Enriching Electronic Health Record with Semantic Features Utilising Pretrained Transformers

Lena AlMutair[1,2] , Eric Atwell[2], Nishant Ravikumar[2]
[1] Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, SA.
[2] School of Computing, University of Leeds, Leeds, UK.
{scla,e.s.atwell,n.ravikumar}@leeds.ac.uk

Abstract

Electronic Health Records (EHRs) have revolutionised healthcare by enhancing patient care and facilitating provider communication. Nevertheless, the efficient extraction of valuable information from EHRs poses challenges, primarily due to the overwhelming volume of unstructured data, the wide variability in data formats, and the lack of standardised labels. Leveraging deep learning and concept embeddings, we address the gap in context-aware systems for EHRs.

The proposed solution was evaluated on the MIMIC III dataset and demonstrated superior performance compared to other methodologies. We addressed the positive impact of the latent feature combined with the note representation in four different settings. Model performance was evaluated using a case study conducted with BertScore, assessing precision, recall, and F1 scores. The model excels in Medical Natural Language Inference (MedNLI) with an 89.3% accuracy, further boosted to **90.5%** through retraining the embeddings using International Classification of Diseases (ICD) codes, which we formally designate as *ClinicNarrIR*. The ClinicNarrIR was tested with 1000 randomly sampled notes, achieving an $NDCG$@10 score of approximately 0.54 with $accuracy$@10 of 0.85. The study also demonstrates a high correlation between the results produced by the proposed representation and medical coders. Notably, in all evaluation cases, the optimal base pretrained model that emerged was BlueBERT .

1 Introduction

In healthcare, Electronic Health Records (EHRs) have gained widespread adoption due to their potential to enhance patient care, minimise medical errors, and promote efficient communication among healthcare providers. However, extracting relevant information from these extensive records remains a challenging and time-consuming task for clinicians. The primary challenge, besides the lack of labeled data, is also the absence of an effective and context-aware system that can semantically retrieve valuable information from EHRs based on practitioner queries (Li et al., 2022a; Wang et al., 2018b). Current approaches often rely on structured data or keyword matching/expansion and overlook contextual nuances, clinical history, and natural language intricacies. Retrieving the information is essential beyond patient care; it is vital for secondary clinical use cases, including clinical trial matching (Mc Cord and Hemkens, 2019; Jin et al., 2023; Goldstein, 2020), drug repurposing (Liu et al., 2021), quality assurance (Hanna von Gerich, 2022; Huang et al., 2018), clinical decision support (Mills, 2019; Sutton et al., 2020), research access (Raman et al., 2018), and population health management (Kruse et al., 2018). These applications demand a comprehensive approach that goes beyond traditional methods, contributing not only to improved patient care but also to significant advancements and breakthroughs in the field.

To address this gap, our research aims to take a significant step towards the development of an automated clinical information retrieval system. The primary objective is to design a context-enhanced network capable of semantically searching Electronic Health Records (EHRs), returning relevant clinical notes in response to practitioner queries, and embedding clinical narratives with semantic features. A *clinical narrative* refers to the free-text entries made by healthcare practitioners to track a patient's progress from admission to discharge (Pakhomov et al., 2007). On the other hand, a *clinical concept* refers to specific entities within the clinical notes, which could include diagnoses or diseases, serving as semantic features.

The research explores deep learning pretrained models to represent these embeddings and investigates the impact of an additional element (medical concept) through four different settings. Additionally, it enhances medical language representation by training a network to learn the clinical narrative vector as the ground truth query. These contributions aim to advance the field by improving clinical information extraction and testing model performance. This comprehensive approach ultimately leads to enhanced clinical decision-making and patient care

The utilisation of Medical Natural Language Processing (MLP) in processing EHRs is essential due to the unique characteristics of clinical notes (Ford et al., 2016). These notes often contain abbreviations and deviate from standard English grammar rules (Ford et al., 2016). In contrast, discharge summaries adhere to more formal language standards, making techniques developed for non-medical text sources potentially applicable (Ford et al., 2016).

Information Extraction (IE) in NLP automates the identification of concepts, entities, events, and their relations in free text (Wang et al., 2018a). IE encompasses various subtasks, including Named Entity Recognition (NER), co-reference resolution, and relation extraction (Wang et al., 2018a). NER is especially significant in the medical field, where it is used to classify entities such as genes, medicines, and diseases within unstructured text (Perera et al., 2020). Conditional Random Fields (CRF), a well-known model, contributes to concept extraction by considering words before and after the entity (Huang et al., 2015)

BERT, a leading NLP model (Devlin et al., 2018), has gained traction in the health informatics field with the development of domain-specific versions like ClinicalBERT (Alsentzer et al., 2019). ClinicalBERT, pretrained on clinical text from the MIMIC-III database (Johnson et al., 2016), provides word embeddings for research without introducing novel training techniques. It outperforms general BERT in clinical tasks like Named Entity Recognition (NER, i2b2) and medical natural language inference (Uzuner et al., 2011).

