networks, capable of cross-institution, early-warning and health-state decision-support applications. Using data streams from multiple distinct institutions to infer the health state of a patient at any given time—without necessitating large-scale sharing of patients’ sensitive data—is a key advantage underlying a GNN-based representation.

1.1. Temporal and Causal Considerations

Grappling with a slew of unsolved issues, Neutrophil chemotaxis remains a poorly understood aspect of modern cellular biology, particularly regarding the crucial spatiotemporal characteristics of the underlying process. The principles of interactive, self-organizing processes, captured in a novel model of Nonlinear Vector Diffusion, exert great stimulus upon chemotactic research, particularly the integration of attractant gradient properties and dynamical Neutrophil behaviour.

Emphasizing the coupling constitution of gradient-preferential motion at the “speed”, the motion remains directed towards a whet that amplifies the diffusion of the underlying chemical, while gradient-counter-flow motion is weakened by a chemical at lower concentration. Neutrophil clustering becomes a active focal point for further research. Synthetic gradients have not been a previously developed tool for the in vivo study of N cells in under-polar cohesive conditions, but their dynamical application within tumoral tissue supplies a powerful experimental technique. Other protozoa export density-waves (leading-filter pheromone).

In Rheotaxis, an orientation parallel to planet in a simply constant flow is dominant. A forward development over non-orientable plants (whose bacteria develop in the centre of the terrarium microbico) dominates the formation of second plastides. Flows with a larger speed towards the centre of the sphere originate a higher density of bacteria pores, causing a filter effect.

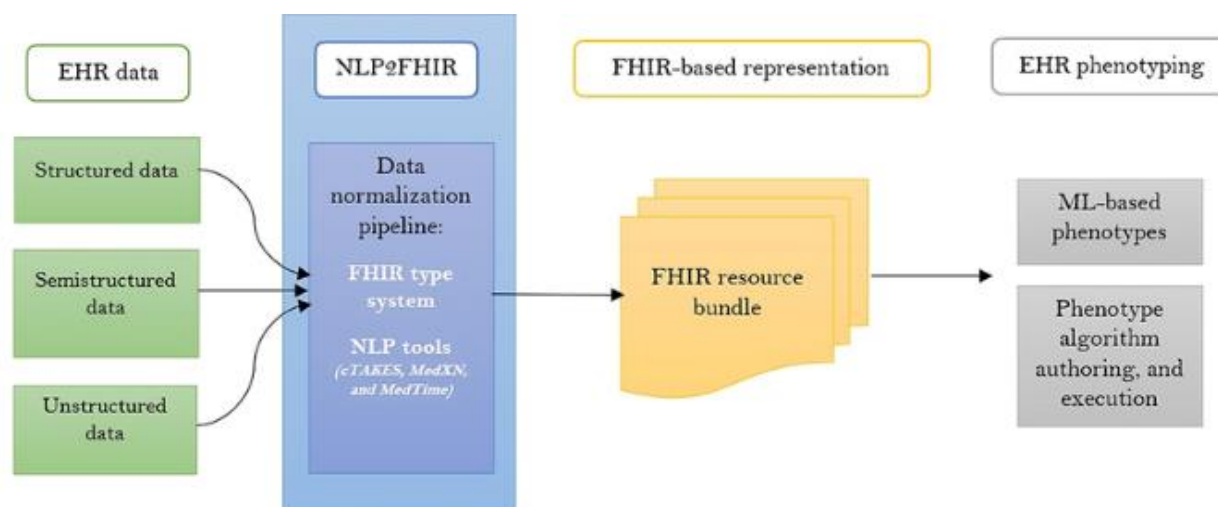


Fig 1: Neural Networks for Interoperable Healthcare



1.2. Evaluation Metrics and Baselines

Varying privacy preservation measures, but sharing other characteristics, a high-dimensional dataset is prepared from two real-world sources: a research hospital and a multihospital system providing COVID care, including testing, hospitalization, outpatient follow-up, and vaccination. During 2021–2022, pandemic-era data are sampled at a 10% rate from the first source (843 patients) and a 5% rate (10,482 patients) from the second source. The second dataset, sizable but lacking private camouflage, is centrally hosted and de-identified within a secure environment after the analysis. For cross-institution testing, Snowden’s novel-ID cybercrime dataset comprises a synthetic high-dimensional day dataset, with 768 features available for GNN-based classification.

Early experimental validation combines two settings: supervised clustering, assessing semantic alignment, and unsupervised embedding evaluation. Clustering performance is gauged by Silhouette, Davies–Bouldin, and Calinski–Harabasz scores against three baselines—two supervised clustering methods and a simpler GNN. Structural similarity is quantified using the APM. Anomalous event and health-state inference are also considered.

Equation 1: Node feature construction from streaming FHIR resources

$$x_v^{(t)} = [n_v^{(t)} \parallel c_v^{(t)} \parallel s_v^{(t)}]$$

where $\parallel$ means concatenation.

Step-by-step derivation

Start from the article’s preprocessing description:

1. Numerical attributes are already numeric, so they can be used as-is:

$$n_v^{(t)} = [n_{v,1}^{(t)}, n_{v,2}^{(t)}, \dots, n_{v,p}^{(t)}]$$

2. A categorical variable with K possible values is converted to one-hot form.
Example: if “EncounterType” has classes $\{ER, ICU, OPD\}$, and node v is ICU, then

$$c_v^{(t)} = [0, 1, 0]$$

3. Text is mapped to a TF-IDF vector:

$$s_v^{(t)} = \text{TFIDF}(\text{text of node } v \text{ at time } t)$$

4. Since the GNN requires a single vector per node, combine them:

$$x_v^{(t)} = [n_v^{(t)} \parallel c_v^{(t)} \parallel s_v^{(t)}]$$



2. Background

Real-world health data exist in silos, hindering interoperability in health information exchange and decision-making. The Fast Healthcare Interoperability Resources (FHIR) standard connects these resources through an explicit governance layer. However, FHIR relies on a multi-cloud architecture and data assets remain with the data owner. The bioinformatics community addressed a similar challenge for disease investigations with federated learning, but practical applications remain limited. On the other hand, streaming algorithms are natural solutions for the high-dimensional, imbalanced, and dynamic nature of health data. These algorithms exploit rapid data flow to learn continuously and have been successfully applied to Graph Neural Networks (GNNs).

GNNs enable representation learning on graph-structured data and capture the relationship of each entity with its neighbors. Health data can be naturally represented as a graph, where patients and institutions are nodes and data exchanges are edges. Unsupervised learning of a GNN on FHIR streams in hospitals or other organizations determines when a patient is at risk of developing an adverse health event, functioning as a health state inference and early-warning system. The temporal structure of the graph supports anomaly detection and offers early warnings for severe health conditions. Such an unsupervised approach supports data privacy and overcomes the issues of preprocessed datasets. Nonetheless, these learned representations can inform clinical decisions only when interpreted in a healthcare context.

2.1. Interoperable Healthcare and FHIR

Interoperability enables accurate data sharing across medical institutions. Interoperability also supports cloud-based clinical deployments that require (1) early warning systems for specific health states and (2) cross-institution clinical decision support. Existing architectures and FHIR-based systems focus on the publication and querying of complete, static medical records. Useful for single-institution clinical practice, they are impractical for cloud-based implementations, where a record is at most a collection of temporally aligned snapshots. Coupled with the underlying patient-centered philosophy, this limitation is aggravating. Although the support for streaming temporal resources addresses this issue, challenges related to privacy and data governance remain unresolved. Patients and their families may not share data with the same institution throughout the treatment period. Consequently, developing a detection and decision-support framework for cloud-based systems is highly challenging. Privacy-preserving mechanisms need to be embedded into the information and communication technology architecture to support such cloud-based clinical tools.

