Using Embedding-based Metrics to expedite patients recruitment process for clinical trials

1st Houssein Dhayne
Faculty of Engineering, ESIB
Saint Joseph University
Beirut, Lebanon
houssein.dhayne@net.usj.edu.lb

2nd Rima Kilany
Faculty of Engineering, ESIB
Saint Joseph University
Beirut, Lebanon
rima.kilany@usj.edu.lb

***Abstract*—Despite the unprecedented volumes of Electronic Medical Records (EMRs) generated daily across healthcare facilities, the ability to leverage these data for patient participation in clinical trial remains overwhelmingly unfulfilled. The reason behind this is that matching patient information to the eligibility criteria for clinical trials is a manual, effort-consuming process. Therefore, automating this process is an essential step in improving the number of patients participating in clinical research. To address this issue, we propose a novel framework for automated patients to clinical trials matching. The matching process is based on measuring the similarity score between phrases extracted from patient medical records and the eligibility criterion for a trial.**

Our solution is based on a combination of NLP techniques and modern deep learning-based NLP models. In this context, we follow pre-training and transfer learning approaches to help the model learn task-specific reasoning skills. Additionally, we perform supervised fine-tuning on large Medical Natural Language Inference (MedNLI) and Semantic Textual Similarity (STS-B) datasets. The matching process was performed at semantic phrases level by converting patient information and trial criteria into vector representations. We then used a scoring function that combined cosine similarity and scaling normalization to identify potential patient-trial matches. The experimental results have shown that our framework is highly effective in sorting out patients by their similarity scores.

***Index Terms*—NLP, NLI, EMR, Automated clinical trial eligibility screening, BioBERT, Sentence similarity**

I. Introduction

The widespread adoption and use of electronic medical records (EMRs), together with the development of advanced artificial intelligence models, offer remarkable opportunities for improving the clinical research sector [1]. Furthermore, EMRs offer a wide range of potential uses in clinical trials such as facilitating the clinical trial feasibility assessment and patient recruitment, as well as obtaining main patient health information and medical history prior to their screening visit. The latter is a critical step in reducing the costs and duration of clinical trials [2]. Additionally, linking EMRs with clinical trials has been shown to increase patient recruitment rate [3]. However, there are many barriers to overcome in order to use EMRs for clinical trials.

Even though EMRs were designed to record information in a structured format, such as procedure information, diagnosis codes, drug prescriptions, and lab results, free text remains the most flexible way for physicians to express case nuances and clinical reasoning [4]. These free texts usually contain important facts about patients, but they are rarely available for formal queries [5].

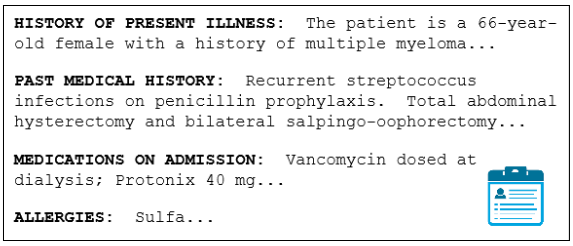
On the other hand, eligibility criteria for a clinical trial describes the characteristics of patients who are qualified to participate in the trial. Each criterion is usually expressed as a descriptive text and specified in the form of inclusion and exclusion criteria. Therefore, free text criteria can not always be transformed into structured data representations.

Authors in [6] confirmed that using only structured data from the EMR is insufficient in resolving eligibility criteria for patient recruitment in clinical trials, and that unstructured data is essential to resolve 59% to 77% of the trial criteria.

However, matching clinical notes with eligibility criteria is still a manually performed task, which makes it an expensive process in terms of time and effort. This slows down clinical trials and may delay new drugs from benefiting patients. As a consequence, it might entail the loss of human lives that otherwise would have been able to benefit from new medication. For these reasons, automated matching of clinical notes with eligibility criteria in the eligibility screening workflow would help overcome the bottlenecks of pre-screening practices in a trial setting.

To tackle the above challenge efficiently, we need to execute a matching process at a semantic sentence level, rather than by just checking for the presence or absence of a lexical criterion. The investigation of the potential use of modern deep learning-based NLP(Natural Language Processing) models, led us to propose a framework that would automate the evaluation of the eligibility of patients to be candidates for a relevant clinical trial. As a first step, the framework splits patient clinical report and clinical trial sentences into comparatively basic phrase units. Secondly, it classifies the phrases into various clinical categories (diagnosis, drug, procedure, observation). Thirdly, the framework converts candidate phrases into vector representations using an appropriate deep learning-based NLP model. Finally, it calculates a semantic matching score between patients and a clinical trial by using a combination of cosine similarity alongside a scaling normalization method.

