Journal of the American Medical Informatics Association, 0(0), 2019, 1–8
doi: 10.1093/jamia/ocz064
Research and Applications

Research and Applications

Optimizing clinical trials recruitment via deep learning

Jelena Gligorijevic,[1,*] Djordje Gligorijevic,[1,*] Martin Pavlovski,[1,2] Elizabeth Milkovits,[3] Lucas Glass,[3] Kevin Grier,[3] Praveen Vankireddy,[3] and Zoran Obradovic[1]

[1]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, Pennsylvania, USA, [2]Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia and [3]IQVIA, Plymouth Meeting, Pennsylvania, USA

*Authors contributed equally

Corresponding Author: Zoran Obradovic, PhD, Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, Pennsylvania, USA (zoran.obradovic@temple.edu)

Received 13 January 2019; Revised 11 April 2019; Editorial Decision 18 April 2019; Accepted 20 April 2019

ABSTRACT

Objective: Clinical trials, prospective research studies on human participants carried out by a distributed team of clinical investigators, play a crucial role in the development of new treatments in health care. This is a complex and expensive process where investigators aim to enroll volunteers with predetermined characteristics, administer treatment(s), and collect safety and efficacy data. Therefore, choosing top-enrolling investigators is essential for efficient clinical trial execution and is 1 of the primary drivers of drug development cost.
Materials and Methods: To facilitate clinical trials optimization, we propose *DeepMatch* (DM), a novel approach that builds on top of advances in deep learning. DM is designed to learn from both investigator and trial-related heterogeneous data sources and rank investigators based on their expected enrollment performance on new clinical trials.
Results: Large-scale evaluation conducted on 2618 studies provides evidence that the proposed ranking-based framework improves the current state-of-the-art by up to 19% on ranking investigators and up to 10% on detecting top/bottom performers when recruiting investigators for new clinical trials.
Discussion: The extensive experimental section suggests that DM can provide substantial improvement over current industry standards in several regards: (1) the enrollment potential of the investigator list, (2) the time it takes to generate the list, and (3) data-informed decisions about new investigators.
Conclusion: Due to the great significance of the problem at hand, related research efforts are set to shift the paradigm of how investigators are chosen for clinical trials, thereby optimizing and automating them and reducing the cost of new therapies.

Key words: clinical trials, electronic health records, deep learning, deep matching, pointwise ranking

INTRODUCTION

Clinical trials constitute a multi-billion–dollar industry aimed to provide an appropriate environment for successful evaluation of drug effectiveness in large patient populations, ultimately resulting in well-tested cures for diseases.[1] A clinical trial is carried out by a large and distributed team of investigators (often physicians) responsible for ensuring that the trial is conducted according to certain regulations. Investigators are responsible for enrolling often hard-to-recruit volunteers with predetermined characteristics, administering the treatment(s), and collecting data on the subjects' health, safety, and treatment efficacy. Enrolling patients for clinical trials can take years,[1] so the success of clinical trials thus heavily depends on choosing top-enrolling investigators; therefore, this is 1 of the primary drivers of drug development cost and delays in health delivery.[2,3] Site selection, the process in which investigators are chosen to participate in a clinical trial, is 1 of the primary tasks of contract research organizations (CROs). The objective of our study is to optimize this process by developing an effective data science tool for ranking investigators by their expected performance, as even a minor improvement in site selection accuracy is worth tens of millions of dollars[4] as well as weeks or months until a drug can be released on the market.

Investigators for new clinical trials are traditionally selected manually by searching through available in-house and public databases from various sources.[5] This selection process is very tedious and fallible due to the infeasibility of searching through every possible record manually. Recruiters thus resort to maintaining short lists of preferred investigators, whereas investigators who wish to be enrolled in future studies need to reach out to recruiters in addition to being registered in public databases.[5]

CROs have a strong interest in the development of automated systems for selecting investigators for new clinical trials primarily because such systems would shorten the enrollment period which they oversee, and a delay of a single day in bringing new drugs to the market equates to between a $1 million and $5 million loss[4] and potentially months of lost health delivery opportunities. As even the most recent advances in enrollment strategies involve infeasible manual selection,[3] the objective of novel systems is to automatically learn from available databases, thereby shifting the core of how selection of investigators has been historically carried out.