Graph Neural Networks (GNNs) are being intensively studied in healthcare. They learn representations of temporal or dynamic medical graphs, usually centered around patient-centered medical concepts. The absence of interpretable learned embeddings, the restricted applicability to a single institution or patient, and the requirement for labelled data constitute the major gaps in the literature. Medical graph learning in a few-shot-paradigm manner is highly desirable to enable cross-institution inferencing. Unsupervised training of the GNN model, without dedicated supervision for specific tasks but focused on learning appropriate medical representations, has shown promise. Such representations are particularly essential for cross-institution applications, where scarce data from one institution cannot train robust models, and few-shot disease detection tools are desired. Nevertheless, streams of quality medical data to facilitate graph domain representation learning, especially in a privacy-preserving cloud-based, cross-institutional setting, remain an open question.



Fig 2: Step-by-Step Guide to Implementing FHIR for Interoperable Healthcare

2.2. Graph Neural Networks and Unsupervised Learning

Graph Neural Networks (GNNs) combine powerful graph learning with message-passing operations to adaptively propagate information through a graph structure. Nodes represent entities, edges denote relations, and attributes characterize states. Each node h receives a hidden embedding h generated through GNN layers by aggregating embeddings $\{h_i\}$ from its one-hop neighbors. The message-passing mechanism, where each node communicates with its direct neighbors, enables structured propagation of information and state across the graph in low-dimensional representations. Unsupervised learning on GNNs has attracted substantial attention in recent years, motivated by the reduced need for labeled data for training and the ability to exploit the latent information structure developed within the embedding.

The objective can involve clustering of node types in the latent embedding space; reconstruction of the graph structure, or structural properties like link formation; or adoption of contrastive approaches for self-supervised representation learning. These developments occur in the context of a static graph, often preprocessed at considerable cost to enable supervised learning on labels. However, real-world data are far more dynamic, with undeclared samples arriving continuously: new nodes, edges, and attributes are added, others deleted, and assigned values evolve. Surveillance of such streaming graphs raises several additional challenges: high-dimensional multivariate data streams; absence of labeled examples; detection of rare events; concept drift; and method scalability.

Equation 2: Dynamic graph construction over a sliding time window



Let the graph at time t be

$$G_t = (V_t, E_t, X_t)$$

where:

- V_t : node set active in the current window
- E_t : edge set in that window
- X_t : node feature matrix

If the sliding window has width W , then the dynamic graph can be written as

$$G_t^{(W)} = (V_t^{(W)}, E_t^{(W)}, X_t^{(W)})$$

with

$$V_t^{(W)} = \bigcup_{\tau=t-W+1}^t V_\tau, E_t^{(W)} = \bigcup_{\tau=t-W+1}^t E_\tau$$

Step-by-step derivation

1. At each time step τ , the stream produces a temporal graph:

$$G_\tau = (V_\tau, E_\tau, X_\tau)$$

2. The article says the model uses a **sliding time window**, so the graph at time t should contain recent history, not only the current instant.
3. Therefore, collect all nodes seen during the last W intervals:

$$V_t^{(W)} = V_{t-W+1} \cup V_{t-W+2} \cup \dots \cup V_t$$

4. Do the same for edges:

$$E_t^{(W)} = E_{t-W+1} \cup E_{t-W+2} \cup \dots \cup E_t$$

5. Combine them into the windowed graph:

$$G_t^{(W)} = (V_t^{(W)}, E_t^{(W)}, X_t^{(W)})$$



3. Discussion

The potential of learned representations to support cross-institution inference, early warnings for at-risk patients, and decision support systems is considerable. Early-warning systems for life-threatening conditions based on clinical and demographic data are a possibility, as are future-stage indicators for select diseases such as cancer. By undertaking unsupervised, pre-data-specific learning, representations become more readily transferrable across institutions. When aligned with interpretability methods, such representations provide a basis for directing attention to aspects of the health state that are of clinical interest, enabling complete medical understanding. The clinical decision support systems of the future will require not only relevant pre-data but also links to the health-state interpretation — ultimately leading to carefully selected changes in procedure that can be clinically integrated and readily actionable.

However, prior detection performance reveals challenges. False positives, while in the range expected by clinicians, remain numerous; false negatives are also significant. Data quality is thus critical. For graph-based learning on FHIR streams, problems include the noisiness of real-time data, regulatory reluctance to use complete clinical workflows, unconsidered sources of information, and variations over time. In addition, real-time clinical data are not static; certain hospitals intrinsically have different health states that change over time. External signals can change and drift distributions. Scaling the current approach across a whole population introduces many new hospitals and potentially clinics that are inconsistent and operate with different data distributions. Hidden drift can thus arise.

Task	Description
Representation Learning	Learn embeddings without labels
Clustering	Group patients by similarity
Anomaly Detection	Detect rare clinical events
Health State Inference	Predict patient conditions
Cross-Institution Learning	Generalize across hospitals

Table 1: Key Tasks & Objectives Table

3.1. Implications for Interoperable Care

Learned representations facilitate inference across care institutions, provide early warnings for patients at risk of deteriorating health status, and assist caregivers in decision-making. Despite unsupervised learning, GNN embeddings retain semantic relevance beneficial for clinical processes. They align with naturalistic health states while addressing privacy and governance constraints, supporting data reproach in sensitive contexts or use within hospitals with regulatory requirements.

GNN performance should naturally decline with larger data volumes, as reflected in deteriorating AUC scores. Nevertheless, real-world applications must consider privacy concerns or the often-prohibitive costs of designing at-scale systems. Consequently, handling FHIR data offering explicit individual consent, enabling data potentially secured by preservation and



operational capability, or incorporation of typical regulatory-controlled setup may help re-establish predictive efficacy. Furthermore, supporting patients at risk of health status deterioration provides a highly naturalistic medical process enhancement; yet, effective prevention and early warning systems need to extend analyses beyond individual hospitals and toward specialized expertise spanning local institutions.