This paper is organized as follows: In section II, we expose the problem definition and review the related works. In sec-



HISTORY OF PRESENT ILLNESS: The patient is a 66-year-old female with a history of multiple myeloma...

PAST MEDICAL HISTORY: Recurrent streptococcus infections on penicillin prophylaxis. Total abdominal hysterectomy and bilateral salpingo-oophorectomy...

MEDICATIONS ON ADMISSION: Vancomycin dosed at dialysis; Protonix 40 mg...

ALLERGIES: Sulfa...

Fig. 1. An example of discharge summary contents and format.

tion III, we describe our framework and illustrate the different challenges. The evaluation of the results and outcomes is discussed in section IV. Finally, we conclude this paper in section V.

II. Background

A. Problem definition

According to our approach, the problem definition of patient-trial matching can be described as follows:

Finding clinical trial participants is the task of matching Patient $P_i(P_i \in EMR)$ represented by a **D**ischarge **S**ummary DS_i to a **C**linical **T**rial CT represented by an **E**ligibility **C**riteria EC. Formally, the solution to this task is to find the top-K highest-values of function $\mathcal{M}$ which computes the matching score denoted by:

$\mathcal{M}(P_i, CT) = v$ which represents the score of matching patient P_i to a CT.

This list of the top-K highest-scores reduces the overall number of patients that will need to be screened by clinicians in order to identify eligible patients.

B. Data representation

1) 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. Authors in [7] 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 automated screening for patient eligibility.

2) Patient medical records: An EMR typically collects various types of patient information, including patient discharge summaries, prior diagnoses, radiology reports, medication history, and so on. Hospital discharge summaries are a physician-authored synopsis of a patient's hospital stay, which serve as the main documents communicating a patients care plan to the post-hospital care team [8]. Discharge summaries are organized in several sections. These sections usually include past medical history and history of present illness as shown in fig.1.

C. Related work

In the recent past, several projects have developed tools and technologies for automated trial-patient matching. Milian et al. [9] used a template-based formalism to extract and represent the semantics of the trial criteria in order to improve their comparability. Patel et al. [10] formulated the matching process as a semantic retrieval problem by expressing clinical trial criterion in the form of semantic query, which a reasoner can then use with a formal medical ontology - SNOMED CT to retrieve eligible patients. Other works such as EliIE [11] and Criteria2Query [12] have focused on identifying standardized medical entities in eligibility criteria using machine learning approaches, the extracted entities being then used to query patient data. Shivade et al. [13] constructed an annotated dataset that determined whether the medical note contains text that meets a criterion or not. Then, they implemented two lexical methods and two semantic methods to determine a relevance score of each sentence with a criterion statement, and found that semantic methods gave better results than lexical methods. Ni et al [14] evaluated a system using a combination of NLP, information retrieval and machine learning methods to identify a cohort of patients for clinical trial eligibility pre-screening. Their system relies on both structured data and clinical notes from EMRs.

III. Framework overview

In this section, we describe the framework we propose for automating the matching process between patients and a clinical trial. This framework takes into account the following different challenges; (i) In order to treat complex sentences in patient's data as well as in clinical trials, we break down paragraphs into sentences and complex sentences are then parsed into phrases. These phrases are the basic units for matching. (ii) To avoid costly comparisons without fault dismissals, phrases are partitioned using classification methods, which limits the number of pairs to match. (iii) To match phrases, we represent them in the form of distributed vectors, which enables calculating similarity for formally different but semantically related phrases. Fig. 2 shows an overview of our Patients to Clinical Trial matching framework. Given a Clinical Trial CT and set of Patients P, our task is to calculate a Matching score $\mathcal{M}(P_i, CT)$.

A. Paragraph and sentence decomposition

In order to measure the similarity between two sentences, we have to deal with a simple sentence representing a linguistically-meaningful unit. This process requires segmenting both paragraph-level and sentence-level structures into phrase-level structures. According to [15], segmentation of paragraphs and sentences is the process of parsing the longer processing units, consisting of one or more words, to further processing stages such as part-of-speech parsers, morphological analyzers, etc.

