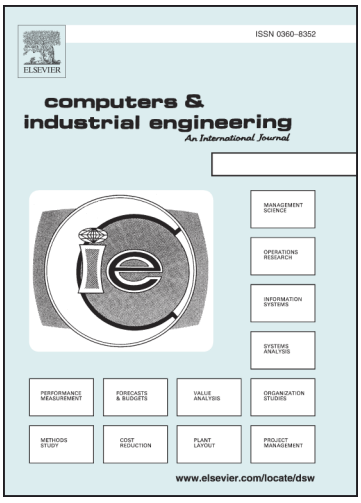
Journal Pre-proofs



EMR2vec: Bridging the Gap Between Patient Data and Clinical Trial

Houssein Dhayne, Rima Kilany, Rafiqul Haque, Yehia Taher

PII: S0360-8352(21)00140-6
DOI: https://doi.org/10.1016/j.cie.2021.107236
Reference: CAIE 107236

To appear in: *Computers & Industrial Engineering*

Received Date: 3 June 2020
Accepted Date: 8 March 2021

Please cite this article as: Dhayne, H., Kilany, R., Haque, R., Taher, Y., EMR2vec: Bridging the Gap Between Patient Data and Clinical Trial, *Computers & Industrial Engineering* (2021), doi: https://doi.org/10.1016/j.cie.2021.107236

EMR2vec: Bridging the Gap Between Patient Data andClinical Trial

Houssein Dhayne, Rima Kilany

Saint Joseph University, Mar Roukos, Beirut – Lebanon

Rafiqul Haque

Intelligencia, 66 Avenue des Champs Elysees, Paris – France

Yehia Taher

David lab, 45 Avenue des Etats Unis, Versailles - France

Highlights (for review)

EMR2vec: Bridging the Gap Between Patient Data and Clinical Trial

- We propose a novel methodology to represent EMR data in a vector space model with clinical trial dimensions.
- A respective geometric measurement is introduced to match a patient with a clinical trial.
- The power of combining machine learning with ontological reasoning techniques is investigated for representing patient profile in a vector space model.
- Dimensionality reduction technique is examined as an effective means of reducing feature vector size and eliminating noise.

Credit Author Statement

Houssein Dhayne: Methodology, Software, Data curation, Writing- Original draft preparation.
Rima Kilany: Supervision, Conceptualization, Project administration, Writing - Review & Editing
Rafiqul Haque: Conceptualization, Writing - Review & Editing.
Yehia Taher: Conceptualization, Writing - Review & Editing.

EMR2vec: Bridging the Gap Between Patient Data and Clinical Trial

Houssein Dhayne[1], Rima Kilany

Saint Joseph University, Mar Roukos, Beirut - Lebanon

Rafiqul Haque

Intelligencia, 66 Avenue des Champs Elysees, Paris - France

Yehia Taher

David lab, 45 Avenue des Etats Unis, Versailles - France

Abstract

The human suffering from diseases caused by life-threatening viruses such as SARS, Ebola, and COVID-19 motivated many of us to study and discover the best means to harness the potential of data integration to assist clinical researchers to curb these viruses. Integrating patients data with clinical trials data is enormously promising as it provides a comprehensive knowledge base that accelerates the clinical research response-ability to tackle emerging infectious disease outbreaks. This work introduces EMR2vec, a platform that customises advanced NLP, machine learning and semantic web techniques to link potential patients to suitable clinical trials. Linking these two different but complementary datasets allows clinicians and researchers to compare patients to clinical research opportunities or to automatically select patients for personalized clinical care. The platform derives a 'bag of medical terms' (BoMT) from eligibility criteria by normalizing extracted entities through SNOMED-CT ontology. With the usage of BoMT, an ontological reasoning method is proposed to represent EMR and clinical trials in a vector space model. The platform presents a matching process that reduces vector dimensionality using a neural

[1]Corresponding author(houssein.dhayne@net.usj.edu.lb)

network, then applies orthogonality projection to measure the similarity between vectors. Finally, the proposed EMR2vec platform is evaluated with an extendable prototype based on Big data tools.

Key words: EMR, Clinical Trial, Medical Data Integration, Neural Network, Semantic Web

2010 MSC: 00-01, 99-00

1. Introduction

The growing awareness of electronic medical records (EMRs) in diverse healthcare institutions has undoubtedly reshaped the way healthcare is delivered and health data is documented, enabling the collection of medical data from millions of patients. EMRs, including diagnosis codes, laboratory results and prescription data, are typically used for the systematic collection of patient health records in a digital format for the purpose of patient diagnosis and treatment. This collected electronic medical data holds great promise; EMRs not only contribute significantly to the provision of health care but can also be linked to datasets collected by other sectors to support a wide range of clinical research [**?**][**?**].

The evidence to support a clinical action or decision may be drawn from a wide variety of sources. Clinical trials are experiments or observations made in clinical research that help to determine the safety and efficacy of new medical treatments, turning them into evidence that can be applied in standard clinical care [**?**]. However, it would be naive to think that a clinical trial provides a definitive answer on its own since one trial is often limited to a selected subset of patients in atypically controlled circumstances and in insufficient numbers to assess the effects [**?**]. For instance, a systematic review of the clinical trials of COVID-19 up to February 9, 2020, resulted in 75 registered clinical trials. Only 11 trials have started recruiting patients, and all of the registered clinical trials are not completed, where most trials have a small sample of recruited patients [**?**].

During the last decade, the concept of evidence-based medicine (EBM) has aroused great interest because it integrates the best available evidence obtained by clinical research with the experience of the practitioner and expectations of the patient [**?**]. Additionally, personalized medicine typically involves a combination of diagnostic steps to provide a patient-specific profile and an actual treatment step. As well as, it is widely recognized that different patients respond differently to the same treatment. Therefore, integrating EMR and clinical trials could be used to apply the outcome of clinical trials into personalized recommendations by identifying patients for whom the benefits of treatment outweigh the harms, which can ultimately be used to enable more personalized clinical care.

Motivation. researchers need to take advantage of any data that is available. For instance, the advent of big EMR data offers an unprecedented opportunity to draw on the characteristics of real-world patients to guide and inform clinical research; this would require the linking and integration of big EMR data with clinical trial datasets. Integrated data could be extremely helpful in supporting investigators; it can provide a better understanding of actual patient populations, optimise the precision, recruitment feasibility and representation of eligibility criteria, and reduce the capture of redundant data. It can also assist in verifying the feasibility of clinical trials, evaluating the efficacy and results of treatment, and carrying out post-marketing surveillance and long-term monitoring [**?**]. Indeed, several studies have described the advantage of leveraging EMRs to improve trial recruitment [**?**].

In clinical trials, the eligibility criteria specify the characteristics of patients for whom a research protocol may be applicable. The criteria differ from one study to another; they can include age, gender, medical history and current health status. More than 74% of eligibility criteria could be evaluated using available structured data elements in the EMR [**?**], the most common categories are disease, symptom or sign (36%), therapy or surgery (13%), and medication (10%). When linking patients to clinical trials, it is helpful to match patient

medical information to eligibility criteria, allowing clinicians and researchers to compare patients to clinical research opportunities or to automatically select patients for personalised clinical care [**?**]. Consequently, there is a need to develop scalable integrated healthcare platforms to manage and link EMR datasets with clinical trials.

In this platform, the linking process identifies the eligibility criteria for each trial and then automatically determines eligible patients based on information from the EMR. Each created link is made up of the clinical trial identity, patient identity and a numeric value. This value represents the similarity score between the trial criteria and the patient's condition.

Challenges. there are significant challenges in linking EMR data to clinical trials, which, to our knowledge, have not all been systematically addressed [**?**]. (1) Eligibility criteria are usually described by a set of free text to be human readable; consequently, they are both syntactically and semantically complex. Computational processing requires the extraction and representation of semantics of the eligibility criteria in a machine-processable manner. (2) There is a semantic 'gap' between current expressions of clinical trial eligibility criteria and clinical data from the EMR; while eligibility criteria are described by coarser (more generic) clinical concepts or by defining their characteristics (attributes), EMR data is presented by granular (more specific) information. This discrepancy requires matching at the semantic concept level rather than verification of the absence or presence of a criterion at the lexical level. (3) Health data comes in many forms: vital signs, diagnosis, procedures, prescriptions and various types of medical reports. While the main forms are structured and could easily be analysed over time, medical reports must be analysed and interpreted using advanced natural language-processing tools.

Proposal. in our previous works, we proposed a semantic-driven engine to integrate structured and unstructured patient data in order to reformulate an entire patients medical record and query patient data across different data sources [**?**]. Furthermore, in another work, we proposed a framework for automated match-

ing of patients to clinical trials based on unstructured data from both datasets [**?**]. The proposed framework used BioBERT, a pre-trained biomedical language representation model, to match unstructured medical data. In this context, we found that vector representing methods such as word2vec, med2vec and BERT were very adequate at representing word and phrase as an embedding, but were not sufficient for representing complex objects such as patient or clinical trial, since these objects are usually composed of different types of information, which are structured and unstructured.