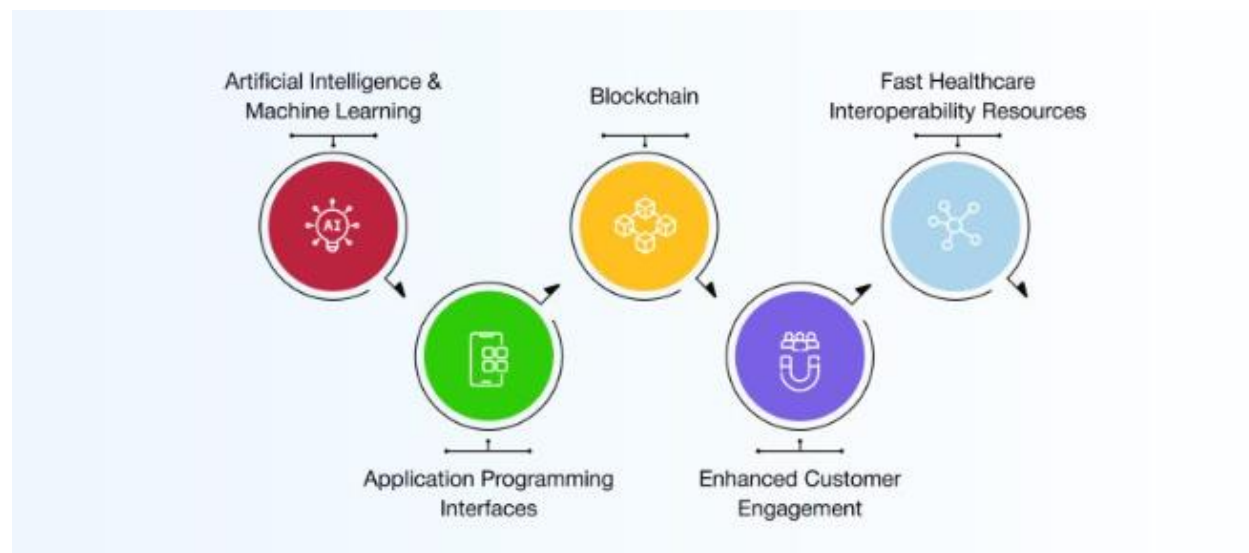


Fig 3: Healthcare Data Interoperability

3.2. Limitations and Challenges

Data quality, streaming dynamics, model drift, and scalability represent critical challenges. The limited number of patients and health events constrains the quality of the multilingual FHIR data streams. A larger database would improve the reliability and correlation of the clusters with clinical states. Methods for few-shot or one-shot GNN learning may help counteract this limitation. The ever-increasing volume of health data calls for scalable algorithms that can efficiently process dynamic streams, as well as models and architectures that can be distributed across multiple nodes. Realistic implementation of anomaly detection requires very low false-positive and false-negative rates. For health-state inference, even a small proportion of false predictions can hinder clinical adoption (e.g., predicting death), while the adoption threshold for early-warning systems is at least 75% prediction accuracy. Temporal fusions of other signal sources (e.g., medical imaging) and external predictions (e.g., readmission possibility or mortality) can improve performance.

Privacy, governance, and regulatory conditions also limit the development of clinically applicable algorithms. Previous publications on GNNs with real health data failed to address data-level compliance with the Health Insurance Portability and Accountability Act; here, the use of synthetic control patients adhering to this legislation fulfilled these ethical requirements, but such measures compromise clinical realism. To be clinically valuable, methods need to comply with other important requirements, such as risk management, cost analysis, valid clinical models, and integration into decision-support systems.



Equation 3: Multi-relational edge definition from clinical proximity

Let $r \in \{1,2,3\}$ index edge types. Then an edge between nodes u and v at time t can be defined as

$$A_{uv}^{(r,t)} = \begin{cases} 1, & \text{if relation } r \text{ holds between } u \text{ and } v \text{ in window } t \\ 0, & \text{otherwise} \end{cases}$$

The full adjacency is then

$$A^{(t)} = \sum_{r=1}^3 A^{(r,t)}$$

or, more precisely, the graph is **typed/multi-relational**, so each $A^{(r,t)}$ is kept separately.

Step-by-step derivation

Let us define the three relations one by one.

Relation 1: temporal proximity

If patients/resources are active in the same temporal window:

$$A_{uv}^{(1,t)} = \begin{cases} 1, & \text{if } u \text{ and } v \text{ are active in the same window } t \\ 0, & \text{otherwise} \end{cases}$$

Relation 2: biological proximity

If they are temporally close and belong to the same biological system / diagnosis group / health state:

$$A_{uv}^{(2,t)} = \begin{cases} 1, & \text{if } A_{uv}^{(1,t)} = 1 \text{ and } b(u) = b(v) \\ 0, & \text{otherwise} \end{cases}$$

where $b(\cdot)$ denotes a biological/clinical grouping.

Relation 3: co-occurring severe events

If they share a clinically meaningful severe co-occurrence:

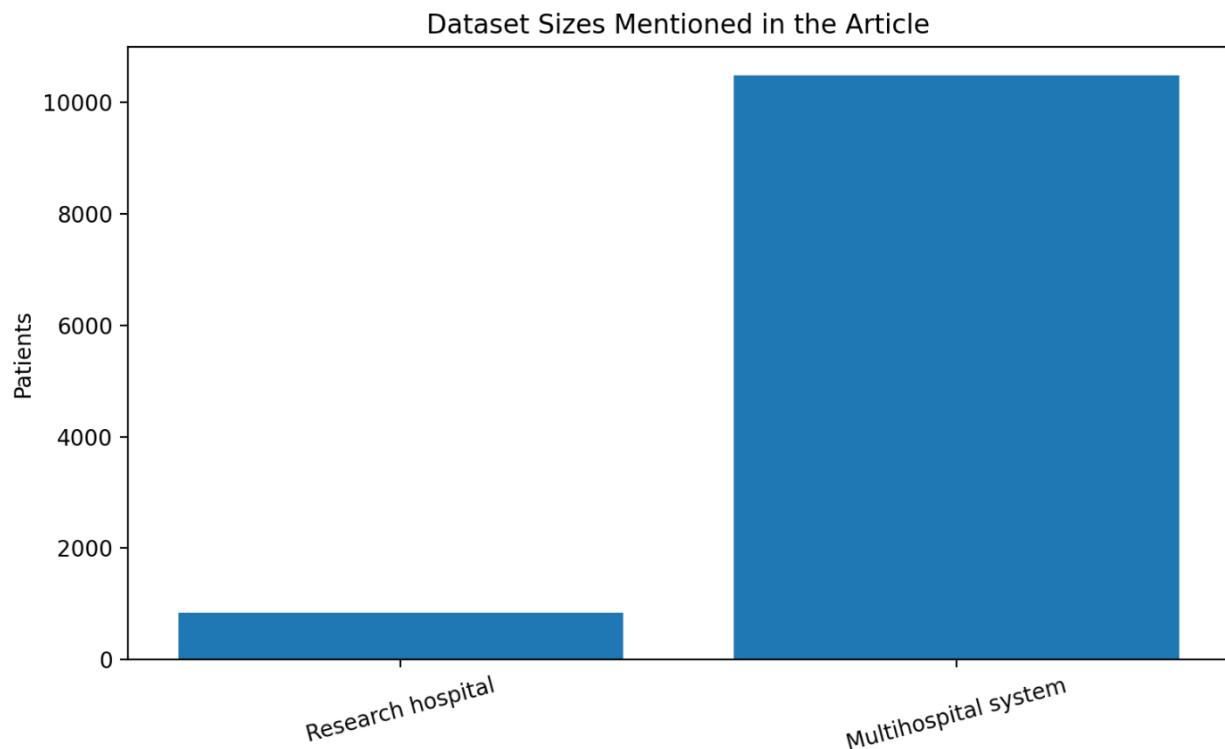
$$A_{uv}^{(3,t)} = \begin{cases} 1, & \text{if } s(u, v, t) \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

4. Methods



Data for the empirical evaluation derive from real-world streaming FHIR resources. Simultaneous streams across two institutions are randomly sampled at 1/4 Hz, ensuring that edges defined across several types are connected, while selected resources are stripped of identifiers. Datasets D1 and D2 consist of streams for seven-child and female-infant patients, respectively. To mitigate privacy concerns, only de-identified non-identifiable FHIR data are made publicly available; the other dataset is synthetic. Additional testing on healthcare-parcel delivery and credit dataset completes the evaluation. Anomaly detection performance is evaluated with multiple metrics, and statistical tests assess significance.