Additionally, BioBERT is another domain-specific BERT model, trained on biomedical research data sourced from PubMed (Lee et al., 2020). Its specialization in biomedical texts enhances performance across various biomedical NLP tasks (Lee et al., 2020).

Key research themes in EHR include Information Extraction, EHR Representation, Outcome Prediction, Computational Phenotyping, and Clinical Data De-Identification (Tran et al., 2015; Li et al., 2015; Huang et al., 2019; Denaxas et al., 2018; Carroll et al., 2011; Stubbs et al., 2015). Information Extraction involves concept extraction, temporal event identification, and abbreviation expansion, primarily using clinical notes. EHR Representation uses medical codes to represent concepts/patients. Outcome prediction includes applications like hospital readmission and mortality prediction (Huang et al., 2019; Che et al., 2018).

SemEHR, a tool for semantic search in a clinical context, uses ontology to locate mentions of biomedical topics in EHRs (Wu et al., 2018). Unsupervised embedding is advantageous, given the challenges of obtaining human labels for clinical records. Combining medical concepts with clinical narrative, as demonstrated by (Huang et al., 2022), outperforms other unsupervised approaches in tasks like mortality prediction and patient relatedness (Huang et al., 2022).

In the upcoming sections, the methodology will be detailed, discussing indexing, retrieval, and the value of retraining embeddings. Subsequently, the evaluation includes a case study, MedNLI assessment, and correlation analysis, followed by discussions of pretrained models, and limitations.

2 Methodology

In Siamese BERT, a fixed-sized sentence embedding is created by adding a pooling layer to the BERT output, facilitating the comparison of input similarity. Siamese Neural Networks consist of multiple identical subnetworks that share parameters and configurations, aiding in various applications (Chicco, 2021).

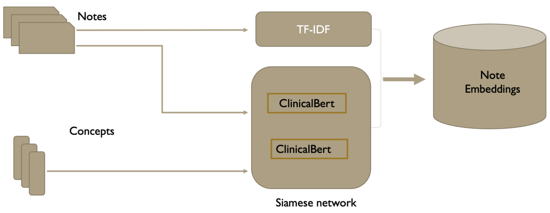


Figure 1: Encoding clinical narrative combined with concept provide more contextual information alongside TF-IDF vectors

The document indexing phase, as illustrated in Figure 1, leverages deep learning and keyword-based methods, such as TF-IDF (Term Frequency - Inverse Document Frequency), to enable the semantic and syntactic searchability of clinical notes at scale. Concept embeddings provide detailed information about clinical entities within the notes, while narrative embeddings capture their overall relevance and context. By integrating these embeddings, the retrieval stage can identify notes with similar clinical concepts, thereby enhancing the relevance of results for clinicians.

During indexing, clinical text n is transformed into vectors $[n_1, n_2, .., n_n]$, where each note is embedded as $[NE_1, NE_2, .., NE_n]$ and combined with concept embeddings $[CE_1, CE_2, ..., CE_n]$. The concepts were extracted using the MedCat tool (Kraljevic et al., 2021) and filtered, retaining symptoms, past diseases, and disorders. This process ensures that the relevance ranking model prioritises clinically significant information while reducing noise and enhancing semantic understanding. By considering the clinical context and incorporating these filtered concepts, the system improves information retrieval, aiding healthcare practitioners in making informed decisions and delivering high-quality patient care.

To evaluate the effectiveness of the network, particularly the added value of concept embeddings, performance was assessed by translating notes into embeddings and exploring the impact of adding concepts. Different settings were examined: using only clinical narrative (setting 1), using multiple concept embeddings (setting 2), combining concept embeddings and averaging them into a single vector, represented as $[CE_1 \ldots CE_z]/z$ where z is the number of concepts (setting 3), and combining the note with concepts (setting 4).

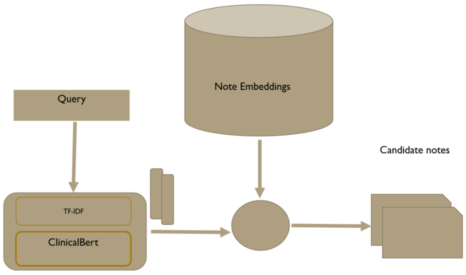


Figure 2: retrieving the best candidates clinical note for a patient by find the highest similarity note with the query

In the retrieval stage depicted in Figure 2 a sorted set of relevant notes is returned from the corpus by computing the cosine similarity between query q and note n.

$$similarity(q, n) = \frac{q.n}{|q|\,|n|} \quad (1)$$

During retrieval using setting 2, Equation 2 is computed and produce multiple vectors with a cosine similarity score that measures the likelihood of a concept being relevant to query q. A higher score indicates greater similarity to the keyword $query$.

$$\forall j \in \{1, 2......n\} \sum_{i=1}^{z} sim(q, CE_i) \quad (2)$$

Where n is the number of notes and z number of concepts in each note.