At the experimental level, the vector space model has proven to be an effective and robust framework for representing entities as vectors and querying about them [**?**]. Therefore, In this research, we explore the potential of using this model to match and link EMR data to clinical trials. With this model, we represent EMR and clinical trials objects as vectors of features values, where each feature corresponds to a dimension in the vector space model. Vector elements are generally represented by weights that describe the degree to which the corresponding feature describes the object (EMR or Clinical trial). In the vector space model, object vector representation plays a crucial role in many tasks, from objects matching and data clustering to similarity measuring [**?**].

We therefore present EMR2vec, a vector space platform in order to link two different but complementary datasets, patient data and clinical trial. To this end, we have customised and combined the advanced technologies of NLP, machine learning and the Semantic web and have derived a *bag of medical terms* (BoMT) from eligibility criteria. Utilising the BoMT, we proposed a method based on the vector space model to represent structured data from EMRs and unstructured eligibility criteria from clinical trials in order to develop an effective matching measure between patients and clinical trials.

Platform overview. fig. 1 shows an overview of the proposed EMR2vec platform. It includes five main stages: BoMT preparation, EMR vectorisation, clinical trial vectorisation, dimensionality reduction and data matching. Two datasets, Clinical Trial and EMR, represent the main entries for the platform.

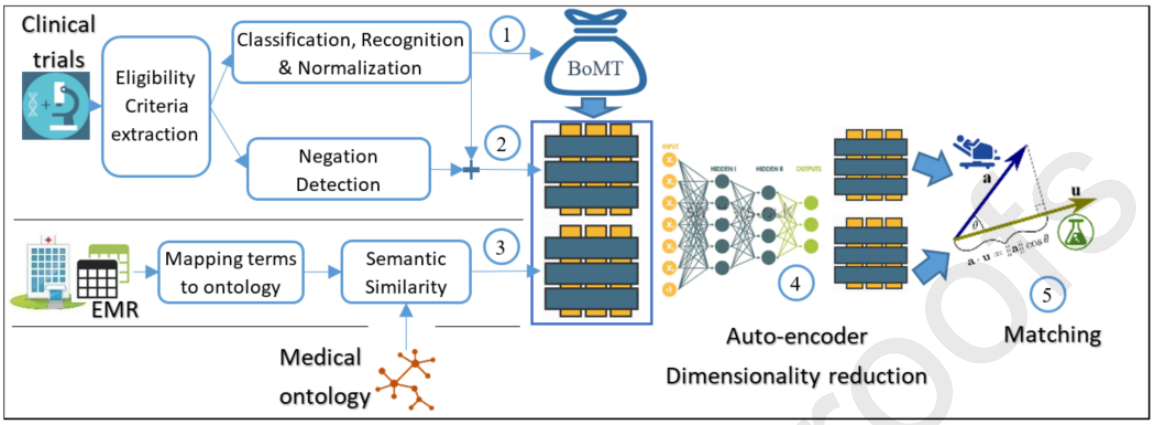


Figure 1: Overall architecture of EMR2vec platform.

1)The BoMT preparation stage consists of extracting medical terms (features) from a set of inclusion and exclusion criteria in order to construct the features of the BoMT. Different processes are applied to these criteria, including; classification, named entity recognition and normalisation. 2) In addition to processes that applied to construct BoMT, negation detection is performed to support representing clinical trials in vector space. 3) At the EMR and clinical trial vectorisation stages, BoMT features are used to represent data in a vector space. Medical ontology is used to convert EMR to vector by inferring relationships, as well as measuring the similarity between EMR terms and BoMT. 4) Since the high dimensionality space of BoMT is susceptible to noise, a dimensionality reduction technique is applied in stage four to reduce feature vector size and eliminate noise from the data. 5) At the data-matching stage, a projection similarity measure is used, whereby an orthogonal projection of the EMR vector onto the clinical trial vector, calculates a value showing the degree of matching between these two vectors.

Contributions. The main contributions of the paper are summarized as follows:

- We first defined a pipeline describing a novel methodology to extract the main SNOMED-CT terms from the criteria of targeted clinical trials.

- Regardless of specific hospital information systems, we described a novel methodology to transform EMR data into vectors in a vector space with dimensions of clinical trials.
- We investigated the power of combining machine learning with ontological reasoning techniques to match structured and unstructured medical data.
- To find links between EMR and clinical trial datasets, we systematically analysed their common medical characteristics, then introduced respective geometric measures to match a patient to a clinical trial.

The remainder of this paper is organized as follows. In Section 2, related background knowledge is covered. Section 3 deals with detailed descriptions of the BoMT preparation. 4 and 5 describe profiling vectors of Clinical trial and EMR, and how they were matched. The experimental evaluations are presented in Section 6, while Section 7 draws conclusions.

2. Background

There are many initiatives in literature aimed at providing solution for effectively matching patients to clinical trials; different approaches and tools are also offered a large body of literature such as [**?**][**?**]. The matching process was developed for accessing EMR data to find eligible patients in a clinical database for a clinical trial[**?**][**?**], or to help volunteers search for trials that may be appropriate for them[**?**][**?**]. Both cases focused on aligning the eligibility criteria with patient data to assess whether a given patient is eligible for a trial. An important challenge for matching clinical trials to patients is to merge the two cases so that clinicians can suggest alternative medications or interventions to the patient, as well as researchers, would have the ability to identify patients who meet eligibility characteristics of a trial and, later, to assess treatment effectiveness of trial outcomes within the stated context of use.

On the other hand, representing data in a vector space has been used in many domains [**?**], including healthcare. Various studies have attempted to convert

data stored in EMR systems into vectors. Deep patient [**?**] is a framework for deriving predictive patient description from electronic health records. The method represents patients by a set of general medical features through a deep learning approach. In order to predict the future hospitalization of a patient, a vector space framework is proposed to learn an interpretable representation for each patient [**?**]. The framework interprets the weight of features using a hierarchical attention mechanism. Authors in [**?**] reuse patients data of enrolled participants in under consideration clinical trial to discover new eligible patients, this work represents patients with four vectors by applying a simple training model based on finding common concepts in each data type. Although in this paper we also use a vector space model, our work distinguishes itself from the cited references in the fact that it exploits both machine learning and semantic web technologies to represent EMR and clinical trial data.

2.1. Electronic Medical Records (EMR)

EMR is an application environment, replacing paper medical records, composed of patient information that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization.

The basic features and functions of EMRs include the following [**?**]:

- Manage patient information including patient problems, medications, allergies, notes, past medical history, and observation results (such as laboratory, radiology, and other testing results).
- Provide substantial benefits to healthcare practitioners such as physicians, clinic practices, to monitor and manage patient information.
- Guide workflow and manage patient-specific care plans, provide appropriate guidelines and protocols, and support clinical decision-making.
- Provide a 360° view of the patient's condition at all times.

2.2. Clinical Trial

A clinical trial is a type of research that provides a longstanding foundation in the practice of medicine and the evaluation of new medical treatments. Each trial has eligibility criteria describing the characteristics according to which a patient or participant must meet all "inclusion criteria" and none of the "exclusion criteria". In this respect, the criteria differ from study to study. The authors in [**?**] analysed 1000 eligibility criteria and showed that 23% of the criteria are simple, or can be reduced to simple criteria, and that 77% of the criteria remain complex to evaluate. Therefore, a formally computable representation of eligibility criteria would require natural language processing techniques as part of various research functions in the era of EMR including evaluating feasibility, cohort identification and trial recruitment.

2.3. Identifying a Cohort of Patients for a Trial

EMR could help physicians to find patients who meet the criteria by searching in the patient information. Extraction of structured information from EMR is performed via traditional database queries to include diagnosis, procedures, laboratory results and medication orders [**?**].

However, matching eligibility criteria with EMR is still a manually performed task, making it a time-consuming and labor-intensive process and could lead to delays in delivering new therapies to the public. For instance, despite the heightened effort made by researchers, it is estimated that less than 5% of all adult cancer patients enter clinical trials [**? ?**] and potentially eligible patients are often overlooked in the manual process [**?**]. This is because the volume and complexity of patient information make it difficult for research staff to conduct a complete and adequate assessment of each patient.

2.4. SNOMED-CT

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [**?**] is a standardized, multilingual ontology of clinical terminology that includes terms of all medical domains and provides the general core terminology for the

EMR. SNOMED-CT has been developed over the past 30 years in a multinational effort and accepted as the global common language for health terms. It is a comprehensive international clinical terminology that is used in over fifty countries. With more than 349,548 unique biomedical terms(concepts) and 1.2 million synonyms grouped into 19 top-level concepts, SNOMED has very good clinical conceptual coverage. The concepts in SNOMED-CT are divided into hierarchies as diverse as body structure, clinical findings, geographic location, and pharmaceutical/biological product [**?**]. The core component types in SNOMED-CT are:

- Concepts that represent clinical meanings organized in hierarchies.
- Descriptions that relate the appropriate human-readable terms to concepts.
- Relationships that link each concept to other related concepts.

3. Bag of Medical Terms (BoMT) Preparation

Matching EMR to a clinical trial or automatically screening a patient for clinical trial eligibility is the task of comparing the clinical features of the patient $fp_i(fpi \in emr)$ to the features extracted from the inclusion and exclusion of eligibility criteria (EC) of clinical trial (ct), $ft_j(ft_j \in ct)$. Where emr and ct are two sets representing clinical features of the patient's EMR and clinical trial EC, respectively. From a practical standpoint, in order to be able to compare emr and ct, the terms in these two datasets should be normalized using the same medical standard or ontology (such as $SNOMED\text{-}CT$).