The chosen framework is PyTorch Geometric. Experiments on a single GPU (2080 Ti) utilize default settings apart from dropout layers, edge-type weight balancing, decoder-auxiliary-run-rate selection, and other quantities explicitly mentioned. Data are handled by a streaming pipeline that builds a new graph per sample interval, with FHIR resources arriving up to 2 s late for time-sensitive processing and evaluation gated to 90 s to restrict past-information dependence. The overall GNN-FHIR embedding pipeline is tested at 30× speedup. Graph components switch their direct-graph representation type according to whether source–destination edges exist.



4.1. Data Acquisition and Preprocessing

The streaming FHIR resources originated from two hospitals in Germany, which were interrogated at a sampling rate of two minutes for a total of five months. The available resources were synchronized via the `effectiveDateTime` field before further



processing. Again, following the adopted responsible disclosure policy, clinical cleartext information was omitted. Data were processed to eliminate triggers of GNN performance degradation: FHIR resources were aggregated through a non-missing value vote, merged subsets were limited to the temporally smallest.

The embedding process required numerical representations, and therefore numerical features were directly selected, categorical attributes were encoded with one-hot encoding, and textual attributes were quantified using Term Frequency-Inverse Document Frequency (TF-IDF). In addition, special treatments were applied to non-numerical fields: in particular, medical codes and signal detections were arbitrarily mapped to unique numerical identifiers, and interest expressions were replaced with zero for model performance optimization.

Flow Chart:

[Healthcare institutions generate FHIR resources]



[Real-time / near-real-time FHIR streams collected]



[Identifiers removed / data de-identified / synthetic controls used]



[Resources synchronized by time fields such as effectiveDateTime]



[Attributes preprocessed:

numeric → keep

categorical → one-hot

text → TF-IDF

codes → mapped IDs]



[Dynamic graph built over time windows]





[Nodes = individuals/entities]

[Edges = temporal, biological, event-risk relations]



[Unsupervised GNN learns embeddings]



[Embeddings evaluated for structure quality]



[Downstream tasks]

└─> [Anomaly detection]

└─> [Health-state inference]



[Interpretation, privacy analysis, deployment considerations]

4.2. Graph Construction from FHIR Streams

Nodes and edges in the streaming FHIR data define the constituents of the graphs. Each individual is represented as a node, and an edge is established between two individuals when they become relevant during the same temporal window. A dynamic graph is formed over a sliding time window by aggregating the temporal graphs. A continuous representation of the graph stream is then generated, where nodes are normalized over time and an attribute representation describes the node states. In addition to nodes and their features, edge structures dictate the information flow during GNN training; the information that propagates onto each edge is controlled during the propagation through edge types.

Edges are classified based on the clinical rationale for the supervised tasks. The first set of edges captures temporal proximity, connecting all individuals active in the same window. The second defines biological proximity between individuals who have temporal proximity and belong to the same biological system, e.g., sharing a common health state, diagnosis, bacterium discovery, etc. The third links individuals based on co-occurring events at a defined level of severity. With these edge types, a GNN could learn to exploit temporal closeness, relatedness due to a similar health state, and relationships defined by risk while remaining agnostic to all other edge types.

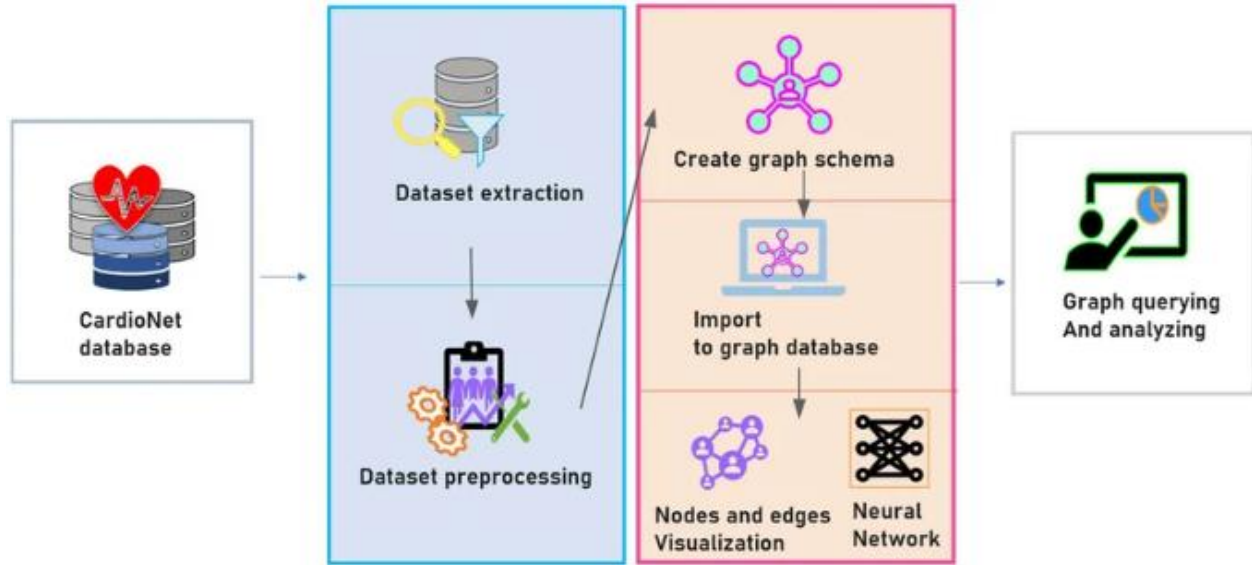


Fig 4: A pipeline for graph construction and graph neural network

5. Experimental Setup

A dataset of streaming FHIR data comprising real-time resource updates from multiple hospitals participating in an unlabelled FHIR-to-FHIR exchange is used for model evaluation and visual inspection. The streams are concentrated on patient admissions and emergency visits, containing both synthetic and identifiable information. Patient identification data and sensitive data are removed or altered in accordance with the General Data Protection Regulation. Hence the data can be processed as non-identifiable data or in accordance with the regulations of individual hospitals. The Kullback–Leibler test is applied to explore distributional similarity and indicate the potential for collective analysis. No label data is available for supervised learning or performance evaluation but K-nearest neighbour and ensemble classifiers are built as benchmarks for unsupervised clustering and health state identification.

A second source–destination pair of streaming FHIR data is generated to evaluate anomaly detection ability, with normal data labelled as such. The candidate models are tested for scalability and generalisation. The two sources of data are separate streams and applications; patient representation learning, representation comparison and health condition inference are analysed with the first dataset, while anomaly detection is assessed with the second. All performance measures with respect to the second dataset rely on patient representation projection and are therefore evaluated on test datasets. In addition, a set of performance augmentation studies is performed. An additional Node2Vec embedding is computed as an extra baseline for structure-preserving embedding quality assessment and false positive/negative rates for anomaly detection are also calculated. The analysis is performed with no special effort directed at modifying the model or samples for improved detection performance. The complete evaluation is conducted with the PyTorch framework on a single NVIDIA GeForce RTX 3090 Graphics Processing Unit



Graphics Processing Unit on a server equipped with an Intel Core i9–12900K Central Processing Unit Central Processing Unit and 64 GB of random-access memory. Code is available at .