In setting 3, the combined concepts $[CE_1 \ldots CE_z]$ for all n notes, denoted as $AN_1..AN_n$. the cosine similarity is calculated between query and $\forall i \in AN$, and the n with the maximum similarity score will be retrieved.

$$\forall i \in \{1, 2......n\}.f(i) = sim(q, AN_i) \quad (3)$$

The approach considers various strategies, such as multiplication, or weighted sum, and averaging to combine TF-IDF vectors with embeddings.

In the following section, we will conduct a case study on a female patient records, evaluate MedNLI performance, and analyse Physician and Coder correlation. These assessments will offer insights into the effectiveness of different settings in real-world clinical applications.

2.1 Retraining Embeddings

The essential process of adapting and refining embeddings is explored. While the initial embedding phase lays the groundwork, this section delves into the process of revitalizing these representations through retraining. These dynamic representations are the core of the information retrieval system, ensuring adaptability in the complex realm of clinical language and healthcare information retrieval.

Throughout the training process, the primary goal is to maximize the similarity between a query q and the relevant note n. To achieve this, random negative examples were introduced, following the approach suggested by (Henderson et al., 2017). These negative samples, denoted as n_j $(\forall i \neq j)$, serve to encourage the model to maximize the dissimilarity between q and n_j. This objective is implemented by using the following formula:

$$L(q, n_i) = \max(0, \delta + sim(q, n_j) - sim(q, n_i)) \quad (4)$$

$sim(x, y)$ represents the similarity function between embeddings x and y.

The loss used in this context is a type of contrastive loss, as introduced in (Logeswaran and Lee, 2018). Unlike traditional contrastive losses, which typically involve comparing one positive example with one negative example for each instance, this approach utilizes multiple negative examples for each positive example. Specifically, the loss penalizes the model when the similarity between the query (q) and a negative example (n_i) is not sufficiently greater than the similarity between the query and another presumably negative example (n_j). The parameter δ in this loss represents the minimum difference required between the similarity of the query and any negative example to ensure that the loss is appropriately applied.

The queries are selected as the ICD-9 code (ICD9 $shorttitle$) from MIMIC III (Johnson et al., 2016) since it serves as a robust indicator of medical notes. This is particularly significant because the task of assigning International Classification of Diseases (ICD) codes is typically carried out by skilled medical coders or clinical documentation specialists (Yan et al., 2022).

In the retrieval stage, the trained model leverages its capabilities to obtain embeddings for the query. In addition to these embeddings, we also utilize the TF-IDF representation to facilitate the retrieval process. This combination enables us to find the most suitable matches among the notes, all of which are embedded using the same model.
The study presents performance results for various scenarios, where experiments were conducted with different hyperparameters. Specifically, experiments were conducted with variable numbers of training epochs, including 10, 15, and 20. Given the limitations of computational resources, the batch size was adjusted, exploring values such as 16, 20, and 24. Care had to be taken about increasing the batch size further due to hardware constraints, with a maximum GPU RAM capacity of 32 GB. It is worth noting that a maximum sequence length of 512 token was chosen for the notes. Additionally, 20% of the dataset was dedicated to validation to assess the model's performance.

The findings indicate that increasing the size of batches and the number of epochs yielded improved performance, aligning with the model's learning capabilities. To assess the retrieval system's effectiveness, we reported results based on three widely recognised information retrieval metrics: Accuracy($Accuracy@10$), Mean Average Precision ($MAP@100$), and Normalised Discounted Cumulative Gain ($NDCG@10$).

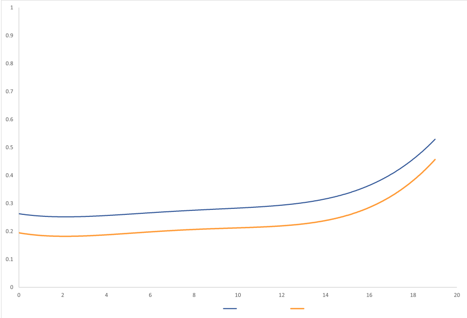


Figure 3: "Figure showing the NDCG@10 (blue) and MAP@100 (orange) curves over 20 epochs, highlighting a peak NDCG@10 score of 0.57 in the validation dataset. The curves illustrate the model's progressive improvement in information retrieval performance during training

Over the course of 20 epoch training process, the cosine similarity score on $Accuracy@10$ reached a peak of 0.9 on the validation dataset, indicating a high degree of similarity between the query and the top retrieved documents. Additionally, the model exhibited significant improvements. The NDCG reached an impressive 0.57, as presented in Figure 3 signifying the model's increasing ability to provide relevant search results.

Furthermore, as the number of training epochs increased, the model consistently improved its performance, with $MAP@100$ exceeding 0.48 on validation dataset. This trend showcases the model's proficiency in ranking and retrieving pertinent information. These observations collectively highlight the model's promising performance. These are standard methods for assessing the significance of documents in the context of information retrieval, highlighting the strong relevance of the retrieved documents to the query.