In this work, we will represent patients and clinical trials in the vector space model, by converting them to features vector with dimensionality equal to the number of different medical terms extracted from all eligibility criteria.

In vector space model, the features characterize the vector instance. Each feature of the vector is assigned a weight, which captures the relative importance of the feature in the vector. Thus, to represent these features, entities extracted

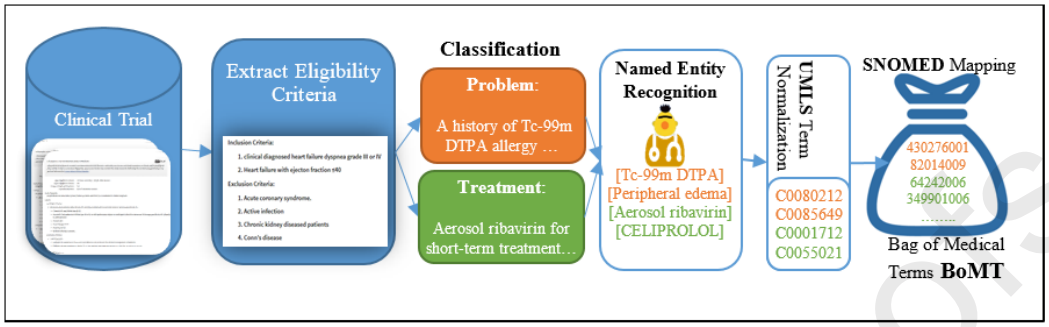


Figure 2: Pipeline to construct features of BoMT.

from all ECs are grouped in one BoMT in which complex medical objects are characterized by feature vectors of medical terms. A BoMT is also referred to as a bag-of-words (BoW) [**?**] in document classification, a bag-of-frames in audio and speech recognition [**?**], and a bag-of-instances in multiple instance learning [**?**]. The BoMT that consists of medical ontology terms(SNOMED-CT), serves as the features to represent the dimensions of EMR and Clinical Trial vectors. The features of each patient is profiled by extracting events from EMR entities such as Diagnosis, Procedures and Prescriptions and mapping them to SNOMED-CT. Each of them is profiled by entities extracted from their ECs with respect to clinical trials.

Fig. 2 presents the pipeline for building the BoMT; the pipeline starts by identifying and classifying each criterion into Problem and Treatment classes. Next, every criterion is processed to extract various medical terms using Named Entity Recognition (NER) - a popular technique of NLP. Each extracted term is replaced with the corresponding normalized term from the SNOMED-CT ontology. The output is stored and will serve as features to represent the dimensions of EMR and clinical trial vectors.

3.1. Eligibility Criteria Classification

The objective of this work is to match clinical trial with data from EMR. Since the primary content of EMR is the medical and treatment history of patients in one practice and that ECs (such as demographic, race,...) will probably

be difficult to retrieve from traditional EMR, we focused in our process to select EC for problem and treatment categories only. As we will see in the next section, classifying sentences into these classes will help reduce ambiguity when identifying and categorizing named entities mentioned in ECs [**?**].

3.2. Medical Entity Extraction

In the previous section, the EC has been classified as a single entity, but given that a sentence from EC contains important medical concepts, we were interested in extracting the concepts (named entities) that express the main idea of the sentence. For instance, in the case of a sentence representing a problem, we aimed to detect the name of the disease. Consider the sentence "the patient suffers from severe cardiovascular diseases", the goal was to detect the terms "cardiovascular disease". To that end, we used medical NER which is an important task of natural language processing (NLP). It enables the detection of a medical entity (problem, treatment, ...) in a sentence. Several methods have been implemented to examine the performance of medical NER. Most of these methods are based on Conditional Random Fields (CRF) and supervised machine learning models which utilize both textual and contextual information.

3.3. Entities Normalization

Named entities recognized from EC have to be compared to EMR structure data represented by standard coding systems. Therefore, entities extracted from EC must be normalized according to a standard ontology for medical concepts. We choose to use SNOMED-CT ontology.

Term normalization was performed by querying MetaMap for an entity and retrieving the corresponding concept with its Concept Unique Identifier (CUI) from SNOMED-CT . As a result, for each clinical trial, we ended up with a list of EC presented by a classification type, a normalized term, and a CUI (fig.2). By linking ECs to a medical ontology, it becomes possible to match semantic terms between Clinical Trial and EMR using automated inference over medical concepts.

3.4. Data Description & Experiments

3.4.1. Classification

A total of 20000 ECs were randomly extracted from Clinical Trials. The majority of these ECs can be mapped either to the concept Problem or to the concept Treatment.

MetaMap [**?**] is a medical natural language processing tool able to map biomedical text to the Unified Medical Language System (UMLS). We used MetaMap to treat and annotate these ECs with concepts from UMLS and preserve sentences containing only these two concepts. UMLS is a compendium of many controlled vocabularies in the biomedical sciences, each concept is assigned one or more semantic types, which are linked to each other by semantic relationships.

To filter these ECs, we proposed to regroup some semantic types from UMLS into a larger concept [**?**]. The group Problem consists of the following semantic types: Disease or Syndrome, Finding, Neoplastic Process, Pathologic Function, Sign or Symptom, Laboratory or Test Result, Mental or Behavioral Dysfunction. The group Treatment consists of the following semantic types: Health Care Activity, Research Activity, Therapeutic or Preventive Procedure, Antibiotic, Hormone and Pharmacologic Substance. Fig. 3 illustrates the hierarchy of this grouping.

Each sentence of the input dataset was submitted to the MetaMap tool. From the generated results produced by MetaMap, we considered the semantic types and the mapping scores. It is worth noting that for the same sentence, MetaMap provides multiple semantic types with different scores. Once a semantic type matched to one of the semantic types mentioned above, the sentence was accepted in that group. If none of the semantic types provided by MetaMap was found in the above list, this sentence was discarded. If for a sentence, MetaMap provided multiple semantic types included in our list, the one with the highest score was considered eligible.

Eventually, we obtained an imbalanced dataset of 5000 sentences where the

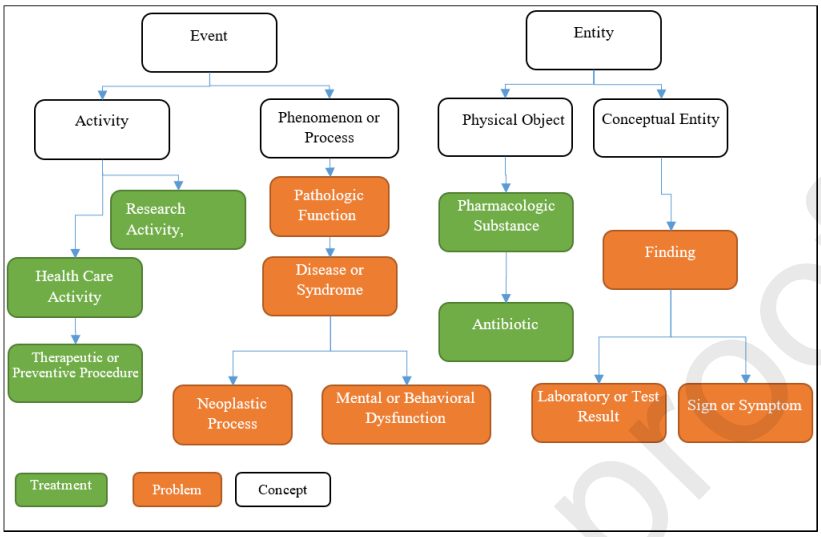


Figure 3: Selected UMLS semantic types clustered into Problem and Treatment concepts.

Problem group dominates in more than 4000 sentences. In order to account for this imbalance, we manually removed most of the Problem's sentences as well as the similar ones. The final EC dataset containing 1500 EC was prepared for verification and annotation in order to be manually labelled by a nurse and a data scientist according to the two classes: Problem and Treatment. The following criteria were applied:

- Problem class: includes the patient's complaints, symptoms, diseases, and diagnoses.
- Treatment class: includes medications, surgeries and other procedures.

In the case of multi-entities recognition in the sentence, the context of the sentence must be checked. For example, an ambiguous sentence, such as "Concurrent Medication: Allowed, Aerosol ribavirin for short-term treatment of RSV" (NCT00000961) discusses patient's medication, therefore it was labelled by Treatment despite the presence of a disease entity (Respiratory Syncytial Virus infection- RSV).

Sentences Classification

Once the EC were labelled, the next step was to train a classifier. For that purpose, we splitted the dataset into 80% samples for the training set and 20% for evaluation and testing. In our experiment, we explored and empirically compared five methods which are widely used in classification as the baseline of our classification: SVM, CNN, LSTM, C-LSTM, BioBERT.

- Word embedding: Word embedding is a set of low-dimensional continuous vectors that attempt to grasp the meaning of words and the context in their values. PubMed-and-PMC-w2v [**?**], is a 200-dimensional word embedding trained by an unsupervised language model using 23 million PubMed abstracts and 700,000 complete PubMed Central documents. The word embedding is initialized with Word2vec via gensim [**?**] and generates 4,087,446 vocabularies relevant to the biomedical domain. We used PubMed-and-PMC-w2v and the average word embedding to turn each EC into a vector representation form, which could be manipulated by machine learning algorithms [**?**].