In our model, we handle each phrase as a primitive semantic unit and find matching phrases between patient and clinical

TABLE I
EXAMPLE OF SENTENCES SEGMENTATION INTO PHRASES

Paragraph	**Phrases**
Eligibility Crieteria NCT03484780	
Previous open laparotomy or contraindications to laparoscopy, as determined by implanting physician.	*1- Previous open laparotomy*
	2- contraindications to laparoscopy
	3- determined by implanting physician
Discharge Summary	
History of paroxysmal atrial fibrillation with anticoagulation in the past. History of coronary artery disease status post myocardial infarction	*1- History of paroxysmal atrial fibrillation*
	2- with anticoagulation in the past.
	3- History of coronary artery disease
	4- status post myocardial infarction

trials by calculating the similarity of each phrase in the discharge summary to each phrase in Eligibility Criteria (EC).

We used paragraph and sentence segmentation of MetaMap [16]. MetaMap was provided by the National Library of Medicine (NLM) to map Medical Language Processor (MLP) text to the UMLS Metathesaurus concepts [17]. MetaMap breaks text into paragraphs, sentences, and then phrases. Table I presents a simple example of segmenting sentences into phrases. The first refers to the eligibility criteria (NCT03484780) and the second illustrates an example from a patient discharge summary.

B. Phrases classification

A discharge summary report contains information about different topics. Therefore, the large number of heterogeneous phrases extracted from the patient reports may affect the efficiency and effectiveness of pairwise phrase matching [18].

To minimize the number of required comparisons, we applied a filtering methodology. The latter aims to filter all the classes of phrases that do not correspond to a given class, which limits the number of pairs to match.

Data classification techniques could support achieving this filtering by separating phrases extracted from patient data and clinical trial into different medical categories. This classification filters-out non-matching pairs prior to verification, which increases the efficiency of phrases similarity matching with high precision and without sacrificing recall.

In our study, a total of 1500 eligibility criteria were extracted from a Clinical Trials database[1] and were manually labelled by a certified nurse and a data science master student according to four classes (diagnosis, drug, procedure, observation).

In this work, we have empirically explored and compared four methods widely used in classification as our baseline: SVM, CNN, LSTM, C-LSTM [19], in order to identify the ones with the best performance. For SVM and CNN models, we initialized word embeddings by the average of the word embedding over all words in the sentence via PubMed-and-PMC-w2v [20].

Our experiment indicates that CNN + w2v model has the best prediction performance in comparison to the other models selected in our exploration, with a Precision of 0.87, a Recall of 0.88, and a F1-score of 0.875. We therefore adopted CNN + PubMed-and-PMC-w2v to perform this classification task and were able to categorize the phrases into the four pre-mentioned categories.

C. Phrase vector representations

The purpose of this work is to allow the matching of patients data and clinical trials by comparing unstructured data from both datasets. Our claim is that by measuring the similarity of primitive semantic medical units (medical phrases) of a patient's Discharge Summary and Eligibility Criteria, we can generate a score value supporting the matching task.

There are plenty of measures of semantic similarity between sentences used in NLP. Unsupervised and supervised methods have been used to calculate the semantic similarity between two sentences in the biomedical domain [21]. Recently, a number of novel approaches have been proposed to address this problem by producing sentence vectors [22]. As an example, Neural sentence-embedding methods [23] have been shown to outperform traditional approaches, such as TF-IDF and word overlap based measures.

1) Universal sentence embeddings: The concept of universal sentence embeddings has grown in popularity as it leverages models trained on large text corpora. These pre-trained models can be used in a wide range of downstream tasks, such as providing versatile sentence-embedding models that convert sentences into vector representations. Notable works include ELMo [24], GPT [25], and BERT [26].