The databases generated by these organizations have accumulated vast information over the past several decades covering up to 95% of the medical claims data in the United States as well as enrollment information from thousands of clinical trials. Such sources of data provide a massive opportunity for novel data-driven approaches to give rise to automated site selection. Claims data, as 1 source of information in this work, were successfully used to address several high-impact health care tasks,[6,7] especially through deep learning models.[8,9]

An additional challenge to clinical trial planning is the nuanced nature of clinical trials: no 2 studies are alike, and studies are very difficult to parameterize. Understanding the nuances by learning from free-text data that describe existing medical processes has been a long-standing challenge of research communities, as there are no clear standard methods or tools for analyzing medical text data yet.[10] Several approaches including a symbolic rule-based approach were proposed in the past for predicting cancer stages from free-text pathology reports.[11,12] Such models, however, require handcrafting features, which is long and tedious work, and often require to be revisited to further improve the obtained features. Notable success was achieved with the bi-RNN algorithm[13] used for detection of medical events based on textual features from medical records.[14] Another information-rich source in medical and health care applications are electronic health records (EHRs) that contain data about the patients' diagnoses, procedures, and medications. Learning distributed representations of medical concepts from EHRs was recently proposed in Gligorijevic et al[15] and Choi et al.[16] The goal of these studies was to evaluate the quality of such representations by discovering disease associations.[15]

Furthermore, it has been recently shown that deep architectures obtain cutting-edge results on very sensitive medical tasks. For instance, Gligorijevic et al. proposed a deep learning model that outperformed a hospital's triage staff at predicting triage resource allocation in emergency departments.[9]

Finally, since many predictive problems can be formulated as matching problems, deep learning models that perform matching explicitly were developed. For matching of 2 items, *siamese network architectures* were proposed.[17] More recently, such architectures were further advanced.[18–20]

Motivated by these advances, we propose *DeepMatch* (DM), a novel deep learning model designed specifically to rank investigators for clinical trials through pointwise regression on their estimated enrollment. DM learns deep representations of (1) investigators given their specialty indications and patients' history and (2) clinical trials given their official description, primary indication (PI) and primary therapeutic area (PTA).

These 2 separately learned representations are then matched to obtain a joint representation. This was later used to predict the investigators' enrollment scores which, in this study, are the z-normalized counts of patients enrolled by each investigator for a certain clinical trial. Therefore, DM is

- capable of capturing complex relations between investigators and trials through a dedicated matching layer incorporated within its architecture;
- able to rank investigators for a specific trial from multiple partially observed data sources;
- applicable to investigators with no previous experience in enrolling patients in clinical trials, as well as to investigators with whom the company has an existing relationship.

An extensive experimental evaluation was conducted providing evidence that the proposed framework outperforms the current industry standard, as well as other deep and shallow learning alternatives, on the following 2 tasks: (1) ranking investigators for new clinical trials, and (2) detecting bottom 30% and top 30% performing investigators.

METHODS AND MATERIALS

Site selection for clinical trials

This study's aim is to rank investigators for site selection in order to maximize the number of patients enrolled in clinical trials. More precisely, for an upcoming study, the objective is to determine a list of top-enrolling investigators based on the historical enrollment performance of all investigators on past studies. This requires a system design that (1) involves collection of data relevant to the problem, (2) uses these data to model the investigators' enrollment performance based on past trial records, and (3) has the ability to make reliable predictions of investigators' enrollment scores for future studies. Figure 1 shows how the data for investigators, including self-reported specialty and patient history data, can be matched to clinical trials data (detailed descriptions of a trial's purpose, treatment, and desired population for treatment testing) through reports containing past investigators' performance. Such data contain rich information that can be used by the CROs to rank investigators for upcoming clinical trials according to their performance, thus facilitating better clinical research overall.