- Support Vector Machine (SVM): SVM is a regulated machine learning algorithm that is widely used for classification challenges due to its high accuracy. SVM aims to create a hyperplane or set of hyperplanes to classify all inputs in a high-dimensional space. We took advantage of pre-trained PubMed-and-PMC-w2v to create a vector representation of EC, which will be the input of the SVM algorithm.

- Convolutional Neural Networks (CNN): Word embeddings and CNN have attracted considerable attention in various classification tasks. Recently, using word embeddings with CNN proved effective for various classification tasks, such as sentence classification [**?**]. CNN can be defined as a type of feed-forwarding neural network. Generally, a basic CNN includes the input layer, convolution layer, pooling layer, full connected layer, Softmax classification, and output layer.

- Long Short-Term Memory LSTM: LSTM is an effective type of recurrent neural network (RNN) architecture. The basic unit of an LSTM network is the memory block, which is able to learn long-term dependencies.

- C-LSTM: The authors in [**?**] combine CNN and LSTM architectures to create a hybrid C-LSTM model for text classification. In this method, CNN is used at the word level to extract a sequence of higher-level sentence representations and then the representations are used by an LSTM for text classification.

- BioBERT: BERT (Bidirectional Encoder Representations from Transformers) model applies the bidirectional training of Transformer to language modelling. The model is trained on BookCorpus [**?**] and English Wikipedia, which has in total more than 3,500M words. By adding a new simple layer to the main model, BERT can serve a wide variety of language tasks, such as (classification, Named entity recognition, Question Answering). BERT takes an input of tokens where the first token is always [CLS] which represents the special classification embedding. Taking the same architecture as BERT, Lee et al. [**?**] proposed the BioBERT language model trained by biomedical corpora. BioBERT model shows promising results of tasks in the biomedical domain. To adapt BioBERT for EC classification, a simple Softmax classifier is added to the top of BERT that takes the final hidden state of the first token [CLS] as the representation of the whole sequence then predicts the probability of a class.

Experiment

We measure the final experiment result using three different criteria: Recall, Precision and F1-score (eq.1), which are common criteria used for evaluating

Table 1: The performance of Eligibility criteria classification with different methods

	Precision	Recall	F1-score
SVM + W2v	0.86	0.86	0.86
CNN +w2v	0.87	0.88	0.87
LSTM	0.83	0.82	0.82
C-LSTM	0.82	0.82	0.82
BERT	0.92	0.91	0.91

the model performance.

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad F_1 = 2.\frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

Table 1 shows the attained classification results from 5 methods for classifying ECs in two classes. The results of our experiments indicated that the classification accuracy of BERT reaches more than 91% on the test data, which greatly outperformed other classifiers.

3.4.2. Medical Entity Extraction

We used the tagged corpus developed in 2012 i2b2(Informatics for Integrating Biology & the Bedside) challenge that includes concept extraction task. The i2b2 challenge focuses on extracting problems, treatments, clinical departments and occurrences in discharge summaries. While i2b2 dataset uses its own representation schema, most NER models use BIO annotation scheme, where 'B', 'I' and 'O' denote that a token/character is at the **b**eginning, m**i**ddle and **o**utside of an instance, respectively. Therefore, we implemented a code that parsed i2b2 2012 dataset into BIO representation schema. In our work, we focused only on entities about Problem(2,989 entities) and Treatment(2,269 entities), which support the mapping between clinical trial data and patient data.

Named Entity Recognition *(NER)*

In order to select the best neural model, we explored and empirically evaluated two models which are widely used in NER: BI-LSTM-CRF and CRF tagger on top of BioBert

- Conditional Random Fields (CRF): CRF provides a probabilistic framework for labeling and segmenting sequential data, based on a conditional probability approach, which showed several advantages over the Hidden Markov Model (HMM) and Support Vector Machine (SVM) [**?**]. CRFs are applied within a wide variety of domains, including natural language processing, such as Named Entity Recognition (NER)[**?**].

- Bidirectional LSTM-CRF: BI-LSTM-CRF Model for Sequence Tagging [**?**] focuses on the performance improvement after adding the CRF layer at the top of the word LSTM for named entity recognition task[**?**]. The strength of bidirectional LSTM in combination with CRF is that Bi-LSTM networks function as the encoder, while CRF is used to decode the entity labels to find optimal tag sequences.

- BioBert-CRF: BioBERT described in section (3.1) fosters new research pathways in sequential labeling. To adapt BioBERT for NER, a CRF layer is applied to the outputs of the BioBERT-based model. A public available implementation of BioBERT-CRF[2] was used to evaluate the architecture.

Experiment

Table 2 shows the experiments we made. The F1-scores depict that BERT-CRF performed order of magnitude better than Bi-LSTM-CRF. In other words, BERT-CRF outperforms Bi-LSTM and hence a sharp difference between these technologies is observed .

[2]https://github.com/dmis-lab/biobert

Table 2: The performance of NER with different methods

	Precision	Recall	F1-score
Bi-LSTM-CRF	0.675	0.665	0.665
BioBERT-CRF	0.94	0.92	0.93

3.4.3. Evaluation

As a conclusion, due to the small size of our datasets, it is evident that combining word and sentence embedding with any classification and NER algorithms produced more convincing results. We can state that neural network-based embedding is capable of effectively capturing semantic relation among words in order to cover medical terms which are not included in the training phase. As a result of the experiments explained above, BERT embedder who produces contextualized embedding has proven to be the most effective among other embeddings.

4. Vector Space Model for EMR & Clinical Trial

Profiling medical data is a very challenging task. In this section, we first introduce the detailed method for converting patient profile to vector space model using ontological reasoning. Subsequently, we propose an approach to represent clinical trial with the same space model.

4.1. BoMT Based Profiling of Patient and Clinical Trial

Different methods are proposed to profiling patient from EMR [**?**][**?**]. One often-used representation is the vector space model. In the vector space model, an instance (patient or clinical trial) could be represented as an n-dimensional vector, where each dimension corresponds to a specific medical event(feature) and n is the total number of features existing in these instances.

Illustrating EMR in BoMT vector space model requires computing the weights of the vector. In the case of document representation, Term Frequency-Inverse Document Frequency (TF-IDF) algorithm has been usually used with Bag Of

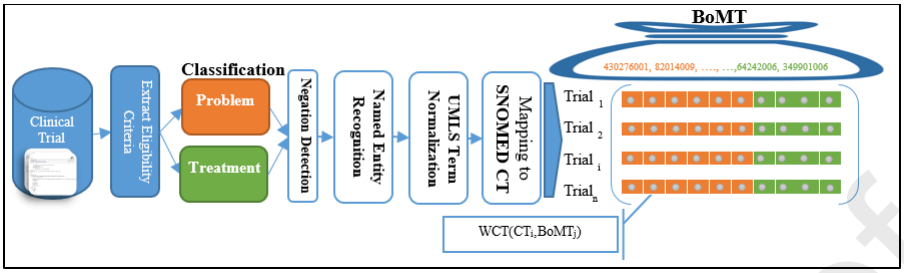


Figure 4: Pipeline to construct clinical trial vector. The input is a set of eligibility criteria from a clinical trial, the clinical trial vector is constructed by following the same steps to create the BoMT and add the negation step.

Word (BOW) representation model to calculate the weight of a document [**?**]. The main drawback is that this method ignores any semantic similarity between terms. In the following section, we detail our proposed similarity weight that operates on patient feature and BoMT.

4.2. Clinical Trial Profiling

As stated above, Clinical Trials have to be constructed as a profile in a vector space model with dimensionality equal to the number of terms in BoMT. As presented in fig. 4, the clinical trial vector could be constructed by carrying out the same steps that were done to create the BoMT and by adding the negation step.

Although the aim of this work is to study a group of a predetermined set of patients and clinical trials, a new clinical trial could be represented in this vector space model, provided that all medical terms extracted from the new clinical trial have to exist in the prepared BoMT. Therefore, a "global BoMT" could be created by considering all current eligibility criteria extracted from existing clinical trial datasets, which greatly ensures the possibility of accommodating any new trial.

4.2.1. Negation Identification

Many eligibility criteria contain negation, and this is an important aspect of the sentence given that it changes its meaning and therefore changes a patient

eligibility status. We used NegScope (Automatic biomedical negation scope detection algorithm)[**?**] to detect negation sentences. NegScope uses conditional random fields (CRFs), a supervised machine-learning algorithm, to detect negation cue phrases and their scope in biomedical literature. NegScope was built by training the model on the BioScope corpus. The model performed better than all baseline systems and NegEx, achieving an F1-score of 98% and 95% in clinical notes, and an F1-score of 97% and 85% in biological literature [**?**]. As an example, NegScope detects the negation *"No severe cardiac failure"* from the following EC *"uncontrolled or persistent hypercalcemia Cardiovascular: No severe cardiac failure"* (NCT00010088). In the equation, negated inclusion criteria detected by the negation method will be treated as Exclusion EC, and vice versa, for the negation Exclusion.