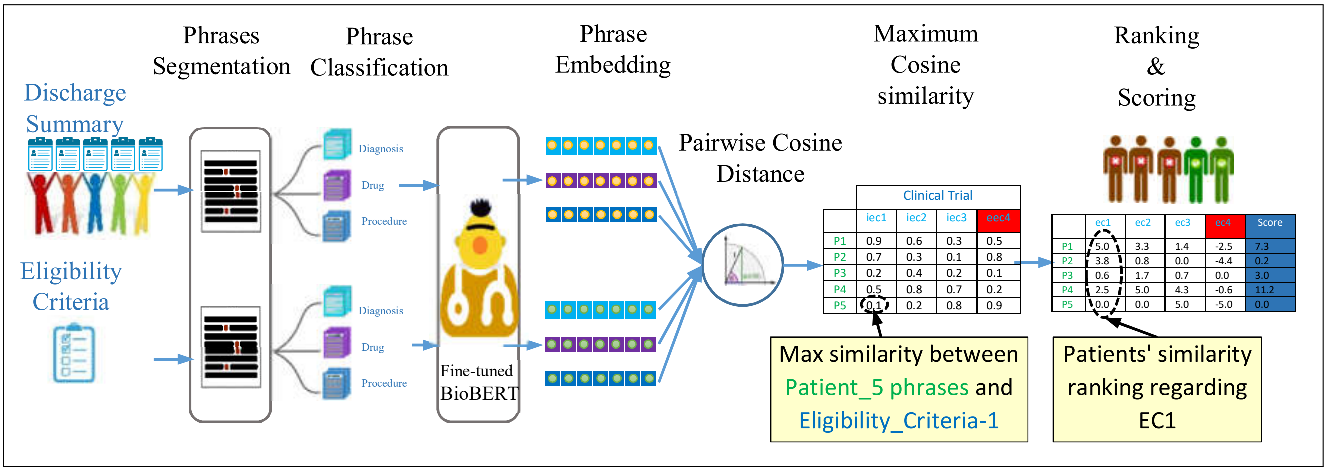
2) BioBERT: BERT (Bidirectional Encoder Representations from Transformers) is a neural network language model trained on plain text for masked word prediction and next sentence prediction tasks. BERT applies multi-layer bidirectional transformer encoder with self-attention. According to [27], BERT overall achieved state-of-the-art performances in many Natural Language Processing tasks and was significantly better than other models. However, compared against more recent models, XLNet [28] outperforms BERT and achieves better prediction metrics on the GLUE benchmark [29], but is not yet widely used in the medical field. Applying the same architecture as BERT, Lee et al. [30] proposed the BioBERT language model trained on biomedical corpora including PubMED and PMC. The BioBERT model showed promising results in the biomedical domain.

3) Phrase embedding: In this respect, to generate context-rich phrase embeddings, we chose BioBERT as the language model in conjunction with the Bert-as-service library [31]. Bert-as-service is a feature extraction service based on BERT which uses two strategies to derive a fixed-sized vector. In the default strategy, Bert-as-service does average pooling of all of the tokens of second-to-last hidden layer, while the second uses the output of the special CLS token and is recommended only after fine-tuning BERT on a downstream task.

D. Phrases Similarity Measures

The similarity between two vectors can be evaluated using various similarity measures such as Cosine similarity, Eu-

[1] https://clinicaltrials.gov/



	Clinical Trial			
	iec1	iec2	iec3	eec4
P1	0.9	0.6	0.3	0.5
P2	0.7	0.3	0.1	0.8
P3	0.2	0.4	0.2	0.1
P4	0.5	0.8	0.7	0.2
P5	0.1	0.2	0.8	0.9

	ec1	ec2	ec3	ec4	Score
P1	5.0	3.3	1.4	-2.5	7.3
P2	3.8	0.8	0.0	-4.4	0.2
P3	0.6	1.7	0.7	0.0	3.0
P4	2.5	5.0	4.3	-0.6	11.2
P5	0.0	0.0	5.0	-5.0	0.0

Max similarity between Patient_5 phrases and Eligibility_Criteria-1

Patients' similarity ranking regarding EC1

Fig. 2. Framework overview

clidean distance, and Manhattan distance. Since these similarity metrics are a linear space in which all dimensions are weighted equally, we perform here the similarity matching metrics of different phrases by ranking these phrases according to the cosine similarity. Therefore, the rank of similarity can be obtained by the equations presented in (1) and (2).