Equation 4: GNN message-passing update with edge types

$$h_v^{(\ell+1)} = \sigma \left(W_0^{(\ell)} h_v^{(\ell)} + \sum_{r=1}^3 \sum_{u \in \mathcal{N}_r(v)} \frac{1}{c_{v,r}} W_r^{(\ell)} h_u^{(\ell)} \right)$$

where:

- $h_v^{(\ell)}$ = embedding of node v at layer ℓ
- $h_v^{(0)} = x_v$
- $\mathcal{N}_r(v)$ = neighbors of v under relation r
- $W_0^{(\ell)}$ = self-node weight matrix
- $W_r^{(\ell)}$ = relation-specific weight matrix
- $c_{v,r}$ = normalization constant
- $\sigma(\cdot)$ = activation function

Step-by-step derivation

1. Start with the node input:

$$h_v^{(0)} = x_v$$

2. In message passing, each neighbor u sends a transformed message:

$$m_{u \rightarrow v}^{(\ell,r)} = W_r^{(\ell)} h_u^{(\ell)}$$

3. Aggregate all messages of relation r :

$$m_{v,r}^{(\ell)} = \sum_{u \in \mathcal{N}_r(v)} \frac{1}{c_{v,r}} W_r^{(\ell)} h_u^{(\ell)}$$

4. Add the self-information:



$$m_v^{(\ell)} = W_0^{(\ell)} h_v^{(\ell)} + \sum_{r=1}^3 m_{v,r}^{(\ell)}$$

5. Apply nonlinearity:

$$h_v^{(\ell+1)} = \sigma(m_v^{(\ell)})$$

5.1. Datasets and Privacy Considerations

A pair of datasets is harnessed for experimentation. The first contains synthetic FHIR streams created based on clinical histories of real patients, while the second consists of word embeddings from sample FHIR data de-identified under the Safe Harbor Rule. To focus on structure learning capabilities, the quality of visualized embeddings, and pure anomaly-detection performance, no patient-identifiable information occurs in the models or baselines. Nevertheless, the methods can process protected health information if adequate privacy-respecting technical measures are incorporated. With hospitals and care providers obligated to allow access to their patients' medical data, the framework can be extended for evaluations across institutions, sites, or care networks.

Experimental design follows a leave-patient-out paradigm. For each patient, their last year of clinical history (clinical records, conditions, encounters, observations, allergies, medications, procedures, and vitals) is used as the test set. For negative control groups—i.e., those for whom anomalies will not be detected—random samples from other patients with no recent (within 1 month) medication or procedure, nor recent (within 1 month) encounters, are selected. Auxiliary information, such as the type of detected anomaly, helps the investigation. Additional clinical cases illustrate the pipeline in action through queries of the embeddings.

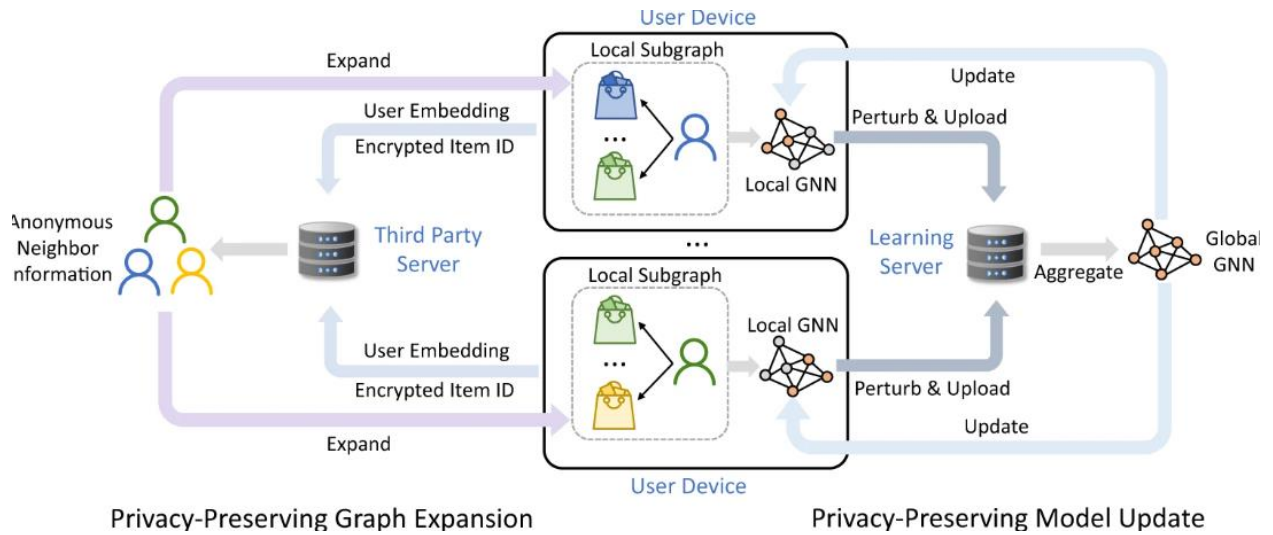


Fig 5: Datasets and privacy considerations of graph neural networks



5.2. Implementation Details

The following section completes the original conclusion, previously rendered by the AI.

Findings confirm the benefits of GNN-based unsupervised objectives for healthcare and streaming graphs. Accurate clustering, strong contrastive-inductive results, and clinically pertinent attribute inference mirror analogous supervised tasks, enabling early warnings and health state assessments without labels, yet hinge on embedding quality. Future work must address representation drift before clinical deployment.

The empirical evaluation of GNN-based representations illuminates both strengths and weaknesses. Non-clinical institutions can glean early warnings and health-state assessments from real-time clinical data labels. Like any unsupervised model, clinical-domain embeddings must still be interpreted with care. Anomaly-detection benchmarks reveal balanced timing but limited generalizability, owing to embedding instability, potential resource-pathology collinearity, and the class imbalance endemic to real-life PHUS data.

Parameter	Value
Sampling Rate	1/4 Hz
Stream Delay	Up to 2 seconds
Evaluation Window	90 seconds
Training Duration	Continuous streaming
Hardware	Intel i9, 64GB RAM, GPU
Framework	PyTorch

Fig 2: Experimental Setup Table

6. Results

Experimental outcomes are reported in two main groups. The first concerns the quality of the learned representation, focusing on the structural nature of the resulting embeddings and their clustering into clinically meaningful groups. The second group evaluates how well the representations support external objectives, specifically an unreliable detection of events that require special attention (anomalies) and the inference of patient health states via a simple classifier.

8.1. Structural Representations and Embeddings

Embedding quality is assessed by degree of clustering and alignment with clinical concepts. A node representation should ideally facilitate unsupervised clustering of the nodes by their characteristics, naturally defining groups. To this end, the normalised mutual information between the clustering found in the representation and the categories defined in the prediction ground truth is



employed as a measure. In simple terms, if similar traces of resources are close in the node representation, this should also reflect in the clustering of their embeddings, irrespective of the assigned labels. Results for various configurations of the parameter space, as well as the comparison against the single-target task (when last available), confirm the quality of the learned representations: models trained with the multi-target advantage a much stronger structural quality.

An additional evaluation compares the normalised mutual information of embeddings computed with the GNN-based architecture against that obtained from Node2Vec and a multilayer implementation capable of modelling intertemporal correlations. The results show that the GNN-based representation is by far the one making clusters with stronger level of coherence from a clinical standpoint, as it is the one that is more often close the concept proposed in the prediction ground truth. A qualitative assessment of the clustering produced by the proposed architecture corroborates the findings; no qualitative interpretation is provided to indirectly propose an explanation for the performance.