4.2.2. Vector Representation

Based on the section 3 that represents each EC in the form of SNOMED-CT terms, the clinical trial is represented as a vector of zeros except for the position corresponding to the SNOMED terms found in the eligibility criteria. The weight of features in Clinical Trial Vector is calculated as follows:

$$WCT_{ij}(CT) = WCT(CT_i, BoMT_j) =$$

$$\begin{cases} 1 \ if \ the \ term \ present \ in \ Inclusion \ EC \\ -1 \ if \ the \ term \ present \ in \ negation \ Inclusion \ EC \\ -1 \ if \ the \ term \ present \ in \ Exclusion \ EC \\ 1 \ if \ the \ term \ present \ in \ negation \ Exclusion \ EC \\ 0 \ otherwise \ (the \ term \ does \ not \ appear \ in \ any \ EC) \end{cases} \quad (2)$$

4.3. Patient Profiling

In order to represent a patient in the vector space model, patient information is collected from different admissions (visits) from the EMR, regardless of the admission timeline. As we detailed in Section 3.1, the primary content of EMR is the medical and treatment history of patients. Therefore, the target

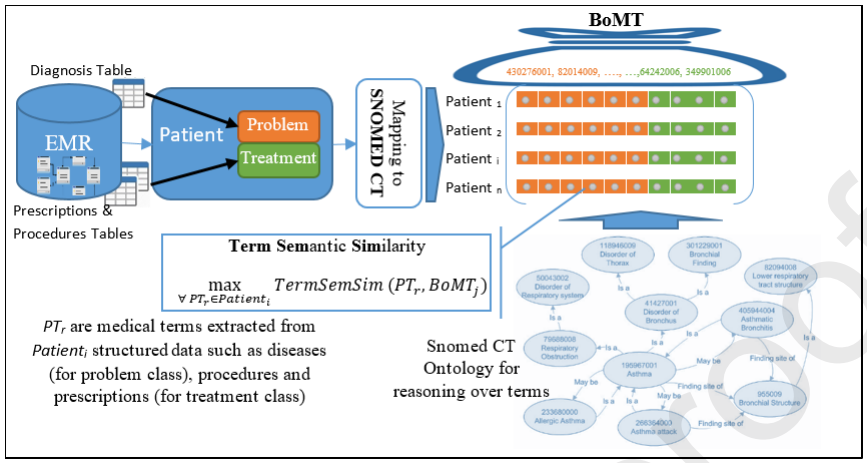


Figure 5: Pipeline to construct Patient Vector. Various EMR tables could be explored in this pipeline to extract patient profile. SNOMED-CT medical ontology was used to standardize medical terms as well as to measure the semantic similarity.

patient information is generated by creating two lists; problems and treatments, which combine terms extracted from EMR structured data such as diseases (for problem list), procedures and prescriptions (for treatment list). Other patient information (like age, gender) is not included in the EMR vector as it is easy to use to filter EMR database before starting the data linkage process.

In NLP, TF-IDF method ignores any semantic similarity between terms. Indeed, terms could be different enough to be considered as different features, although they can be semantically similar. For example "Cardiac Insufficiency" represents a feature of a patient in the EMR, while "Heart Failure" is an entity extracted from the EC "Clinical diagnosis of heart failure" (NCT03390088). TF-IDF considers these two terms to be different and, therefore, patients with "Cardiac Insufficiency" could not be mapped to any EC features. Thus, semantic similarity between terms yields better results for applications such as biomedical information retrieval. To determine each feature weight of a patient vector, unlike TF-IDF that reflects how important a feature is to a document in a corpus, we propose a similarity metric among clinical features(fig.5). The new Weight EMR metric approach ($WEMR$) calculates the maximum **Term**

Semantic **Sim**ilarity ($TermSemSim$) between the **P**atient **T**erms PT of patient P_i and BoMT features.

$$WEMR_{ij}(EMR) = WEMR(P_i, BoMT_j) = \max_{\forall TP_r \in Pi}(TermSemSim(PT_r, BoMT_j) \tag{3}$$

4.3.1. Ontology Reasoning for Term Semantic Similarity Metric

While EC is described by coarser (more generic) clinical concepts or by defining their characteristics (attributes), EMR data is presented by granular (more specific) information. To overcome this matching challenge, ($TermSemSim$) uses semantic reasoning by incorporating knowledge from SNOMED-CT ontology, such as "is a" and "has a-type" to semantically match both datasets. For example, using "is a" relation, the patient with the *"Cardiac arrhythmia"* condition will be mapped to the EC *"Underlying Heart disease"* (NCT02217267), as well as, a reasoning task could be performed to map the patient's condition "Benign tumor of lung parenchyma" to EC *"History or presence of any benign neoplasm considered by the investigator to be clinically significant"*(NCT01839279) using *"Has associated morphology"*.

4.3.2. Vector Representation

Based on the above, we define ($TermSemSim$) as an equation to calculate similarity between patient feature and BoMT by:

$TermSemSim(PT_r, BoMT_j) =$

$$\begin{cases} 1 \ if \ (PT_r \ is \ same\text{-}as \ to \ BoMTj) \\ 1 \ if \ (PT_r \ is \ is\text{-}a \ BoMTj) \\ 1 \ if \ (PT_r \ is \ is\text{-}a \ Term(T) \\ \quad and \ T \ \ has\text{-}characteristic \ BoMT_j) \\ \frac{2\times IC(LCS(c_1,c_2))}{IC(c_1)+IC(c_2)} [\textbf{?} \ ] \ if(\frac{2\times IC(LCS(c_1,c_2))}{IC(c_1)+IC(c_2)} \geq \xi) \\ 0 \ otherwise \end{cases} \tag{4}$$

Lin [**?**] defined the similarity measure between two concepts (c1, c2) as: dividing twice the Information content (IC) of the Least Common Subsumer (LCS) of concepts by sum of the individual IC of each concept. This measure is probably the best known and most widely used method to compute concepts similarity between terms in SNOMED-CT, as it showed a high correlation with expert judgments in empirical evaluations [**?**]. The threshold ($\xi \in [0, 1]$) is used as an acceptable weight similarity between two entities, and chosen based on what appeared to produce a reasonable valid matching.

5. Linking EMR to Clinical Trial

Measuring the patient-clinical trial matching score requires the use of similarity measures between vectors. More precisely, the more common features two vectors share, the larger the value of matching will be. However, the number of features is very large, therefore it is common to reduce the original dimensionality before measuring the patient-clinical trial similarity.

5.1. Dimensionality Reduction

By collating all patients vectors together, the EMR dataset is represented as a matrix $EMR_m \in R^{n*m}$, where n is the number of patients and m is the number of BoMT terms(features). This applies also to clinical trials which is represented by a matrix $CT_m \in R^{l*m}$, where l is the number of clinical trials.

BoMT represents terms extracted from all ECs of Clinical Trial, thereby EMR and CT matrices are high dimensional and very sparse due to the existence of a lot of possible medical concepts. The high dimensionality and sparsity cause considerable noise effects which affect the matching process. Dimensionality reduction techniques have been an effective means of automatically extracting latent concepts by removing noise and reducing the complexity in processing high dimensional data [**?**].

Among different approaches, such as principal component analysis (PCA), linear discriminant analysis (LDA) and Laplacian Eigenmaps, the autoencoder

approach is shown to outperform linear approaches for dimensionality reduction. In general, the autoencoder is a type of artificial neural network (ANN) where the input layer and the output layer have the same dimension and are connected with several hidden layers in between. The hidden layer with the minimum number of neurons is known as the bottleneck of the autoencoder and represents the dimension of the low-dimensional data [**?**].

In this work, we applied the autoencoder model to reduce the dimensionality of features and obtain a low-dimensional approximation that would extract the main features of BoMT and eliminate the noise of the data.

5.2. Matching Vectors Based on Projection Similarity

Let the following definitions be given, describing the content of the BoMT, EMR and clinical trial:

- $BoMT$ is represented with m features $f_1, f_2, ..., f_m, BoMT = (f_j | f_j$ extracted from criteria)
- An electronic medical record emr_i of a patient i with r features $fp_1, fp_2, ..., fp_r$ is represented as an m-dimensional vector, i.e. $P_i = (e_{i1}, e_{i2}, ..., e_{im} | e_{ij} = WEMR_{ij}(EMR) = WEMR(P_i, f_j))$.
- A clinical trial ct_i with l features $ft_1, ft_2, ..., ft_l$ is represented as an m-dimensional vector, i.e. $ct_i = (t_{i1}, t_{i2}, ..., t_{im} | t_{ij} = WCT_{ij}(CT) = WCT(ct_i, f_j))$

The criteria from the clinical trial are the ones that establish the matching process, since the presence of a feature in the trial requires researching it in the patient, and the absence of a feature in the trial should never affect the matching process. Therefore, considering a feature $f_j \in BoMT$, the following *"Matching rules"* should be respected when computing the matching score between an *emr* vector and a clinical trial *ct* vector:

1. The matching score should increase when f_j appears in both emr and inclusion criteria.

2. The matching score should decrease when f_j appears in both emr and the exclusion criteria.
3. The matching score should not to be affected when f_j appears neither in the inclusion nor in exclusion criteria.

The score calculation rules between emr and ct vectors mentioned above are incompatible with certain similarity measurement metrics such as Cosine, Euclidean and Jacquard distances. Consider, for example, two patients $P_1(1, 0.5, 1)$ and $P_2(1, 1, 1)$ and a clinical trial $CT_1(1, 0, 1)$. With Cosine, the similarity between P_1 and CT_1 is 0.9, which is greater than the similarity between P_2 and CT_1, 0.8. The same result was found using Euclidean distance measure. The distance between P_1 and CT_1 is 0.5, which is lower than the similarity between P_2 and CT_1, 1. The results show that these two measures do not satisfy rule 3.