It is important to note that experiments were conducted to assess the performance of various pretrained transformers, including ClinicalBERT(Alsentzer et al., 2019) and BlueBERT (Peng et al., 2019). The findings consistently showed that BlueBERT outperformed others in a variety of tasks/settings, as demonstrated in the following section, making it an excellent choice as the base for the model. The proposed model will be referred to as *ClinicNarrIR*, which stands for 'Clinical Narrative Information Retrieval'

A visual analysis is conducted using a scatter plot, our specific focus centred on the top 20 queries (as detailed in Appendix A.2) from the

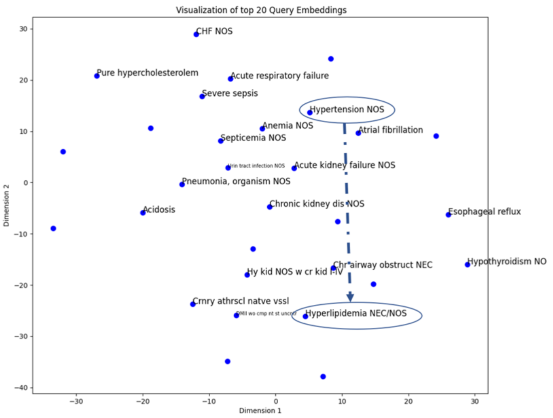


Figure 4: The base model representing the 20 queries in Appendix A.2 without any training.

MIMIC III dataset. In Figure 4, the embeddings' relatedness is showcased before the model was trained.

Interestingly, after training the model for 20 epochs, hyperlipidemia was found to be closely associated with hypertension in the embedding space as depicted in Figure 5. This suggests that the model learned this relationship from the clinical notes it was trained on, even though query similarities were not explicitly part of the training process.

This finding aligns with earlier research from 1991 by Ames (1991), which noted the prevalence of hyperlipidemia in hypertension, though the cause of this association remained unknown at the time. Additionally, a recent study by Tang et al. (2022) highlighted a strong link between high blood cholesterol levels and the development of arterial plaque. This plaque, in turn, can lead to the narrowing and stiffening of blood vessels, resulting in increased blood pressure.

In summary, the model's learned embeddings provided concrete validation of the connection between hyperlipidaemia and hypertension, reinforcing the practical significance of the results and their alignment with existing research.

3 Experiments and Evaluation

Evaluating unsupervised problems, especially in the context of information retrieval (IR), presents distinct challenges. The evaluation involves a case study where we apply BertScore to assess precision, recall, and F1. We also measure accuracy in MedNLI using the proposed approach and its variation. Additionally, we investigate the correlation with physicians/coders in drawing inspiration from (Huang et al., 2019) evaluation.

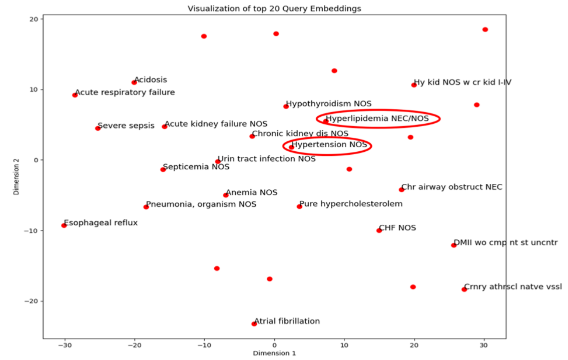
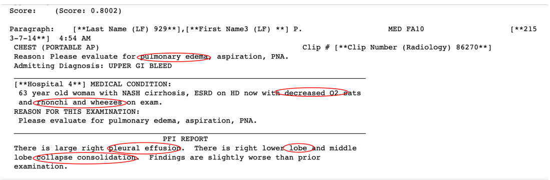


Figure 5: The newly trained model reveals previously unrecognised relationships between the queries in Appendix A.2. This finding suggests that the model has the ability to learn and discover new connections between embeddings.

3.1 Case study: Female with pneumonia infection

In the case study, our patient selection process focused on mirroring real-world healthcare dynamics, emphasizing practitioners' concentration on individual cases rather than aggregate records. Evaluating the patient's medical history using BERTScore, a benchmark known for its correlation with human judgement (Zhang et al., 2019). We applied it to key queries related to prevalent diseases in the MIMIC-III dataset (Johnson et al., 2016), these queries presented in Table 1. Leveraging pretrained ClinicalBERT, we calculated BERTScore, assessing F1, precision, and recall. This analysis ensured contextual relevance between queries and clinical notes, augmenting our investigation of the individual patient's health.