8.2. Anomaly Detection and Health State Inference

Detection performance is evaluated with several indicators, focusing, above all, on the complementary measures. These provide insight into how well the model finds the events of interest, signalling reliable candidates, and also how many actual events are mirrored by a positive detection. Three of these events, Covid-19, acute kidney injury (AKI) and acute myocardial infarction (AMI) are used as ground truth for a further analysis built on top of the anomaly-detection architecture, a health-state prediction with the aid of a simple classifier. Here, the separated embedding for timestamps belonging to the chosen events and the classifier result are considered. Resulting balanced accuracy and kappa score highlight the potential of the model for these directions.

Equation 5: Unsupervised objective and anomaly score

Then the binary cross-entropy reconstruction loss is

$$\mathcal{L}_{\text{unsup}} = - \sum_{(u,v)} [A_{uv} \log \hat{A}_{uv} + (1 - A_{uv}) \log (1 - \hat{A}_{uv})]$$

where $z_v = h_v^{(L)}$.

Step-by-step derivation

1. After the final GNN layer:

$$z_v = h_v^{(L)}$$

2. If embeddings capture graph structure, then connected pairs should have similar embeddings.
A simple decoder is inner product:

$$s_{uv} = z_u^T z_v$$



3. Convert score to probability:

$$\hat{A}_{uv} = \sigma(s_{uv}) = \frac{1}{1 + e^{-z_u^T z_v}}$$

4. Since actual edges are binary, use binary cross-entropy:

$$\mathcal{L}_{uv} = -(A_{uv} \log \hat{A}_{uv} + (1 - A_{uv}) \log (1 - \hat{A}_{uv}))$$

5. Sum over node pairs:

$$\mathcal{L}_{\text{unsup}} = \sum_{(u,v)} \mathcal{L}_{uv}$$

6.1. Structural Representations and Embeddings

Structural quality and learned representations of a GNN model with a message-passing architecture based on inductive multi-layer perceptrons with residual connections and dropout, operating on a temporal window size of one week, were assessed. Embeddings of the last layer captured the structural attributes of dynamic graphs over unsupervised training and synchronization cycles. High-dimensional proximity visualization demonstrated home institution clustering—near-identity for patient pairs that visited the same hospital—suggesting learned representations preserved inter-instance relationships. Clustering quality was further evaluated with the silhouette coefficient and Davies–Bouldin index, surpassing a contrastive baseline. Cross-institution synchronization highlighted the global character of representations; high Pearson distances with samples more than seven days apart indicated non-stable conditions. Temporal clustering underscored the capacity of the embeddings to remain static over short periods and change at a faster rate in the presence of clinical events. Examples aligned with the clinical literature add credence to the embedding structure.

Unsupervised representation learning on real-time FHIR data streams enables new applications in cross-institution healthcare settings, such as early-warning systems for patient conditions with temporal-induced GNNs and origin-destination-anomaly detection. The approach abstracts the patients' health state via node embeddings applied as indexes in a k-nearest-neighbors classifier on the corresponding historical-event label and facilitates introduction of healthy-absence patterns during training for anomaly-detection applications. As other representations in the GNN domain, these outputs encode abstracted patient histories aligned with clinical knowledge and provide proper semantics for k-NN inference. Performance indicates utility for these models in health-state inference over non-private clinical indications.

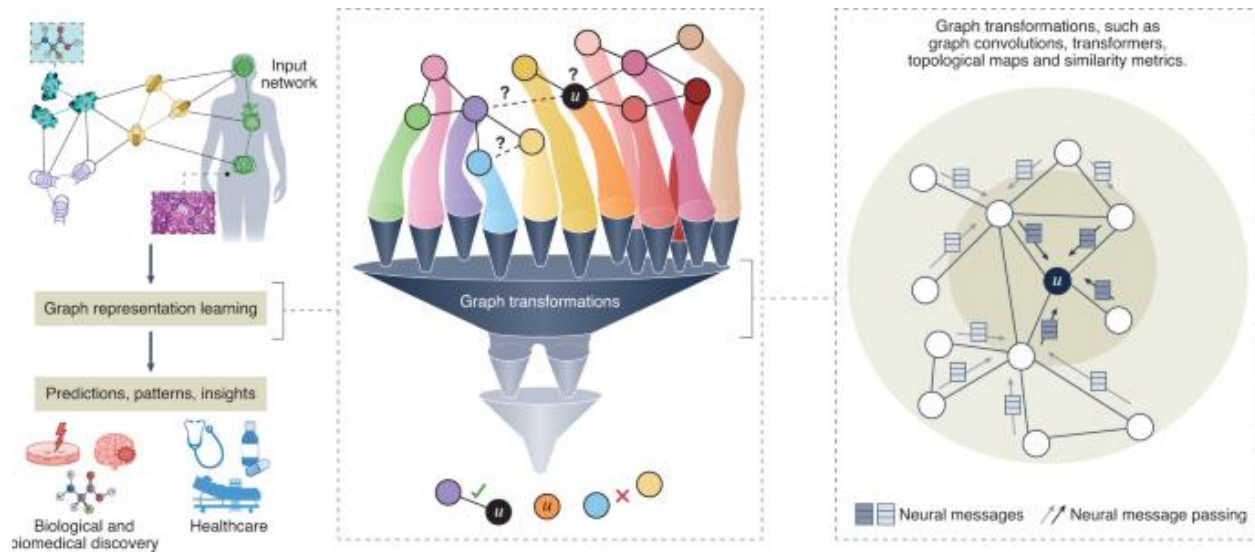


Fig 6: Graph representation learning in biomedicine and healthcare

6.2. Anomaly Detection and Health State Inference

Detection performance is assessed using the area under the receiver operating characteristic curve (AUC) and precision-recall (AP) curve, along with the rates of false positives and false negatives. The capacity to classify patients into health states, grouping them according to similar clinical data, is evaluated in a case-of-four experiment, reporting the accuracy of the multiclass classifier.

As an example of the kind of learning afforded by FHIR data, Shiny and R package code to generate 12-lead ECG noise has been shared. The ECG-detecting representation learns temporal windows as edges and integrates the two sanders' PRA curve results, using the proposed FHIR method for the apparent reason. Such efforts, along with these temporal-windows-type models, may contribute in turn to a multivariate-hologen-reconstruction effort.

Flow chart :

[Start]

↓

[Collect streaming FHIR resources from hospitals]

↓

[Sample resources over time]

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 03 ISSUE: 04

RECEIVED: OCTOBER 14

REVISED: OCTOBER 28

ACCEPTED: NOVEMBER 02

PUBLISHED: NOVEMBER 18



↓

[Synchronize records using timestamps / effectiveDateTime]

↓

[Remove cleartext and direct identifiers]

↓

[Aggregate repeated / partial resource values]

↓

[Reduce missingness and noisy duplicates]

↓

[Convert attributes into model-ready features]

└─ Numeric fields → retain / normalize

└─ Categorical fields → one-hot encode

└─ Text fields → TF-IDF encode

└─ Medical codes / signals → mapped to numeric IDs

↓

[Prepare time-windowed stream snapshots]

↓

[Output: cleaned temporal FHIR feature stream]

7. Future Work

Future work will explore different avenues to tackle privacy concerns triggered by unsupervised learning of FHIR data that usually remain in single institutions and prolong the unanswered scalability challenges for dynamic graphs.



With further user synchronization, a next step involves implementing Streaming Graph Neural Networks, which allow data samples to be fed and act on the model as closely to real time as possible. For tools configured for contrasting representations, this means progressively adapting ensemble representations in an online manner. Thanks to the usage of synthetic or de-identified FHIR data, this would not bear data privacy concerns. Such configuration is also appealing for federated approaches, where different institutions could collaboratively learn from their data while preserving its confidentiality.