Data

The data used in this study contain information related to both investigators and studies collected over the past 20 years. The following provides a brief description of data collected and aggregated from several heterogeneous data sources:

- *Investigator performance data.* A proprietary data set collected by the CRO that bridges investigators with their history of performed clinical studies. All major biopharma companies and CROs maintain similar proprietary data sets. Therefore, this data set contains information on both investigators and the studies they participated in. The former relates to investigators' areas of specialty, while the latter holds a textual format and includes

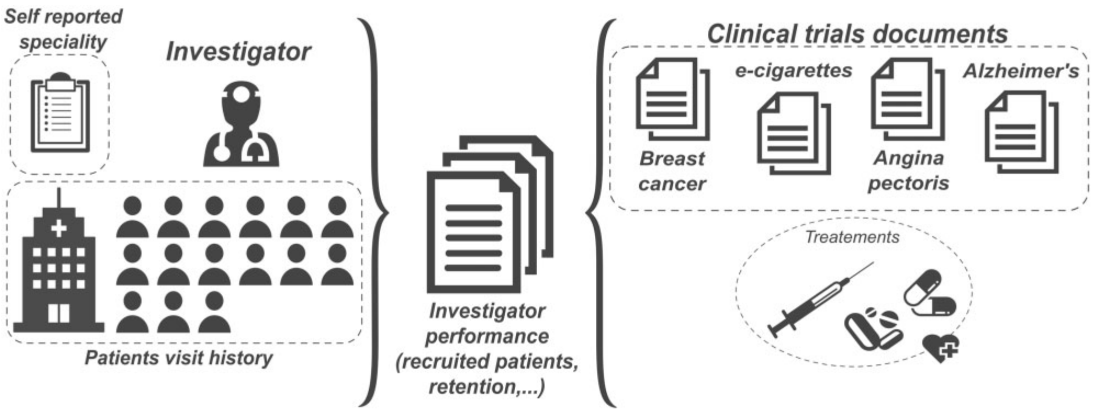

Figure 1. The proposed system for matching clinical trials to investigators. Investigators (on the left) provide self-reported specialties, and their patient visit history may be available. Clinical trials (on the right) contain documents describing the treatment and purpose of the trial, primary indication (PI), primary therapeutic area (PTA), and desired population. The matched pairs of investigators and trials contain investigator performance reports from historical data used by the considered models to learn to rank investigators.

information about a study's primary indication and therapeutic area. In addition, for each investigator, the number of patients that he or she enrolled to each of their past studies is also provided. Consequently, since each investigator–study pair has an enrollment score, if a record of a given investigator is not present in this file, such an investigator has no enrollment history (ie, has never participated in a clinical study).

- *EHR data.* An integrated electronic health record (EHR) database of patients' medical claims and prescription history that covers 65% of all physician claims and 92% of all prescription claims in the US. These records include information about patients' visits over time in terms of the diseases they were diagnosed with (up to 15 per visit), procedures they underwent (up to 15 per visit), and the medications that were prescribed by investigators (up to 5 per visit).
- *Public study data.* Detailed study-related free-form text data crawled from clinicaltrials.gov—the largest registry of clinical studies throughout the world.

Upon collection, the data from the described sources were integrated into 2 separate views:

- *Investigator data view.* To create this view, investigators were retrieved from the investigator performance data. Thereafter, each investigator was represented by a vector of most frequent codes of diagnosed diseases, conducted procedures, and prescribed medications during all visits to the corresponding investigator stemming from the EHR data source.
- *Study data view.* Each record in this view includes the primary indication and therapeutic area of a single study stemming from the investigator performance data and further textual description of the study obtained from clinicaltrials.gov.

For a complete system overview, refer to Figure 2.

Proposed approach DeepMatch

The architecture of the proposed DM model is illustrated in Figure 3. The following text describes the building blocks of DM, including all hyperparameters, in detail.

Building blocks

Medical concepts embedding layer. From the EHR data we chose the most frequently diagnosed, procedures, and medications prescribed by each investigator, resulting in an input of length $l_i^{(1)} = 130$. In particular, $l_i^{(1)}$ thresholds from the most frequent 50 diagnoses, 50 procedures, and 30 prescriptions for each investigator, where threshold values are chosen based on basic descriptive statistics of medical concept codes across all investigators. An embedding lookup layer is built to learn distributed representations of medical concepts (having a dimensionality of $d_m = 200$), after which 2 fully connected (FC) layers with ReLU nonlinearities are used to learn higher-order interactions between these embeddings, thus capturing complex relations between diseases, procedures, and medications. The final representation of this data source has dimensionality $l_i^{(1)} \times \ d_p$, where $d_p = 50$ is the dimension of a joint representation in the same space with all input data components.