Figure 6 highlights a pertinent note for 'dyspnea,' revealing contextual understanding without explicit keywords. Table 1 presents compelling results in Setting 4, utilizing BERTScore. Noteworthy semantic alignment between high blood pressure and hypertension is depicted in Figure 7. This success, coupled with promising results from Setting 3, underscores the importance of merging contextualized embeddings and domain-specific models in clinical information retrieval. The achievements extend to the top 20 patient presentations in Appendix A.1, demonstrating the efficacy of Setting 4 (Table 2). Additionally, preprocessing, involving stemming and text cleaning, achieves a notable 12% performance improvement, enhancing accuracy in clinical decision-making for the benefit of patient care and healthcare systems.



```
Score:    (Score: 0.8002)

Paragraph:   [**Last Name (LF) 929**],[**First Name3 (LF) **] P.              MED FA10            [**215
3-7-14**]  4:54 AM
 CHEST (PORTABLE AP)                                            Clip # [**Clip Number (Radiology) 86270**]
 Reason: Please evaluate for pulmonary edema, aspiration, PNA.
 Admitting Diagnosis: UPPER GI BLEED
 ______________________________________________________________________________
 [**Hospital 4**] MEDICAL CONDITION:
  63 year old woman with NASH cirrhosis, ESRD on HD now with decreased O2 sats
  and rhonchi and wheezes on exam.
 REASON FOR THIS EXAMINATION:
  Please evaluate for pulmonary edema, aspiration, PNA.
 ______________________________________________________________________________
                                 PFI REPORT
 There is large right pleural effusion.  There is right lower lobe and middle
 lobe collapse consolidation.  Findings are slightly worse than prior
 examination.
```

Figure 6: The first top note retrieved by the query *Dyspnea*. The retrieved note has no indication of *dyspnea*, not even the word *lung*, but it explicitly mentions pulmonary edema, a decrease in O2, wheezing sound, and collapse consolidation in the lobe, which might be a good cause for having dyspnea.

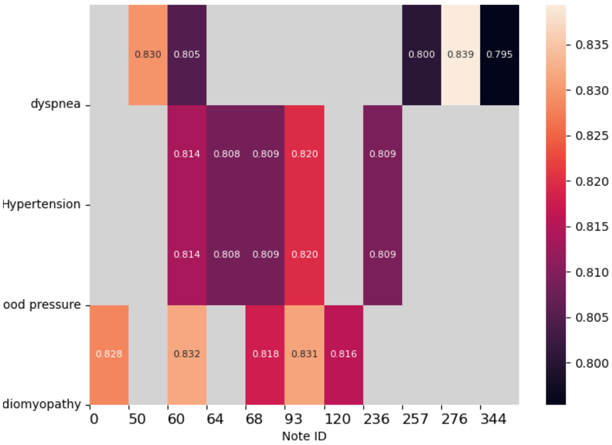


Figure 7: Using concept and clinical narrative (Setting 4), the figure displays the top retrieved note IDs on the horizontal axis and queries on the vertical axis. Lighter cells indicate higher relevance to the query

Query	**F1**	**Precision**	**Recall**
Dyspnea	0.656	0.635	0.678
Hypertension	0.685	0.642	0.734
Cardiomyopathy	0.690	0.648	0.738
High-blood pressure	0.670	0.623	0.724
Performance	0.675	0.634	0.719

Table 1: BERTscore over the top retrieved note with pre-processing and text cleaning

Approach	**F1**	**Precision**	**Recall**
Setting 1	0.630	0.590	0.677
Setting 2	0.630	0.582	0.688
Setting 3	0.673	0.610	0.751
Setting 4	0.699	0.648	0.770

Table 2: BERTscore system evaluation for 4 settings for the top 20 concepts in Appendix A.1 MIMIC-III

3.2 Evaluating ClinicNarrIR

An evaluation is conducted on the proposed model that was built on top of BlueBERT using a set of 1000 clinical notes that were randomly sampled and not part of the training data. The selection of these notes ensured a fair assessment of the model's generalisation capabilities. The top 30 ICD codes from MIMIC III were chosen for the evaluation based on their popularity in the dataset, determined by the number of notes combined with a particular ICD.

The evaluation of the 'ClinicNarrIR' model, as depicted in Table 3, reveals promising results. The model achieves a notable accuracy of 85% in the top 10 retrieved results, demonstrating its ability to retrieve notes closely aligned with practitioner queries. Additionally, it scores 0.544 in NDCG@10, signifying reasonable proficiency in relevance and ranking within Electronic Health Records. The MAP@100 score of 0.463 suggests effectiveness in retrieving pertinent information, particularly within the top 100 documents.

These metrics collectively emphasise the model's effectiveness in returning and ranking relevant clinical information. These results are similar to those in the field of biomedical articles in their first round, which achieve higher scores in subsequent rounds. This difference can be attributed to the challenging nature of Electronic Health Records (EHR) as a context compared to other biomedical articles (Roberts et al., 2020).

In the context of clinical narrative retrieval, finding directly comparable evaluation datasets can be a challenge due to the specialised nature of the task. As a result, we sought to validate its performance in different clinical tasks, specifically in the MedNLI (Medical Natural Language Inference) task, which is further elaborated in the following section. In this context, the proposed model demonstrated ex-

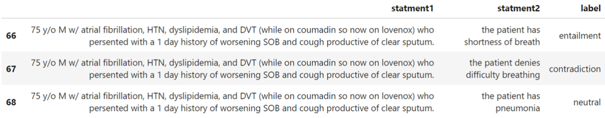
Metric	**Result**
Accuracy@10	0.850
NDCG@10	0.544
MAP@100	0.463

Table 3: ClinicNarrIR Evaluation Metrices

ceptional performance, achieving a score of 90.5%, the highest among all the approaches considered. The results are presented in Table 4, showing the competitiveness and effectiveness of the proposed model in addressing clinical natural language inference tasks.