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} \tag{1}$$

$$\begin{gathered} if \quad \cos(\boldsymbol{A}, \boldsymbol{B}) > \cos(\boldsymbol{A}, \boldsymbol{C}) \\ then\ \boldsymbol{A}\ is\ more\ similar\ to\ \boldsymbol{B}\ than\ \boldsymbol{C}. \end{gathered} \tag{2}$$

Whereas a pre-trained BioBERT knowledge often shows a good performance for certain tasks, as we shall see later on, this prior knowledge is not sufficient to compute the similarity of sentences based on their embeddings. Indeed, we first tried to compute the cosine similarity of sentences, annotated by experts, using extracted embedding from pre-trained BioBert, without any fine-tuning. The result of the comparison was unsatisfactory and unacceptable (table II). The most significant sentence is the exact opposite, for example; the most similar sentence of ”*History of CVA*” was ”*patient has normal brain MRI*” with similarity value of 0.91 which was annotated by experts as ”contradiction”, and the ”Entailment” sentence ”*patient has history of stroke*” appears in the second place with similarity value of 0.89. Therefore, foregoing experiments reinforced our belief that it is necessary to fine-tune BioBERT on our downstream task.

1) Supervised Fine-tuning: Transfer learning is the process of extending a pre-trained model by leveraging data from an additional domain for a better model generalization [32]. The most common transfer learning techniques in NLP is fine-tuning. Fine-tuning involves copying the weights from a pre-trained network and tuning them using labeled data from the downstream tasks. BERT is a fine-tuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level tasks, with pre-trained representations reducing thus the need for many heavily-engineered task-specific architectures.

In the context of natural language understanding (NLU) technology, comparing the relationship between two sentences is based on several downstream tasks such as Natural Language Inference (NLI) and Semantic Textual Similarity (STS) [29]. Besides that, authors in [33] have shown that fine-tuning BERT on NLI and STS datasets creates sentence embeddings which achieve an improvement of 11.7 points compared to InferSent [34] and 5.5 points compared to the Universal Sentence Encoder [22]. In this context, we first fine-tuned BioBERT on STS-B dataset that generated our BioBERT-based model. We then further fine-tuned on MedNLI dataset. We used the fine-tuning classifier from BERT systems [35].

- MedNLI [36]: is a large, publicly available, expert annotated dataset drawn from the medical history section of MIMIC-III. MedNLI includes a set of clinical sentence pairs(14,049 pairs). They were annotated with one of three classes: entailment, contradiction, and neutral.
- STS-B [37]: is a collection of sentence pairs selected from news headlines. The dataset consists of paired sentences (8,628 pairs) labelled by humans with a similarity score of 1 to 5 denoting how similar the two sentences are in terms of semantic meaning.

2) Evaluation of fine-tuned BioBERT: We evaluated the new BioBERT model by computing the cosine similarity between the phrase embeddings. We observed that the model, was not just able to rank phrases in terms of similarity, but also gave a more appropriate cosine value. A representative sample of the results is depicted in Table II.

E. Matching Patients to Clinical Trials

After fine-tuning the BioBERT model for optimized cosine similarity and creating both Discharge Summary and Clinical Trial phrases embeddings, we proceeded to find Clinical Trial participants from an EMR dataset.

Formally, we denote:

TABLE II
NLI AND COS SIMILARITY BEFORE AND AFTER FINE-TUNING OF BIOBERT

		Experts	Pre-trained BioBERT		Fine-tuned BioBERT	
Phrase 1 (P1)	**Phrase 2 (P2)**	**NLI(P1, P2)**	**Cos(P1, P2)**	**Rank**	**Cos(P1, P2)**	**Rank**
History of CVA	patient has history of stroke	Entailment	0.89	1.53	0.87	**3.00**
	patient has normal brain mri	Contradiction	0.91	**3.00**	0.75	0.00
	patient is hemiplegic	Neutral	0.86	0.00	0.77	0.38
Per report ECG with initial qtc of 410 now 475, QRS 82 initially, now 86 rate= 95.	Patient has abnormal EKG findings.	Entailment	0.89	2.05	0.82	**3.00**
	Patient has normal EKG.	Contradiction	0.90	**3.00**	0.80	2.30
	Patient has angina.	Neutral	0.88	0.00	0.73	0.00
History of hypercholesterolemia and peptic ulcer disease s/p gastric bypass some years ago was involved in a low-speed MVC.	the patient was in a MVC.	Entailment	0.89	**3.00**	0.82	**3.00**
	the patient has no medical history.	Contradiction	0.88	2.48	0.53	0.00
	the patient has no significant injuries.	Neutral	0.86	0.00	0.67	1.51

- $DS_i = \{ph_{i,1}, ph_{i,2}, ..., ph_{i,r}\}$ as the phrases extracted from **D**ischarge **S**ummary of patient P_i.
- $IEC = \{iec_1, iec_2, ..., iec_p\}$ as the phrases extracted from **I**nclusion **E**ligibility **C**riteria.
- $EEC = \{eec_1, eec_2, ..., eec_q\}$ as the phrases extracted from **E**xclusion **E**ligibility **C**riteria.
- $EC = \{ec_1, ec_2, ..., ec_l\} = IEC \cup EEC \mid l = p + q$ as all phrases extracted from **E**ligibility **C**riteria.
- $\mathcal{S} \in [0,1]^{n*l}$ as the cosine **S**imilarity matrix, where n and l are the number of Patients and EC elements, respectively.