Medical and trial terms embedding layer. Investigators' specialties, public trials text, and trials PIs/PTAs have a common vocabulary of medical terms. Thus, the inputs for the investigators' specialties ($l_i^{(2)} = 10$), trials PIs/PTAs ($l_s^{(2)} = 10$), and public trials text ($l_s^{(1)} = 100$) were embedded using a medical term embedding layer with a dimensionality of $d_w = 300$. The investigators' specialties and the trials PIs/PTAs are passed through 2 fully connected layers, respectively, building dense representation vectors of medical term interactions, each of size $1 \times \ d_p$. As public trials text contains multiple elements concatenated into a single document, it is worthwhile to learn relations of terms in different segments. Thus, we pass the obtained representation of public trials text into a bi-long short-term memory[13] layer such that the model can learn complex relations between terms (rather than single directional relations). The final representation of the public trial data become $(l_s^{(1)} \times d_p)$-dimensional.

Investigators and clinical trials matching tensor. The investigator–trial matching is obtained in this layer. Using the learned higher-order representations of investigators H_i of shape $(1 + l_i^{(1)}) \times d_p$ and clinical trials H_s of shape $(1 + l_s^{(1)}) \times d_p$, a tensor is built for their implicit matching. A matching tensor H_{is} of shape $(1 + l_i^{(1)}) \times \left(1 + l_s^{(1)}\right) \times d_p$ is created in the following manner:

$$H_{is}(m, n, :) = H_i(m, :) \ \odot \ H_s(n, :),$$

where $m, n \in \mathbb{N}, \ 1 \leq m \leq \left(1 + l_i^{(1)}\right), \ 1 \leq n \leq \left(1 + l_s^{(1)}\right)$ and $\odot$ represents the element-wise product of $H_i(m, :)$ and $H_s(n, :)$. The first 2 dimensions of the tensor H_{is} are $\left(1 + l_i^{(1)}\right)$ and $\left(1 + l_s^{(1)}\right)$, respectively. Here, each investigator's medical concept is matched to