3.3 Medical Natural Language Inference MedNLI

Clinical embeddings are evaluated through the **MedNLI dataset**, derived from the MIMIC-III database. This dataset, created by Romanov and Shivade (2018), pairs clinical notes with hypotheses and labels for three potential categories: entailment, contradiction, or neutral. In Figure 8, a snippet of the MedNLI dataset is presented, comprising 11,232 training samples, 1,395 development samples, and 1,422 testing samples.



	statment1	statment2	label
66	75 y/o M w/ atrial fibrillation, HTN, dyslipidemia, and DVT (while on coumadin so now on lovenox) who persented with a 1 day history of worsening SOB and cough productive of clear sputum.	the patient has shortness of breath	entailment
67	75 y/o M w/ atrial fibrillation, HTN, dyslipidemia, and DVT (while on coumadin so now on lovenox) who persented with a 1 day history of worsening SOB and cough productive of clear sputum.	the patient denies difficulty breathing	contradiction
68	75 y/o M w/ atrial fibrillation, HTN, dyslipidemia, and DVT (while on coumadin so now on lovenox) who persented with a 1 day history of worsening SOB and cough productive of clear sputum.	the patient has pneumonia	neutral

Figure 8: Sample of *MedNLI* dataset with three labels

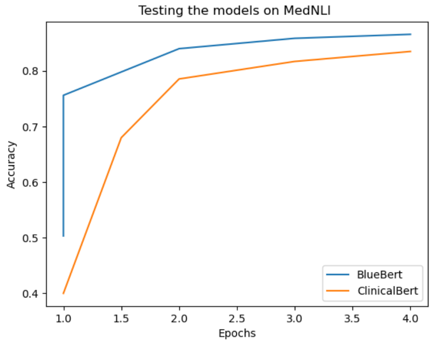


Figure 9: *MedNLI* accuracy over the four epochs for both BlueBERT and ClinicalBERT on the validation dataset

The objective of this work is to train a Siamese network(Chicco, 2021), using ClinicalBERT/BlueBERT. Incorporating a pooling layer to generate fixed-sized sentence embeddings, Siamese Neural Networks are designed to bring queries closer to relevant note vectors. During training, the network is optimized using the loss equation mentioned in Equation 4 on the MedNLI task. Different batch sizes $(16/32)$ were experimented with, and the results were evaluated over four epochs. The accuracy improvements with more epochs and a larger batch size are illustrated in Figure 9.

Approach	Accuracy
Our (ClinicalBERT)	83.4%
Our (BlueBert)	**86.66%**
Our (Large BlueBERT)*	**89.31%**
BLUE (large*)(Peng et al., 2019)	83.8%
BLUE (Base)(Peng et al., 2019)	84%
ClinicNarrIR	**90.5%**
Handcrafted Features[1]	52%
InferSent[1]	73.5%
LongTransformer(Li et al., 2022b)	84%

* with 24 layers
1 (Romanov and Shivade, 2018)

Table 4: Comparison of Approaches and Accuracy

In our evaluation, we utilized setting 4 to assess various pretrained models. Among the proposed models, namely 'Our (BlueBERT)' and 'Our (Large BlueBERT),' we observed impressive accuracy rates of 86.66% and 89.31%, respectively. Detailed results are provided in Table 4. These models outperformed ClinicalBERT and baseline models like BLUE, due to their extended training on both MIMIC III and PubMed articles. Remarkably, $ClinicalNarrIR$ achieved an impressive accuracy of 90.5%, showcasing its effectiveness in capturing clinical language nuances. These results underscore the potential of these approaches to enhance clinical information retrieval accuracy and their applicability across a wide range of clinical tasks.

3.4 Physician and Coder correlation

In the investigation to evaluate semantic similarity capturing within a dataset assessed by physicians and coders, we emulated the approach by (Huang et al., 2019). We measured the Pearson correlation between the physician ratings and the cosine similarity scores produced by the system. We achieved a moderate positive correlation ($r = 0.46$) with physician ratings, accompanied by a significant p-value of 0.0001, although slightly lower than the original study (see Table 5).

To enhance performance, 'Our(BlueBert)' was implemented in setting 4, yielding a robust correlation of 0.61 with physician ratings (p-value = 0.0003), surpassing ClinicalBERT in capturing semantic similarities. Additionally, the model 'Our fine-tuned on MedNLI,' as detailed in Section 3.3, achieved an impressive correlation coefficient of 0.72 with coder ratings (p-value = 6.91e-06). This remarkable improvement over the results reported

Table 5: Semantic Similarity Correlations with Medical practitioner Ratings

Model	**Correlation**		**p-value**	
	Physician	Coder	Physician	Coder
ClinicalBERT	0.46	0.51	0.0001	0.0036
Our (BlueBert)	0.61	0.65	0.0003	0.0002
Our fine-tuned on MedNLI	0.64	0.72	0.0003	6.91e-06

by Huang et al. (2019) highlights the model's effectiveness in enhancing semantic similarity tasks. This exploration aimed to assess the performance of diverse models in capturing semantics within the dataset created by (Pedersen et al., 2007) and evaluated by medical professionals.