1) Matching Patient to Eligibility Criteria: Once phrases embedding are computed for the patients and the clinical trial eligibility criteria, we calculate the similarity between phrases of the same class (Diagnosis, Drug, Procedure,...) as defined in sub-section III-B. An element $s_{i,j}$ of S represents the similarity between patient criteria P_i and single eligibility criteria ec_j. The similarity function is defined by calculating the cosine between each phrase $ph_{i,r}$ extracted from DS_i and ec_j, then only the higher cosine value of similarity is retained for $s_{i,j}$ and all other values are discarded.

$$s_{i,j} = \max_{\forall ph_{i,r} \in DS_i} (\cos(ph_{i,r}, ec_j)) \qquad (3)$$
$$i \in [1, n] \,\&j \in [1, l]$$

Once the similarity values obtained, the final representation of $\mathcal{S}$ would be as follows:

$$\mathcal{S} = \begin{bmatrix} \max_{ph_{1,r}}(\cos(ph_{1,r}, ec_1)) & . \\ . & . \\ . & \max_{ph_{n,r}}(\cos(ph_{n,r}, ec_l)) \end{bmatrix}$$

2) Ranking and Scoring Patients: The semantic cosine similarity calculated in the previous paragraph enables a proportional similarity instead of exact text semantic matching. Therefore, when we compare similarity values obtained for different features (eligibility criteria) in the generated matrix $\mathcal{S}$, we notice that just because the value of similarity is higher, that does not mean that the similarity with the patient is greater. For example if $s_{x,1}$ and $s_{y,2}$ represent the highest value of the features ec_1 and ec_2, respectively, and if $s_{x,1} > s_{y,2}$, this does not mean that P_x has a phrase more similar to ec_1 than P_y for ec_2 (as a noticed in equation 2), but only means that P_x and P_y are ranked respectively at the top similar of the list for ec_1 and ec_2. The same logic applies for the lowest value, which represents the last order of similarity.

This variation in the similarity values between features requires a range normalization step to enable rank similarity instead of cosine similarity, which supports perfectly the computation of a matching score between patients and the Clinical Trial. To this end, we generated a new matrix $\mathcal{R}$ by applying the following feature scaling normalization:

$$r_{i,j} = \begin{cases} n \times \frac{s_{i,j} - \min_{\forall i}(s_{i,j})}{\max_{\forall i}(s_{i,j}) - \min_{\forall i}(s_{i,j})} & ; \quad ec_j \in IEC \\ (-n) \times \frac{s_{i,j} - \min_{\forall i}(s_{i,j})}{\max_{\forall i}(s_{i,j}) - \min_{\forall i}(s_{i,j})} & ; \quad ec_j \in EEC \end{cases} \qquad (4)$$

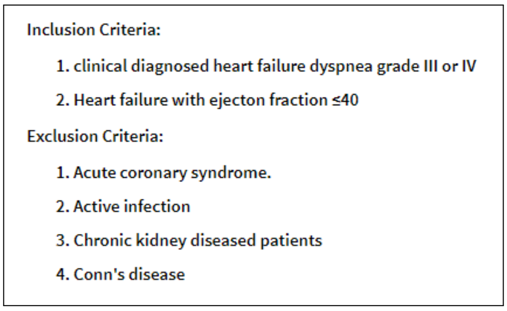
Finally, the matching score $\mathcal{M}$ of Patient P_i with a Clinical Trial is determined by:

$$\mathcal{M}(P_i, CT) = \sum_{j=1}^{l} rij. \qquad (5)$$

IV. EVALUATION

To validate our framework, we used two datasets; MIMIC-III (Medical Information Mart for Intensive Care) [38] comprising information relating to patients admitted to critical care units, and Clinical Trials [2] a Web-based resource providing access to information on supported clinical studies.