Therefore, with Cosine and Euclidean similarity, even features that were not present in clinical trials and have values in EMR vector affect the similarity measure. To avoid these influencing features, we found that the calculation of a matching score should depend on how much of emr vector is pointing to the same direction as the ct vector. Our proposal to address this issue is to use the projection similarity [**?**]. Projection similarity provides a mean to compute the level of similarity of an EMR in the dimensions of the clinical trial features, which takes into consideration features required only in a clinical trial by an orthogonal projection of the EMR vector on the ct vector (fig.6).

The similarity projection of emr onto ct is the vector denoted $proj_{ct}emr$ which is represented by:

$$\mathcal{S} = proj_{ct}emr = \cos\theta|emr| = \frac{emr \cdot ct}{|ct|} = \frac{\sum_{i=1}^{m} e_i t_i}{\sum_{i=1}^{m} t_i^2} \tag{5}$$

Where "$\cdot$" denotes the dot-product of two vectors and θ the angle between emr and ct. Therefore, given a clinical trial ct and a patient document emr, the dot-product assigned to the pair (emr, ct) is a score in the interval [-1,1]. And thus, the longer the projection of emr vector on the ct vector, the higher

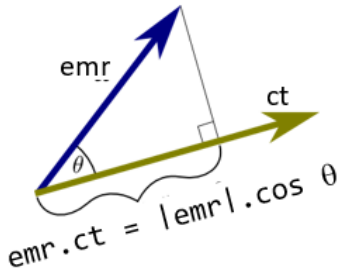


Figure 6: By orthogonally projecting the EMR vector onto the ct vector, the vector similarity takes into account only features required in a clinical trial.

the similarity score, the more likely is the patient eligible. The result of the computation process of each clinical trial is a list of patients sorted according to the matching score equation. Clinicians then select the top-ranking candidates who can be screened to identify eligible patients. Thus, the threshold is the score of the last candidate on this list.

The proposed matching derived from vector projection captures the proposed *matching rules*, which was not the case of the Cosine and Euclidean distance measures. Therefore, the proposed matching is considered the most appropriate.

6. Experimental Evaluation

6.1. Prototype Implementation

The growing amount of data in the healthcare industry has made it impossible to perform data integration using traditional tools. For instance, legacy data warehouses are unfit to handle data with high volume, high variety and high velocity. Thus, to ensure scalable load capacity, we have to adopt Big Data tools when developing an application for integrating and analyzing healthcare data [**?**].

To meet this challenge, we adopted a Data Lake infrastructure for developing our prototype. A Data Lake is a massively scalable storage technology which enables us to answer specific analytical questions, by simplifying the processing

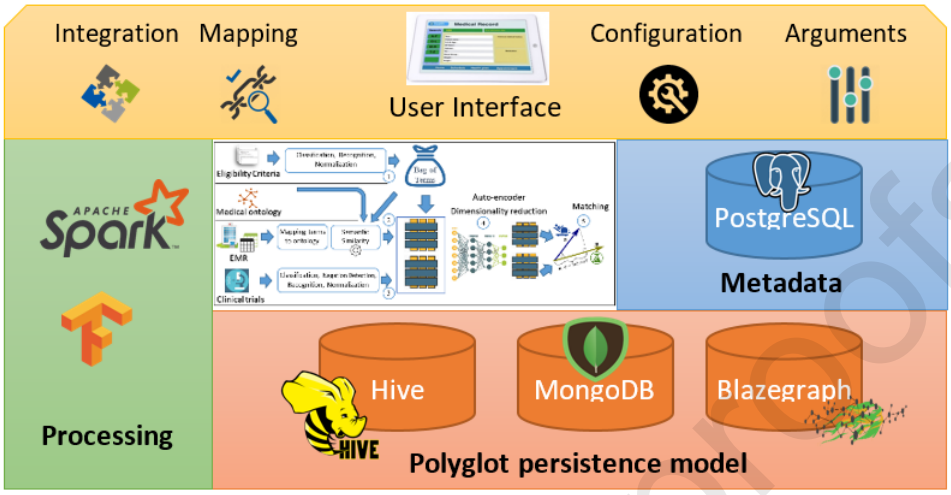


Figure 7: Overall architecture of the prototype.

of data variety using modern integrated tools [**?**]. In this context, the Data Lake concept realizes the polyglot persistence model by collecting data from huge heterogeneous datasources and providing an integrated view of the data without any predefined schema [**?**].

The data is stored in its raw original format and is processed whenever required for a particular analysis task to meet a specific request. We used respectively Hive, MongoDB and Blazegraph for stroring and accessing relational data, semi-structured data and RDF(Resource Description Framework) data. PostgreSQL is used as a metadata repository where we maintain the schema information and the similarity score results from datasets matching. An overview of our prototype is shown in fig. 7. The main function of the prototype is to allow a data analyst to integrate and query data from EMR and clinical trials datasets.

For instance, some researchers have successfully used EMRs as supportive tools to facilitate the assessment of clinical trial outcomes [**?**]. To illustrate our platform's work in an assessment outcome, let us consider the example where a researcher(user) is interested in assessing the effectiveness of a diabetes treatment among a number of trials. Therefore, as a first step, he needs to match

and link patients over 15 years old with "Family history of diabetes mellitus" (ICD9 code = v180) to several clinical trials related to testing drugs for treating "Diabetes Mellitus, Type 1". The process of creating links between patients and clinical trials is initiated by the user, who should express these two queries through the use of the platform's provided graphical interface. The role of these two queries is to select two subsets from EMR and clinical trial respectively. The first one filters patients from Hive and the second one filters clinical trial from MongoDB. In addition, the user defines the semantic similarity threshold ξ with a value between 0.5 and 1, as well as the different mapping files that provide the mapping between EMR coding systems (such as ICD9) and SNOMED-CT. These two queries are then executed by our plateform's pipeline which processes their result, produces the BoMT, as well as, creates representation vectors and matches patients to clinical trials. Finally, the pipeline stores the matching scores in the metadata of the data lake. Once the matching scores have been populated, the researcher can begin investing in the new linked datasets. Unlike many existing matching tools, the new linked datasets generated from the pipeline will allow researchers to ask ad-hoc analytical queries. Fig.8 illustrates the user interface for preparing data and configurations.

6.2. Datasets

To test our platform, we used two datasets; MIMIC-III (Medical Information Mart for Intensive Care) [**?**] comprising information relating to patients admitted to critical care units, and Clinical Trials [3] a publicly available dataset providing access to information on supported clinical studies.

6.2.1. MIMIC-III

It is a database comprising anonymized electronic health records of about 40000 patients admitted to critical care units. The dataset consists of a PostgreSQL database with 26 relational tables linked by identifiers: SUBJECT_ID

[3] https://clinicaltrials.gov/

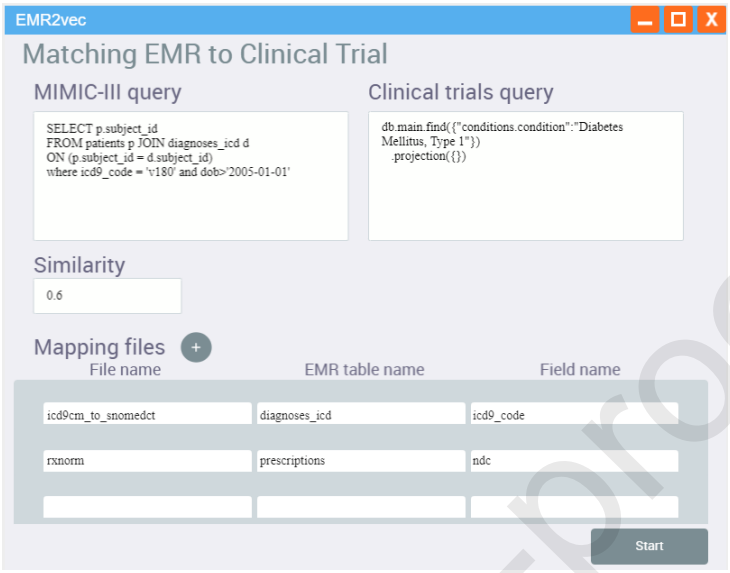


Figure 8: Prototype user interface for data and configuration preparation.

refers to a single patient and HADM_ID refers to a single admission. In our experiment, we used 3 tables of events (diagnosis, procedures, prescriptions) and imported data from PostgreSQL into Hive using Apache Sqoop.

Diseases and procedures are encoded using the International Classification of Diseases version 9 (ICD-9) codes, while prescriptions use various coding systems for drug representation, including Generic Sequence Number (GSN) and National Drug Code (NDC).

6.2.2. Clinical Trial

The ClinicalTrials.gov is the preferred resource for analyzing knowledge from clinical trials. After downloading the XML-encoded data from the official NCT website[4], we automatically processed and converted it into JSON format in order to import it and store it in MongoDB by using Apache NiFi.

The xml package contains all information about every study registered in ClinicalTrials.gov. However, it does not contain discrete eligibility features and

[4]https://clinicaltrials.gov/AllPublicXML.zip

therefore does not automatically support the required analysis of eligibility criteria. The following section describes how to prepare eligibility criteria for the processing step.