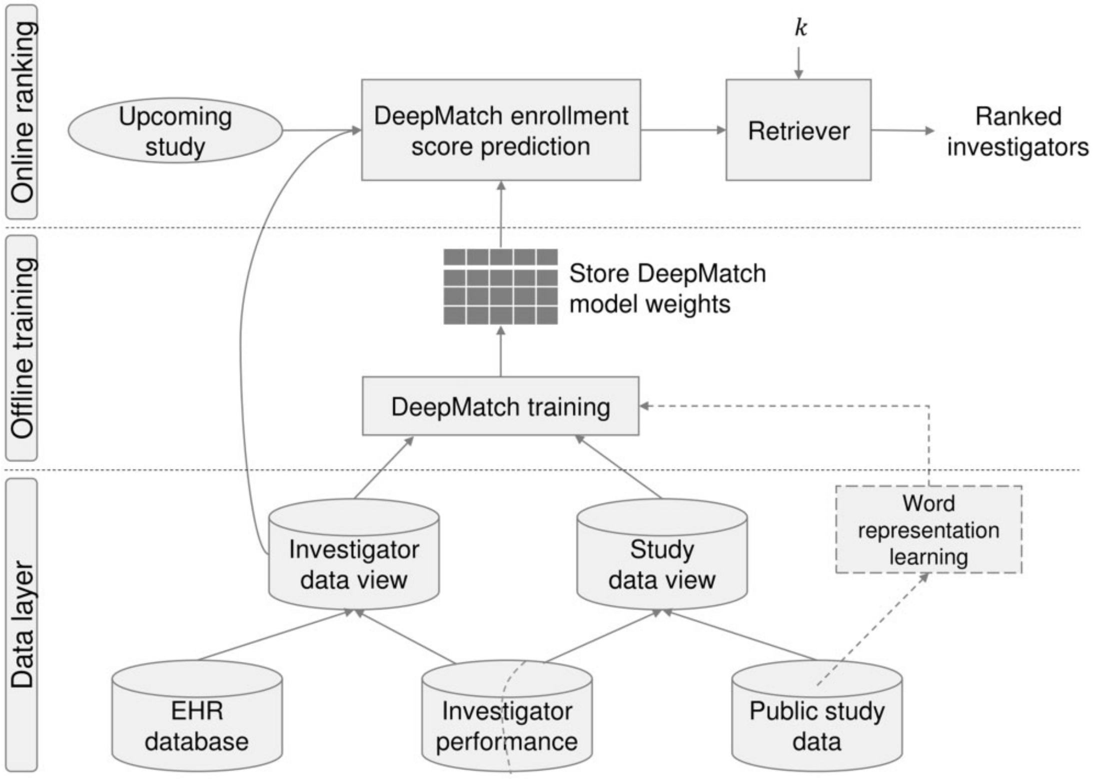

Figure 2. System overview. The system has several layers consisting of multiple components employed for carrying out separate tasks: *1. Data layer*. The data sources are integrated to create separate views for investigators and studies. Word2vec[21] is employed to learn vector representations of relevant medical concepts occurring in all free-form public study texts (clinicaltrials.gov). All contents of the study data view, along with the learned medical term representations, are passed as an additional input to the offline training layer. *2. Offline training*. DeepMatch (DM) learns representations for the medical terms driven to predict the investigators' enrollment scores. DM's weights are then stored on a distributed file storage. *3. Online ranking*. DM's weights are loaded in this layer and ranking is performed for upcoming studies in an on-the-fly fashion. An upcoming study is passed to the matching component where it is matched against existing investigators from the investigator data view; their enrollment scores with regards to the given study are then predicted; finally, the top *k* most eligible investigators are returned by the retriever.

each trial-related term, and the element-wise product of their vectors represents a d_p-dimensional thread in the matching tensor. With this operation, we aim to capture intra-relations between investigators' medical concepts along with their specialties and clinical trials medical terms.

Learning to predict enrollment scores from matched representations. The matching tensor H_{is} from the previous block is convolved through the entire depth $d_p = 50$ by 3 convolutional blocks with different filter sizes: $3 \times 3,\ 3 \times 4,\ 3 \times 5$. The choice of this combination was determined by filter size fine-tuning. The number of filters is fixed to 6 for both the first set of convolution blocks and the final convolutional layer. Such cross-convolutional operators were successfully applied in the past to tasks of modeling similarities between general sentences.[22] Interaction representations between medical concepts and trial-related terms are learned here, and they are passed through the ReLU layer, after which another 1 ×1 convolution with ReLU is used prior to the 2-dimensional max-pool layer that embeds a whole investigator–study pair into a single vector. Finally, this vector is fed into a fully connected layer and passed through a squared-loss ($L = (y - f(x))^2$, where $f(x)$ is the network described above) layer to predict enrollment scores for investigator–study pairs.

The enrollment scores for each investigator per clinical trial are finally used for obtaining the investigator rank.

Experiments

We designed an extensive empirical setup to assess the performance of DM and address the following research questions:

[Q1]: How do the models compare when ranking investigators for clinical trials?
[Q2]: Does adding additional deep layers boost the generalization performance of DM?
[Q3]: Is DM capable of capturing complex relations between investigators and trials through its matching layer better than using simpler strategies (a simple concatenation of the investigators' and trials features)?
[Q4]: How does DM perform when trained with and without the trial-related text data in addition to the internal EHR and performance data?
[Q5]: How do the models perform for investigators with no previous trial participation history?
[Q6]: How do the models compare for detecting the bottom/top 30% performing investigators?