[2] https://clinicaltrials.gov/



Inclusion Criteria:

1. clinical diagnosed heart failure dyspnea grade III or IV
2. Heart failure with ejecton fraction ≤40

Exclusion Criteria:

1. Acute coronary syndrome.
2. Active infection
3. Chronic kidney diseased patients
4. Conn's disease

Fig. 3. The eligibility criteria specified in the NCT04078425 clinical trial

TABLE III
RANKS AND SCORES OF MATCHING 10 PATIENTS WITH 6 ELIGIBILITY CRITERIA (NCT04078425)

	iec1	**iec2**	**eec1**	**eec2**	**eec3**	**eec4**	**Score**
P-1	9.46	9.84	-8.47	-2.74	-1.02	-1.02	*6.03*
P-2	5.02	7.44	-3.43	-3.12	-8.42	-8.42	*-10.93*
P-3	9.08	8.65	-10.00	-2.38	-10.00	-10.00	*-14.65*
P-4	0.00	4.09	-2.96	-5.76	-5.24	-5.24	*-15.12*
P-5	3.43	4.02	-6.26	-2.69	-1.09	-1.09	*-3.69*
P-6	5.19	0.00	0.00	-1.42	0.00	0.00	*3.77*
P-7	5.65	2.95	-3.86	-2.72	-0.15	-0.15	*1.72*
P-8	7.26	10.00	-7.52	-5.76	-6.98	-6.98	*-9.99*
P-9	6.43	9.14	-4.44	-10.00	-2.70	-2.70	*-4.27*
P-10	10.00	7.44	-6.27	0.00	-10.00	-10.00	*-8.83*

A. *Text processing*

MIMIC III Clinical Dataset is a critical care database that contains 2,083,108 medical reports from 46,520 patients. We experimented with a randomly selected dataset of 100 Discharge Summaries from patients last visit, excluding patients whose ages are under 18. The segmentation stage produces an average of 400 phrases per report.

We selected a clinical trial that identifies the role of Aldosterone antagonist in patients of heart failure with preserved ejection fraction (NCT04078425). Fig. 3 shows the five eligibility criteria of this clinical trial.

B. *Evaluation of the obtained results*

Table III presents the results for a sample of ten patients. In order to evaluate the clinical correctness of patients matching to the clinical trial(NCT04078425), a validation task was performed manually by a nurse and a computer science student. The noteworthy fact is that the evaluation of the matching does not reveal false positives in the score results. Indeed, the similarity scores reflect the order of matching between patients and the clinical trial. The score distribution ranged from (-15) to (8), and eligible patients to be retained for further screening by experts were those with a score greater than 5.

We should note that the scores would be more realistic if the segmentation process was more accurate. For instance, the sentence ”you were thought to have a blood clot in your right leg” was segmented by Metamap into ”a blood clot in your right leg” which would result in a false outcome.

V. CONCLUSION

EMRs contain a large portion of unstructured data that need to be matched with eligibility criteria for trial-patient enrollment. Indeed, the gradual improvement of artificial intelligence technology could reduce the number of physician-hours spent in screening patient eligibility. To tackle the problem, we proposed a framework designed to automatically recommend the most suitable patients for a clinical trial. The framework adopts a pre-trained language model (BioBERT) and uses STS-B and MedNLI datasets to improve the accuracy of the model via transfer learning. This work verified that the fine-tuning of BioBERT shows better performance in calculating the similarity between two medical sentences using embedding-based metrics. In future works, we will also explore EMRs structured tables in order to significantly improve the performance and accuracy of our trial-patient matching framework.

ACKNOWLEDGMENT

The authors would like to thank Marvin Moughabghab for his efforts and contributions to this work.

REFERENCES

[1] H. Dhayne, R. Haque, R. Kilany, and Y. Taher, “In search of big medical data integration solutions-a comprehensive survey,” *IEEE Access*, vol. 7, pp. 91 265–91 290, 2019.

[2] G. De Moor, M. Sundgren, D. Kalra, A. Schmidt, M. Dugas, B. Claerhout, T. Karakoyun, C. Ohmann, P.-Y. Lastic, N. Ammour *et al.*, “Using electronic health records for clinical research: the case of the ehr4cr project,” *Journal of biomedical informatics*, vol. 53, pp. 162–173, 2015.