4 Challenges & Limitation

In healthcare-based deep learning and information retrieval, several challenges emerge. Pretrained models demand high GPU memory, and there's a delicate balance to be struck between sequence length and batch size, essential for efficient training. Training embeddings for clinical notes that are associated with numerous ICD codes can result in suboptimal similarity between the note and the ICDs. This is because the training process may introduce conflicting rewards in one batch and penalties in another batch, potentially affecting the expected level of similarity. Inaccurate ICD code assignments by coders to a note can also contribute to this issue. Designing hard negatives sample for ICDs may offer a solution. The choice of evaluation metrics is critical, as they must align with specific research goals and be mindful of their limitations.

Substantial GPU memory, preferably over 32 GB, is required for effective work, which can pose challenges for those with limited access to high-end hardware. Training data may exhibit variations in query exposure, impacting result relevance, necessitating strategies for consistency.

Beyond technical concerns, researchers struggle with limitations on data availability imposed by patient privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). Additionally, they encounter challenges in overcoming issues of data scarcity, quality, and labour-intensive annotation processes. Navigating these complexities, which span technical, ethical, and data-related dimensions, is essential to advancing their work in this crucial field.

5 Conclusion

This research introduces a novel approach to enhance clinical information retrieval from Electronic Health Records (EHRs) by employing a context-enhanced network that leverages deep learning and concept embeddings. It demonstrates the effectiveness of various settings and the importance of integrating clinical narratives with concepts. Notably, the study highlights the model's practical application through a case study of a patient with pneumonia infection. The findings reveal that setting 4, incorporating concept embeddings, consistently outperformed other configurations. Moreover, retraining the embeddings using ICD significantly boosted accuracy to 90.5% in the MedNLI dataset. The *ClinicNarrIR* model achieved an $NDCG@10$ score of 0.54, showing some potential in retrieving clinically relevant information from EHRs, ensuring that the most critical patient data is readily accessible.

The analysis of Physician and Coder correlation demonstrated a strong correlation coefficient of 0.72 with coder ratings (p-value = 6.91e-06). Furthermore, the effectiveness of different pre-trained transformers was investigated, with BlueBERT consistently delivering superior results in all cases.

Leveraging MIMIC-III in this study proves advantageous to the research community, providing realistic data, benchmark status, open access, and ethical de-identification. This contribution establishes a valuable benchmark for fellow researchers.

However, this research acknowledges several challenges and limitations in healthcare-based deep learning and information retrieval, including the need for substantial GPU memory, data privacy regulations, and data scarcity. Despite these challenges, this research paves the way for more accurate and efficient clinical decision-making, benefiting both healthcare practitioners and patients. The comprehensive assessments and results presented herein offer valuable insights into the effectiveness of different settings in real-world clinical applications, reinforcing the significance of this work.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Richard P. Ames. 1991. Hyperlipidemia in hypertension: causes and prevention. *American Heart Journal*, 122(4, Part 2):1219–1224. Managing Cardiovascular Risk and Left Ventricular Hypertrophy New Prospects in Antihypertensive Therapy.

Robert J Carroll, Anne E Eyler, and Joshua C Denny. 2011. Naïve electronic health record phenotype identification for rheumatoid arthritis. In *AMIA annual symposium proceedings*, page 189. American Medical Informatics Association.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12.

D. Chicco. 2021. Siamese neural networks: An overview. *Methods Mol Biol*, 2190:73–94.

Spiros Denaxas, Pontus Stenetorp, Sebastian Riedel, Maria Pikoula, Richard Dobson, and Harry Hemingway. 2018. Application of clinical concept embeddings for heart failure prediction in uk ehr data. *arXiv preprint arXiv:1811.11005*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

Benjamin A Goldstein. 2020. Five analytic challenges in working with electronic health records data to support clinical trials with some solutions. *Clinical Trials*, 17(4):370–376.

Laura-Maria Peltonen Hanna von Gerich. 2022. Assessment of health service quality scoping review. *Challenges of Trustable AI and Added-Value on Health*, page 520.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Michael Z Huang, Candace J Gibson, and Amanda L Terry. 2018. Measuring electronic health record use in primary care: a scoping review. *Applied clinical informatics*, 9(01):015–033.

Xiaolei Huang, Franck Dernoncourt, and Mark Dredze. 2022. Enriching unsupervised user embedding via medical concepts. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 63–78. PMLR.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Jimeng Sun, and Zhiyong Lu. 2023. Matching patients to clinical trials with large language models. *arXiv preprint arXiv:2307.15051*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, and Angus Roberts. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Clemens Scott Kruse, Anna Stein, Heather Thomas, and Harman D Kaur. 2018. The use of electronic health records to support population health: A systematic review of the literature. *Journal of Medical Systems*, 42.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2022a. Neural natural language processing for unstructured data in electronic health records: a review.

Li Li, Wei-Yi Cheng, Benjamin S Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P Bottinger, and Joel T Dudley. 2015. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, 7(311):311ra174–311ra174.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022b. A comparative study of pretrained language models for long clinical text.