Experimental setup

The experiments were conducted in a manner that reflects a real-world situation from the clinical trial business such that the same trial–investigator pairs cannot appear in the training, validation, and test data subsets. In particular, all models were tested on studies

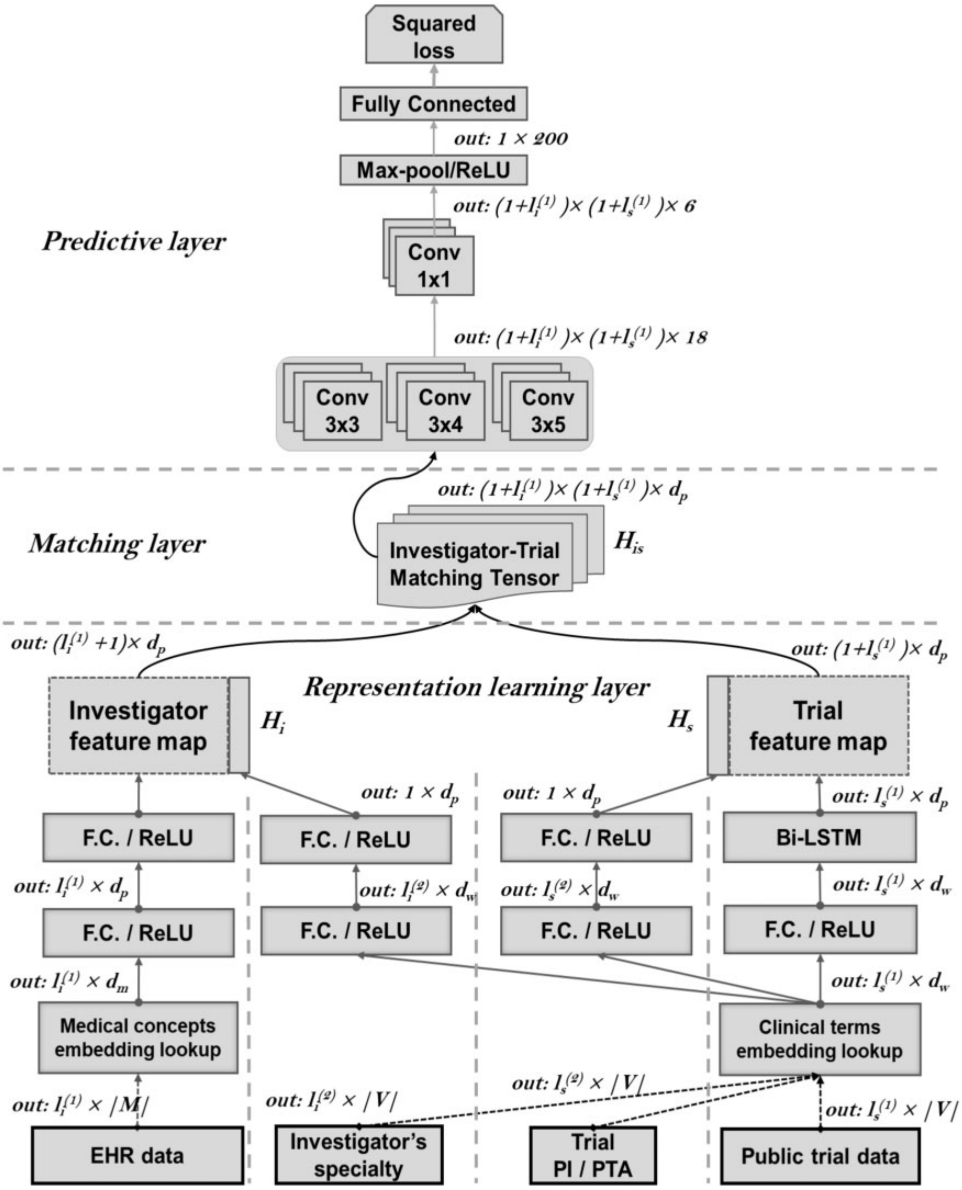
Squared loss
Fully Connected
out: 1×200
Max-pool/ReLU
out: $(1+l_i^{(1)}) \times (1+l_s^{(1)}) \times 6$
Predictive layer
Conv 1x1
out: $(1+l_i^{(1)}) \times (1+l_s^{(1)}) \times 18$
Conv 3x3
Conv 3x4
Conv 3x5
out: $(1+l_i^{(1)}) \times (1+l_s^{(1)}) \times d_p$
Matching layer
Investigator-Trial Matching Tensor
H_{is}
out: $(l_i^{(1)}+1) \times d_p$
out: $(1+l_s^{(1)}) \times d_p$
Representation learning layer
Investigator feature map
H_i
H_s
Trial feature map
out: $1 \times d_p$
out: $1 \times d_p$
out: $l_s^{(1)} \times d_p$
F.C. / ReLU
F.C. / ReLU
F.C. / ReLU
Bi-LSTM
out: $l_i^{(1)} \times d_p$
out: $l_i^{(2)} \times d_w$
out: $l_s^{(2)} \times d_w$
out: $l_s^{(1)} \times d_w$
F.C. / ReLU
F.C. / ReLU
F.C. / ReLU
F.C. / ReLU
out: $l_i^{(1)} \times d_m$
out: $l_s^{(1)} \times d_w$
Medical concepts embedding lookup
Clinical terms embedding lookup
out: $l_i^{(1)} \times |M|$
out: $l_i^{(2)} \times |V|$
out: $l_s^{(2)} \times |V|$
out: $l_s^{(1)} \times |V|$
EHR data
Investigator's specialty
Trial PI / PTA
Public trial data