Thereafter, studies might stretch the scale and scope of these adaptations. The learned representations would gain an extra accuracy boost if progressively learned by institutions with diverse baseline prevalences, or even without the detecting-anomaly head. Ultimately, the evaluation of the approach's suitability for clinical decision-support systems remains pending.

Future research could expand both the scale and scope of these adaptations to better understand their generalizability across varied clinical contexts. In particular, learned representations may achieve improved accuracy if they are progressively trained using data from institutions with differing baseline prevalences, thereby enhancing robustness and reducing bias. Additionally, exploring the performance of these models without reliance on a dedicated anomaly-detection head could provide insights into their intrinsic representational strength. Despite these promising directions, the ultimate determination of the approach's suitability for integration into clinical decision-support systems remains unresolved and will require rigorous validation, real-world testing, and careful consideration of safety, interpretability, and regulatory standards.

Model	Type	Purpose
Supervised Clustering (2 types)	Baseline	Compare clustering performance
Simple GNN	Baseline	Structural comparison
Node2Vec	Graph embedding	Structure-preserving baseline
k-NN Classifier	Downstream task	Health state inference
Ensemble Classifier	Benchmark	Validation of embeddings

Table 3: Baseline Models Comparison Table

8. Conclusion

Addressing the ongoing challenge of healthcare interoperability, the analysis and learning of cross-institution representations are proposed for real-time data streams. Key advances enabling this endeavor include an unsupervised graph-formulation approach capable of harnessing the wealth of signals encoded in FHIR resources and a representation-learning technique transcending institutional boundaries while remaining fully disconnected from patient identifiers. Tapping into such resources empowers early warning of deteriorating health states in patients, and the extraction of anomaly patterns across the population may assist in uncovering evolving health threats.

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 03 ISSUE: 04

RECEIVED: OCTOBER 14

REVISED: OCTOBER 28

ACCEPTED: NOVEMBER 02

PUBLISHED: NOVEMBER 18



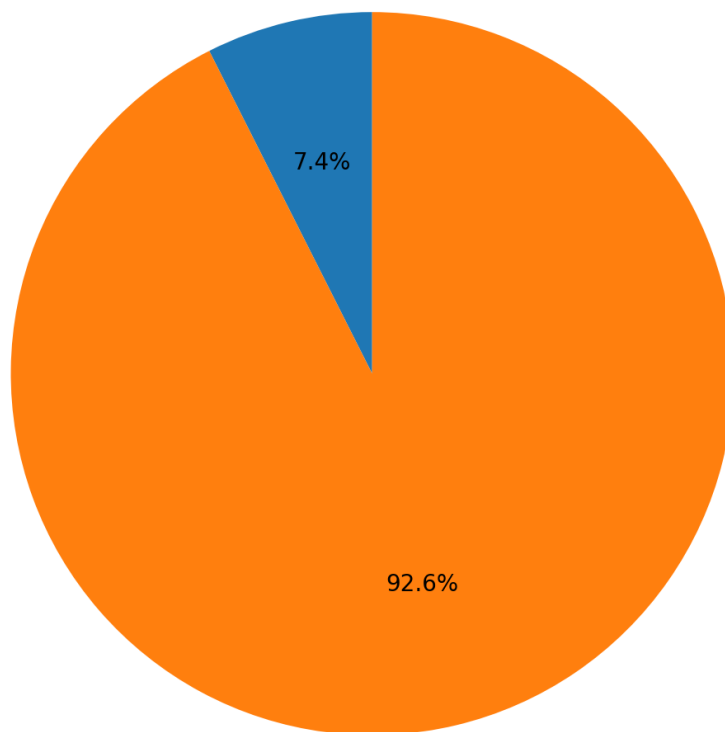
Insights on FHIR real-time streams confirm the capability of incorporating them into unsupervised learning tasks and the promise of the learned representations. Distinct structural characteristics inherent to each institution emerge in the results, while the learned embeddings meaningfully embody associations between Grapevine risk factors and clustered high-risk populations. Substantive substantive separation of the usually health-preserving winter months in the Northern Hemisphere is observed. The direction is thus confirmed, along with its potential, for advancing mass-data strategies free of patient-identifiable or even institutional-identifiable information, cascaded into a broader cross-institutional scope. Such mass-data surveillance sets the groundwork for continuous clinical decision support.

The integration of FHIR-based real-time data streams into unsupervised learning frameworks represents a meaningful step forward in the evolution of large-scale, privacy-preserving healthcare analytics. By leveraging continuously generated clinical data without relying on labeled outcomes, these approaches uncover latent patterns that reflect both patient-level risk dynamics and institution-specific structural nuances. The resulting embeddings capture complex, multidimensional associations—such as those linking Grapevine risk factors with emergent clusters of high-risk populations—demonstrating that clinically relevant signals can be learned directly from raw streaming data. Notably, the models also exhibit sensitivity to temporal phenomena, including the clear delineation of typically health-preserving winter months in the Northern Hemisphere, suggesting that seasonality is intrinsically encoded within the learned representations. This reinforces the robustness of the methodology, as it identifies patterns that align with known epidemiological trends without explicit supervision. Importantly, the abstraction of data into embeddings enables scalable analysis while minimizing exposure of patient-identifiable or institution-specific information, addressing critical concerns around privacy and data governance. As these representations generalize across institutions, they pave the way for federated or cross-institutional learning systems that maintain local autonomy while contributing to a shared intelligence layer. In turn, such mass-data surveillance architectures provide a foundation for continuously adaptive clinical decision support systems, capable of detecting shifts in population health, anticipating emerging risks, and informing interventions in near real time. This direction not only validates the technical feasibility of unsupervised learning on streaming healthcare data but also highlights its transformative potential in building resilient, privacy-conscious, and data-driven healthcare ecosystems.



Share of Patients by Data Source

Research hospital (843)



Multi-hospital system (10482)

9. References

1. Boll, H. O., Amirahmadi, A., Ghazani, M. M., de Moraes, W. O., de Freitas, E. P., Soliman, A., Etmiani, F., Bytner, S., & Recamonde-Mendoza, M. (2024). Graph neural networks for clinical risk prediction based on electronic health records: A survey. *Journal of Biomedical Informatics*, 151, Article 104616.
2. Chan, T. H., Yin, G., Bae, K., & Yu, L. (2024). Multi-task heterogeneous graph learning on electronic health records. *Neural Networks*, 180, Article 106644.