Journal of the American Medical Informatics Association, 30(2):340–347.

Ruoqi Liu, Lai Wei, and Ping Zhang. 2021. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nature machine intelligence*, 3(1):68–75.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Kimberly A Mc Cord and Lars G Hemkens. 2019. Using electronic health records for clinical trials: Where do we stand and where can we go? *Cmaj*, 191(5):E128–E133.

Sherri Mills. 2019. Electronic health records and use of clinical decision support. *Critical Care Nursing Clinics*, 31(2):125–131.

Serguei Pakhomov, Susan A Weston, Steven J Jacobsen, Christopher G Chute, Ryan Meverden, and Véronique L Roger. 2007. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care*, 13(6 Part 1):281–288.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673.

Sudha R. Raman, Lesley H. Curtis, Robert J. Temple, Tomas L. G. Andersson, Justin A. Ezekowitz, Ian Ford, Stefan James, Keith A. Marsolo, Parsa Mirhaji, Mitra Rocca, Russell L. Rothman, Barathi Sethuraman, Norman L. Stockbridge, Sharon F. Terry, Scott M. Wasserman, Eric D. Peterson, and Adrian F. Hernandez. 2018. Leveraging electronic health records for clinical research. *American Heart Journal*, 202:13–19.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Na Tang, Jian Ma, Rongqin Tao, Zhijun Chen, Yide Yang, Quanyuan He, Yuan Lv, Zelong Lan, and Junhua Zhou. 2022. The effects of the interaction between bmi and dyslipidemia on hypertension in adults. *Scientific Reports*, 12(1):927.

Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. 2015. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of biomedical informatics*, 54:96–105.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018a. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, and Sunghwan Sohn. 2018b. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart, and Richard JB Dobson. 2018. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*. *Journal of the American Medical Informatics Association*, 25(5):530–537.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Appendix

In this appendix we are listing popular 30 concept and popular queries that were used for evaluating of the model.

A.1 Popular Concepts

In Figure 1 top 30 patients' presentation (concepts) are extracted in the clinical narrative in MIMIC-III EHR. This can give insight about the type of complaints that the patient presented or was diagnosed with it.

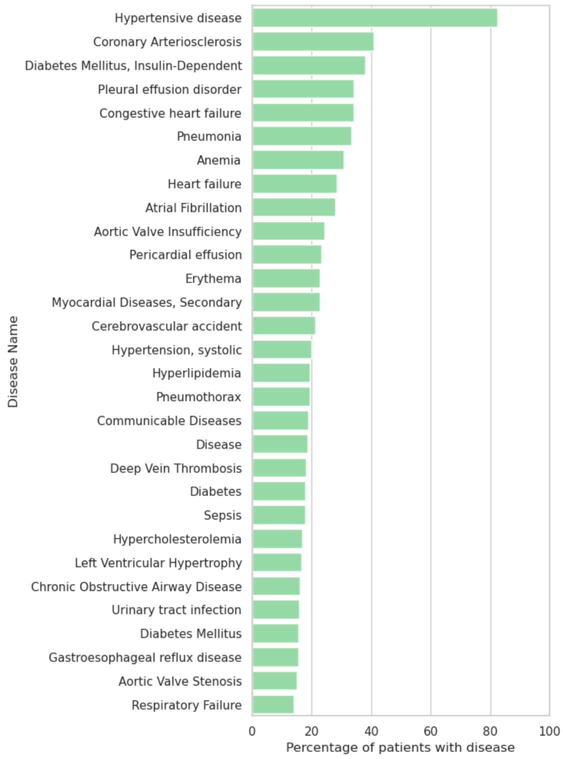


Figure 1: Top 30 patient presentation in EHR

A.2 Most frequent ICD (International Classification of Diseases) that existed MIMIC III

The popularity of the ICD codes in the MIMIC III dataset is determined by the number of notes in which they appear. We conducted an analysis of the dataset to study this popularity, and it follows a ranking from most popular to less popular, as shown in the list below:

1. Hypertension NOS
2. CHF NOS
3. Atrial fibrillation
4. Crnry athrscl natve vssl
5. Acute kidney failure NOS
6. DMII wo cmp nt st uncntr
7. Hyperlipidemia NEC/NOS
8. Acute respiratory failure
9. Urin tract infection NOS
10. Esophageal reflux
11. Pure hypercholesterolem
12. Pneumonia, organism NOS
13. Anemia NOS
14. Hypothyroidism NOS
15. Hy kid NOS w cr kid I-IV
16. Chr airway obstruct NEC
17. Acidosis
18. Chronic kidney dis NOS
19. Septicemia NOS
20. Severe sepsis
21. Food/vomit pneumonitis
22. Ac posthemorrhag anemia
23. Aortocoronary bypass
24. Hyp kid NOS w cr kid V
25. Long-term use anticoagul
26. Old myocardial infarct
27. Mitral valve disorder
28. Subendo infarct, initial
29. Depressive disorder NEC
30. Thrombocytopenia NOS