[3] M. Dugas, M. Lange, C. Müller-Tidow, P. Kirchhof, and H.-U. Prokosch, “Routine data from hospital information systems can support patient recruitment for clinical studies,” *Clinical Trials*, vol. 7, no. 2, pp. 183–189, 2010.

[4] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson, “Data from clinical notes: a perspective on the tension between structure and flexible documentation,” *Journal of the American Medical Informatics Association*, vol. 18, no. 2, pp. 181–186, 2011.

[5] H. Dhayne, R. Kilany, R. Haque, and Y. Taher, “Sedie: A semantic-driven engine for integration of healthcare data,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 617–622.

[6] P. Raghavan, J. L. Chen, E. Fosler-Lussier, and A. M. Lai, “How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?” *AMIA Summits on Translational Science Proceedings*, vol. 2014, p. 218, 2014.

[7] S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim, “A practical method for transforming free-text eligibility criteria into computable criteria,” *Journal of biomedical informatics*, vol. 44, no. 2, pp. 239–250, 2011.

[8] S. Kripalani, F. LeFevre, C. O. Phillips, M. V. Williams, P. Basaviah, and D. W. Baker, “Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care,” *Jama*, vol. 297, no. 8, pp. 831–841, 2007.

[9] K. Milian, R. Hoekstra, A. Bucur, A. ten Teije, F. van Harmelen, and J. Paulissen, “Enhancing reuse of structured eligibility criteria and supporting their relaxation,” *Journal of biomedical informatics*, vol. 56, pp. 205–219, 2015.

[10] C. Patel, J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivas, “Matching patient records to clinical trials using ontologies,” in *The Semantic Web*. Springer, 2007, pp. 816–829.

[11] T. Kang, S. Zhang, Y. Tang, G. W. Hruby, A. Rusanov, N. Elhadad, and C. Weng, “Eliie: An open-source information extraction system for clinical trial eligibility criteria,” *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1062–1071, 2017.

[12] C. Yuan, P. B. Ryan, C. Ta, Y. Guo, Z. Li, J. Hardin, R. Makadia, P. Jin, N. Shang, T. Kang *et al.*, "Criteria2query: a natural language interface to clinical databases for cohort definition," *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 294–305, 2019.

[13] C. Shivade, C. Hebert, M. Lopetegui, M.-C. De Marneffe, E. Fosler-Lussier, and A. M. Lai, "Textual inference for eligibility criteria resolution in clinical trials," *Journal of biomedical informatics*, vol. 58, pp. S211–S218, 2015.

[14] Y. Ni, S. Kennebeck, J. W. Dexheimer, C. M. McAneney, H. Tang, T. Lingren, Q. Li, H. Zhai, and I. Solti, "Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 166–178, 2014.

[15] D. D. Palmer, "Tokenisation and sentence segmentation," *Handbook of natural language processing*, pp. 11–35, 2000.

[16] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium.* American Medical Informatics Association, 2001, p. 17.

[17] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[18] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederee, and W. Nejdl, "A blocking framework for entity resolution in highly heterogeneous information spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 12, pp. 2665–2682, 2012.

[19] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A c-lstm neural network for text classification," *arXiv preprint arXiv:1511.08630*, 2015.

[20] S. Moen and T. S. S. Ananiadou, "Distributional semantics resources for biomedical text processing."

[21] G. Soğancıoğlu, H. Öztürk, and A. Özgür, "Biosses: a semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, 2017.

[22] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[23] Q. Chen, Y. Peng, and Z. Lu, "Biosentvec: creating sentence embeddings for biomedical texts," *arXiv preprint arXiv:1810.09302*, 2018.

[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[25] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] A. Talman and S. Chatzikyriakidis, "Testing the generalization power of neural network models across nli benchmarks," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 85–94.

[28] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[29] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[30] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: pre-trained biomedical language representation model for biomedical text mining," *arXiv preprint arXiv:1901.08746*, 2019.

[31] H. Xiao, "bert-as-service," https://github.com/hanxiao/bert-as-service, 2018.

[32] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 2019, pp. 15–18.

[33] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[34] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[35] "google-research/bert: Tensorflow code and pre-trained models for bert," https://github.com/google-research/bert, (Accessed on 09/17/2019).

[36] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," *arXiv preprint arXiv:1808.06752*, 2018.

[37] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:1708.00055*, 2017.

[38] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.