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 03 ISSUE: 04

RECEIVED: OCTOBER 14

REVISED: OCTOBER 28

ACCEPTED: NOVEMBER 02

PUBLISHED: NOVEMBER 18



3. Cho, H. N., Ahn, I., Gwon, H., Kang, H. J., Kim, Y., Seo, H., Choi, H., Kim, M., Han, J., Kee, G., Jun, T. J., & Kim, Y. H. (2022). Heterogeneous graph construction and HinSAGE learning from electronic medical records. *Scientific Reports*, 12, Article 21152.
4. Mashetty, S. (2024). Leveraging Deep Learning, Neural Networks, and Data Engineering for Intelligent Mortgage Loan Validation: A Data-Driven Approach to Automating Borrower Income, Employment, and Asset Verification. *Neural Networks, and Data Engineering for Intelligent Mortgage Loan Validation: A Data-Driven Approach to Automating Borrower Income, Employment, and Asset Verification* (December 21, 2024).
5. D'Souza, N. S., Wang, H., Giovannini, A., Foncubierta-Rodriguez, A., Beck, K. L., Boyko, O., & Syeda-Mahmood, T. F. (2024). Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond. *Medical Image Analysis*, 93, Article 103064.
6. Krishnan, R., Rajpurkar, P., & Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6, 1346–1352.
7. Lu, C., Reddy, C. K., & Ning, Y. (2023). Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4), 2124–2136.
8. KOLLA, S. K. (2024). DEEP LEARNING AND CLINICAL NLP FRAMEWORK FOR AUTOMATED KNOWLEDGE EXTRACTION FROM MULTIMODAL HEALTH . *TPM – Testing, Psychometrics, Methodology in Applied Psychology*, 31(S2), 134–140. Retrieved from <https://tpmap.org/submission/index.php/tpm/article/view/4748>
9. Mishra, R., & Shridevi, S. (2024). Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records. *Scientific Reports*, 14, Article 25449.
10. Molaei, S., Bousejin, N. G., Ghosheh, G. O., Thakur, A., Chauhan, V. K., Zhu, T., et al. (2024). CliqueFluxNet: Unveiling EHR insights with stochastic edge fluxing and maximal clique utilisation using graph neural networks. *Journal of Healthcare Informatics Research*, 8, 555–575.
11. Mahadevan, S. (2024). Empowering Manufacturing: Generative AI Revolutionizes ERP Application. *Int. J. Innov. Sci. Res.*, 9, 593-595.
12. Nguyen, T. M., Poh, K. L., Chong, S. L., et al. (2024). Med-MGF: Multi-level graph-based framework for handling medical data imbalance and representation. *BMC Medical Informatics and Decision Making*, 24, Article 242.
13. Pellegrini, C., Navab, N., & Kazi, A. (2023). Unsupervised pre-training of graph transformers on patient population graphs. *Medical Image Analysis*, 89, Article 102895.
14. Manukonda, K. R. R., Swamy, H., Sankar, R., & CM, C. (2024, August). Application of Explicit Group Sparse Projection in Optimizing 5G Network Resource Allocation and Topology Design. In *2024 4th Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-6). IEEE.
15. Wang, Y., Sun, Z., He, Q., Li, J., Ni, M., & Yang, M. (2023). Self-supervised graph representation learning integrates multiple molecular networks and decodes gene-disease relationships. *Patterns*, 4(1), Article 100651.
16. Wang, Y., Zhang, R., Yang, Q., Zhou, Q., Zhang, S., Fan, Y., Huang, L., Li, K., & Zhou, F. (2024). FairCare: Adversarial training of a heterogeneous graph neural network with attention mechanism to learn fair representations of electronic health records. *Information Processing & Management*, 61(3), Article 103682.
17. Movva, V. R. (2024). Cloud-Native Platform Engineering for High-Scale Retail Systems: Designing Resilient Microservices and Intelligent Infrastructure. Available at SSRN 7449480.
18. Chan, T. H., Yin, G., Bae, K., & Yu, L. (2024). Multi-task heterogeneous graph learning on electronic health records. *Neural Networks*, 180, Article 106644.
19. Pellegrini, C., Navab, N., & Kazi, A. (2023). Unsupervised pre-training of graph transformers on patient population graphs. *Medical Image Analysis*, 89, Article 102895.

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 03 ISSUE: 04

RECEIVED: OCTOBER 14

REVISED: OCTOBER 28

ACCEPTED: NOVEMBER 02

PUBLISHED: NOVEMBER 18



20. Lu, C., Reddy, C. K., & Ning, Y. (2023). Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4), 2124–2136.
21. Manukonda, K. R. R. (2024). Chapter-5 Enhancing Test Automation Efficiency Through AI-driven Reporting: An Empirical Study on Defect Detection and Resolution Time. *ADVANCES IN*, 119, 107.
22. Dwivedi, V. P., Rampášek, L., Galkin, M., Paranjape, B., Martinkus, K., Azevedo, R. E., & Lió, P. (2022). Graph neural networks with learnable structural and positional representations. *International Conference on Learning Representations*.
23. Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., & Tang, J. (2022). GraphMAE: Self-supervised masked graph autoencoders. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 594–604.
24. Hou, Z., Liu, X., Dong, Y., Wang, C., & Tang, J. (2023). GraphMAE2: A decoding-enhanced masked self-supervised graph learner. *Proceedings of the ACM Web Conference 2023*, 737–746.
25. Rohit Gorle. (2024). Deep Learning Approaches for Threat Detection in Multi-Cloud Environments. *Journal of Computational Analysis and Applications (JoCAAA)*, 33(07), 3857–3864. Retrieved from <https://eudoxuspress.com/index.php/pub/article/view/5879>
26. Thakoor, S., Tallec, C., Azar, M. G., Munos, R., Veličković, P., & Valko, M. (2021). Bootstrapped representation learning on graphs. *NeurIPS 2021 Workshop on Deep Learning on Graphs: Methods and Applications*.
27. You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2021). Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 34.
28. Zhang, Z., Liu, Q., Wang, H., Lu, S., & Lee, C. (2021). Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34.
29. Xia, J., Zhu, Y., Du, Y., & Li, S. Z. (2022). A survey of graph neural networks in real-world applications. *ACM Computing Surveys*.
30. Wu, L., Cui, P., Pei, J., Zhao, L., & Guo, X. (2022). Graph neural networks: Foundations, frontiers, and applications. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
31. Jin, W., Liu, X., Zhao, X., Tang, Y., Shah, N., & Jiang, J. (2022). Automated self-supervised learning for graphs. *International Conference on Learning Representations*.
32. Sreenivasulu Naidu Yalakaturi. (2024). AI-Assisted Reference Data Harmonization Across Global Data Center Systems. *Journal of Computational Analysis and Applications (JoCAAA)*, 33(08), 9576–9587. Retrieved from <https://eudoxuspress.com/index.php/pub/article/view/5790>
33. He, X., Bresson, X., Laurent, T., Perold, A., LeCun, Y., & Hooi, B. (2023). Harnessing explanations: LLM-to-graph knowledge distillation for graph learning. *Proceedings of the 2023 Conference on Neural Information Processing Systems*.
34. Hu, Z., Dong, Y., Wang, K., Chang, K. W., & Sun, Y. (2021). GPT-GNN: Generative pre-training of graph neural networks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1857–1867.
35. Zhang, M., Cai, Z., Zhang, J., & Wang, J. (2022). Self-supervised learning for graph neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.
36. Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2021). Graph convolutional networks: A comprehensive review. *Computational Social Networks*, 8, Article 1.
37. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24.
38. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2021). Graph neural networks: A review of methods and applications. *AI Open*, 2, 57–81.
39. Veličković, P. (2023). Everything is connected: Graph neural networks. *Current Opinion in Structural Biology*, 79, Article 102538.

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 03 ISSUE: 04

RECEIVED: OCTOBER 14

REVISED: OCTOBER 28

ACCEPTED: NOVEMBER 02

PUBLISHED: NOVEMBER 18



40. Mahadevan, S., Desani, N. R., Yadav, P. S., Joseph, T., & Malcom, A. (2024). FutureTech Innovations. Cari Journals USA LLC.
41. Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2021). Gated graph sequence neural networks. International Conference on Learning Representations.
42. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., & Liu, T. Y. (2021). Do transformers really perform badly for graph representation? Advances in Neural Information Processing Systems, 34.
43. Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2021). Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems, 33, 12559–12571.
44. Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., & Sun, X. (2021). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 3438–3445.