6.3. Preprocessing

The test was done according to the methodology described in the previous sections, using a random subset of 10,000 patients, paired with a random subset of 10,000 clinical trials.

To fully achieve system interoperability and prepare an infrastructure that supports the measurement of similarity between different classification systems, SNOMED-CT medical ontology was used to standardize medical terms, by linking the extracted terms to standard SNOMED-CT concepts. Resource Description Framework (RDF) version of SNOMED-CT is available for download from the NCBO BioPortal[5].

6.3.1. MIMIC-III

ICD-9-CM is the official system of assigning codes for surgical, diagnostic, and therapeutic procedures associated with hospital utilization. NDC is a universal product identifier for drugs. It is used for Prescription Drug Programs D.0 claims standards. RxNorm provides standardized names for clinical drugs and links their names to many of the drug vocabularies commonly used in pharmaceutical management. RxNorm had the best coverage of NDC codes [**?**].

In this work, SNOMED-CT represents the semantic bridge between various terminologies used in different datasources. Therefore, we needed to map ICD9 to SNOMED-CT using mapping files "ICD-9-CM Diagnostic Codes to SNOMED-CT Map" created by National Library of Medicine (NIH)[6].On the other hand, we also mapped NDC code to SNOMED-CT using RxNorm[7].

[5]https://bioportal.bioontology.org/ontologies/SNOMEDCT

[6]https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

[7]https://www.nlm.nih.gov/research/umls/rxnorm/index.html

6.3.2. Clinical Trial

The eligibility criteria are usually organized as free-text paragraphs or as bullet lists. Therefore, the classification of criteria into classes of treatment and problem requires the extraction of each criterion as an autonomous sentence. In this context, we have developed a module for analyzing and parsing the eligibility criteria paragraphs into a set of inclusion and exclusion criteria. Eligibility sentences derived from all clinical trials have been inserted into a new "inclusion-exclusion criteria" MongoDB collection, including the National Clinical Trial (NCT) number (a unique ID assigned by ClinicalTrials.gov) and criteria type. A csv file containing the above results can be downloaded from our GitHub repository[8]. Having this "inclusion-exclusion-criteria" collection, we can now flexibly apply the BoMT pipeline to each eligibility criteria in order to extract features across clinical trials.

After executing the stages of the BoMT pipeline, including classification, named entity recognition and normalization, 20034 distinct UMLS terms (features) were extracted from the 10000 selected clinical trials. Next, we mapped all of these features to SNOMED-CT using the mapping file provided by the National Library of Medicine (NIH)[9]

6.4. Evaluation Settings

Our long-term goal is to implement a scalable medical data integration platform that easily integrates into the Spark service and runs in a distributed fashion. BigDL is an open-source project distributed under the Apache 2.0 license. It is implemented as a distributed deep learning library on top of Apache Spark for Big Data platforms. BigDL exploits Spark's memory capabilities to cache the RDDs, containing the features and labels datasets, in the memory of each worker, allowing faster access during iterations [**?**]. For this reason, we

[8] https://github.com/housseindh/EMR-Clinical-Trial

[9] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

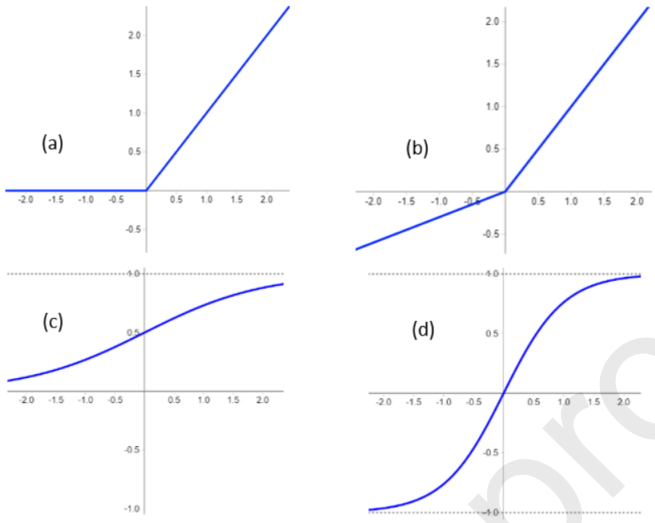


Figure 9: (a) ReLU function; (b) Leaky ReLU function; (c) Sigmoid function; (d) Tanh function.

used BigDL as our reference deep learning library to implement the autoencoder model.

For dimensionality reduction, autoencoder has a hidden layer with 512 dimension units. The input and output layers have exactly the same number of BoMT features with 20,034 units. For the choice of the activation functions, we tested some of the most popular choices in deep learning: "rectified linear units" (ReLU), Sigmoid, hyperbolic tangent (Tanh) and Leaky-ReLU [**?**]. Since the range of our input vectors is $[-1, 1]$, we found that, for this task, the combination of Leaky-ReLU for the hidden layer and Tanh for the output layer achieves better performance than ReLU and Sigmoid, due to:

- Leaky-ReLU assigns a non-zero slope for negative value, in contrast to ReLU in which the negative part is totally removed.
- Tanh function actually is a variant of the Sigmoid function but it maps a real input into the range $[-1, 1]$(see fig. 9).

6.5. Evaluation Metrics

We are interested in evaluating the performance of matching patients to a clinical trial, and therefore it is essential to measure the quality of top patient retrieval results. Thus, we evaluate using P@K based on the fact that the higher the proportion of true positives for a given patient at the top of a ranked list, the better is the platform performance. These techniques are widely used to evaluate text retrieval since the best search results need to be prioritized.

We are interested in evaluating the performance of matching patients to a clinical trial, and therefore it is essential to measure the quality of top patient retrieval results.

Therefore, we have selected 6 clinical trials representing different diseases including Stroke, Osteoarthritis, Thyroid Cancer, Prostate Cancer, Breast Cancer and Obesity. In the clinical domain, physicians were usually used to validate or create gold standard datasets. The authors in [**?**] have shown that high-quality data could by validated with different levels of clinical expertise. So, in order to validate our work, the top 5 patients obtained by our matching process for the 6 clinical trials were manually evaluated by an expert nurse and a master data science student.

To evaluate our matching process, we conducted 3 validation experiments with three different values ($\xi = 0.5$, 0.6 and 0.7) for the Lin equation threshold (eq. 4). In each experiment, we have generated a ranked list of patients based on their relevance for each trial and we evaluated their position in the ranking list. More precisely, two metrics were taken into account in our experience: Precision at K (P@K K=1,2,3,4,5)[**?**] and mean Average Precision (mAP)[**?**], where higher values of scores indicate better performance.

Suppose we have R trials, for each clinical trial ct, p patients are retrieved. Let $Rel(i)$ be an indicator function equaling 1 if the ith retrieved patient is eligible and 0 otherwise. The precision at K (P@K) is the proportion of relevant

patients in the top-k set(Eq. 6).

$$P@K = \frac{\sum_{i=0}^{k} Rel(i)}{K} \tag{6}$$

The mAP score (Eq.8) for a set of clinical trials is the mean of the average precision scores AP (Eq.7) of these clinical trials.

$$AP = \frac{\sum_{k=1}^{p}(Rel(k) \times P@k)}{\sum_{k=1}^{p}(Rel(k))} \tag{7}$$

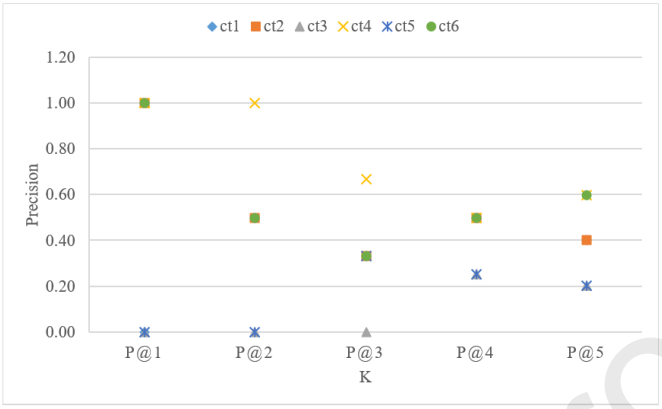
$$mAP = \frac{\sum_{k=1}^{R}(AP_q)}{R} \tag{8}$$

6.6. Evaluation Results

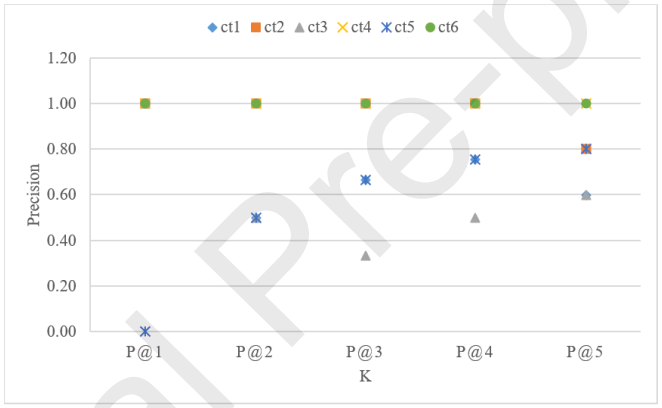
The results of P@K for the 3 experiment evaluations are presented in fig. 10, with the following (ξ) threshold values in equation 4: 0.5, 0.6 and 0.7. As we can observe, the distribution of points which represent the precision at k for the 6 clinical trials suggests that the threshold 0.6 might be better suited than 0.7 and 0.5. For threshold 0.6, the ratio of precision 1.0 is 63% against 20% and 14% for threshold 0.7 and 0.5 respectively.

Fig.11a shows the AP results of the 3 experiments. By investigating the result of different clinical trials, we found that all AP values of experiments increase and decrease in the same way according to clinical trial, this behaviour will be detailed in the discussion section. Furthermore, fig. 11b shows that the best overall mAP 0.86 was obtained using a relevance threshold of 0.6. Consequently, using 0.6 as a semantic similarity threshold leads to the best correspondence between the eligible patients and the clinical trial. Table 3 reports the AP and mAP values for each clinical trial and each threshold experiment.

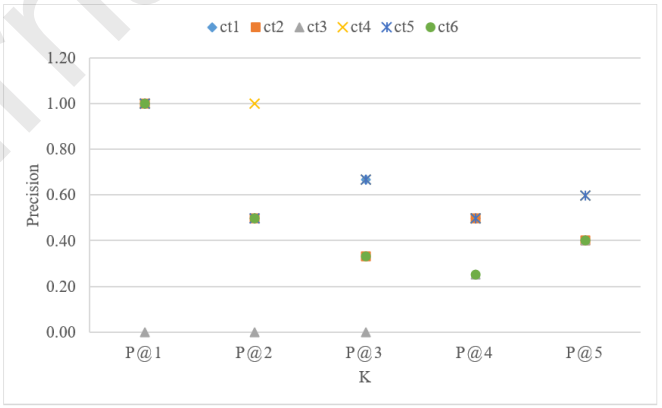
To empirically compare our results with the state of the art of clinical trial-patient matching, we explored the most relevant works. We were able to extensively compare with [**?**], which followed an architecture similar to ours. We found that our platform has the ability to move the relevant matching patients to the top of the ranking with an mAP = 0.86, compared to an mAP = 0.558 in



(a) Threshold 0.5

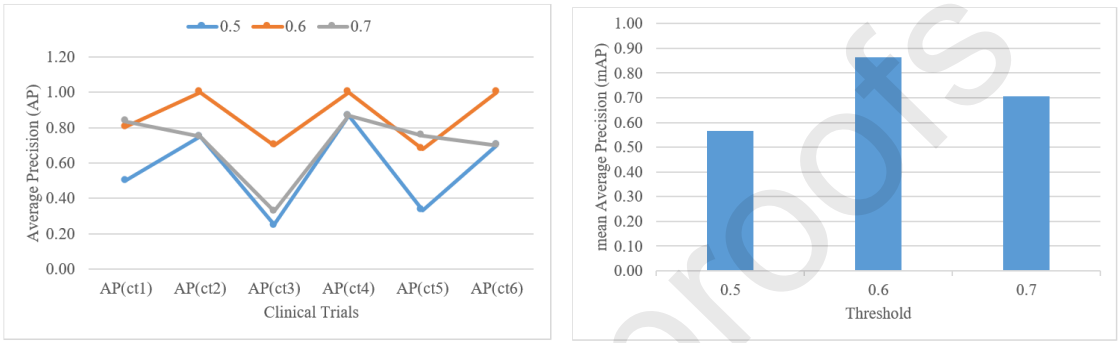


(b) Threshold 0.6



(c) Threshold 0.7

Figure 10: Precision-at-k (k=1 to 5) obtained for 6 clinical trials and 3 semantic similarity thresholds.



(a) Average Precision (AP)

(b) mean Average Precision (mAP)

Figure 11: AP and mAP experimental results from matching patients with clinical trials.

Table 3: Average Precision results (AP) in terms of 3 thresholds 0.5, 0.6 and 0.7, in addition, the mean Average Precision (mAP) over 6 clinical trials.

	$\xi = 0.5$	$\xi = 0.6$	$\xi = 0.7$
AP(ct1)	0.50	0.81	0.83
AP(ct2)	0.75	1.00	0.75
AP(ct3)	0.25	0.70	0.33
AP(ct4)	0.87	1.00	0.87
AP(ct5)	0.33	0.68	0.76
AP(ct6)	0.70	1.00	0.70
mAP	**0.57**	**0.86**	**0.71**

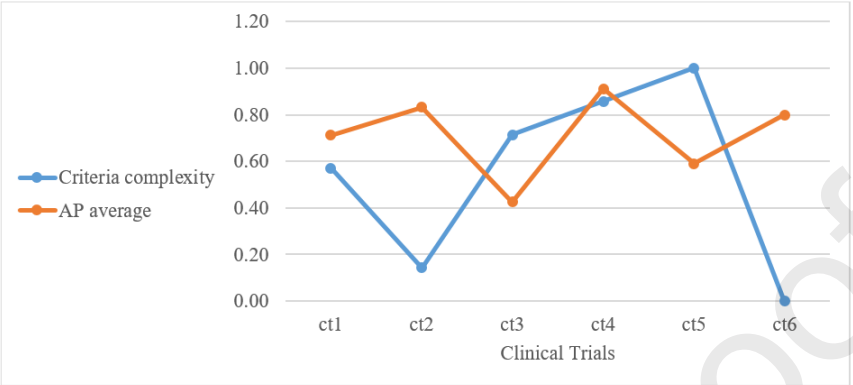


Figure 12: The AP values and the number of medical terms in the trial are inversely related.

[**?**]. Consequently, we can deduce that combining machine learning, semantic web and information extraction techniques, and more precisely, our proposed semantic similarity measurement which uses the SNOMED-CT medical ontology to construct the patient vector, has a very interesting impact in improving the matching performance.

6.7. Discussion

To further explore the insights of our EMR2vec, we investigated the reason why AP in the 3 evaluation experiments changed in a balanced way. We reviewed the eligibility criteria for clinical trials and quantified the number of medical terms in each trial. We found a relationship between the number of medical terms on the one hand, and the value of APs on the other. To corroborate this statement; (1) we normalized between 0 and 1 the number of extracted terms using eq.9, so as a higher value would indicate a more complex criteria, and (2) calculated the average AP of the 3 experiments. Fig. 12 demonstrates how the graph of AP values and the number of medical terms in the trial are inversely related.

$$normValue_i = \frac{NumberTerms_i - \min_{\forall i}}{\max_{\forall i} - \min_{\forall i}} \tag{9}$$

On the whole, the experimental result (mAP=0.86) of our proposed emr2vec platform demonstrates that our algorithm is able to efficiently link EMR and clinical trial datasets. It is now well accepted that vectors representation offer a reliable approach towards automatic matching patients and eligibility criteria. Moreover, automated inference under SNOMED-CT plays a critical role in semantic similarity; Whereas, SNOMED-CT was suitable for automatic reasoning of semantic relations between terms.

6.8. Future Work

Whereas structured data could be the main information representing patients, unstructured clinical text captured in the EMR, including discharge summaries, treatment plans, and progress notes, is certainly also useful for the matching process. Since the matching score should reflect both structured and unstructured data, we propose the integration of a new technique to match unstructured data from both datasets into the current pipeline in order to endow the platform complete control over all data types.

Another set of errors were caused by missing fields in the EMR such as laboratory test observations. Laboratory test results are typically based on a more advanced standard to be compared to the inclusion/exclusion criteria such as "AST/SGOT and ALT/SGPT $<$ 3.0 times upper limit of normal (ULN)". Logical Observation Identifiers Names and Codes (LOINC) is a database and universal standard for identifying medical laboratory observations. Unfortunately, a lack of adoption of LOINC codes in EMR and a lack of LOINC codes experts are major challenges when analszing laboratory data represented by text flags, and values with different units of measure, ranges, and so on. Therefore, to improve the performance of links of EMRs to clinical trials, we plan to extract entities of laboratory tests and their attributes such as observation value, units and result code.

7. Conclusion

In this paper, We presented EMR2vec, a Big data vector space platform for medical data linking. EMR2vec platform allows health researchers to match, link and query two different but complementary datasets, EMR data and clinical trial. To the best of our knowledge, this is the first study aimed at highlighting eligibility of patients for a trial using vector space model approach and combining machine learning and semantic web techniques. EMR2vec features three pipelines that are coupled together to support data matching; 1) BoMT creation, 2) patient data conversion into a vector and 3) clinical trial presentation as a vector. Our Matching process reduces vector dimensionality using neural network, then applies orthogonality projection to measure the similarity between vectors. In this work, we carefully analyzed the element types and data structure of both datasets; as well as we investigated how to handle the diversity of medical nomenclatures, vocabularies, coding and classification systems in order to support a smooth integration of health datasets. We verified the effectiveness of leveraging machine learning and semantic web techniques on EMRs and eligibility criteria. The potential of machine-learning emerged in converting unstructured data to queryable data, whereas semantic web provided data reasoning as well as interoperability between heterogeneous datasets.

We evaluated the performance of the proposed platform by carrying out several experiments. The outcome shows that the vector space model is a reliable approach for medical data matching tasks. More specifically, vector space provides efficient semantic representations of both datasets. To sum up, the proposed EMR2vec platform is a promising approach to bridge the gap between patient data and clinical trial. This can be very effective and efficient in treating patients suffering life-threatening diseases and hence can play a significant role in saving lives.

References