Figure 3. DeepMatch(DM) model architecture. DM takes 2 sets of inputs that correspond to features of an investigator and text of a clinical trial from universes of medical terms $|V|$ and concepts $|M|$. An investigator's features consist of a list of his/her clinical specialty areas represented as an $l_i^{(2)} x\ |V|$ matrix and summarized EHR data of investigators' patients represented as $l_i^{(2)} x\ |V|$. A clinical trial features include a public report for a trial (text) and its PI and PTA; these components are represented as $l_s^{(2)} x\ |V|$ and $l_s^{(1)} x\ |V|$, respectively. The embeddings of both input sets V and M are obtained through a series of layers, including fully connected layers with rectified linear units and bi-directional LSTM layers which learn interactions between words and concepts in each input independently. The element-wise product is then calculated for all pairs of the learned embeddings and organized in a matching tensor. Finally, the constructed matching tensor is passed through a series of cross-convolutional and pooling operators to learn the investigator–trial enrollment scores.

that started in 2016 and 2017. Studies that started in 2015 were used for validation and hyperparameter tuning, whereas all studies that started prior to 2015 were used for training. The characterization of the training, validation, and test set sizes is given in Table 1.

The task is to rank investigators by their expected enrollment.

Baseline methods

The accuracy and relevance of enrollment scores predicted by DM was compared against the following alternatives (representing both current industry standard and state-of-the-art learning approaches):

- *Median enrollment (ME)*. The current industry standard is to take the median enrollment from within the therapeutic area of interest to predict future enrollment scores.
- *Point-wise support vector regression (SVR)*. A support vector regression model[23] that approaches the investigator-to-study matching in a pointwise ranking fashion.
- *Linear model (LM)*. A linear model is employed for predicting investigators' enrollment scores based on their handcrafted features (their specialties, as well as the frequencies of their patients' diagnoses, procedures, and medications) and study text data (summarized by summing up the numerical vectors of clinical trial terms).

Table 1. Training, validation, and test set size characterization: Number of Unique Studies, Number of Unique Investigators, Number of Records, Average Number of Investigators per Study, and Average Number of Patients Recruited per Investigator–Study pair. The relatively small number of recruited patients per investigator reflects the typical number of patients sought per trial type (275 patients per cardiovascular trial, 20 patients per cancer trial [which makes up almost 50% of all trials], 70 patients per depression trial, and 100 per diabetes trial[24])

	Years	UniqueStudies	UniqueInv.	No. of Records	Avg. no.Inv. per Study	Avg. no. Recruited Patients per Inv.
Train.	< 2015	2250	21223	70906	32	5
Valid.	2015	209	1958	2489	11	4
Test	2016, 2017	159	1841	2236	14	4

- *Multilayer perceptron (MLP). A* 2-layer MLP model is used to learn the second-order interactions of the handcrafted features and the summarized trials text to predict the investigators' enrollment scores.
- *DeepConcat (DC).* A model similar to the proposed DeepMatch; however, in DC, trials and investigators' feature maps are summarized into respective vectors which are only concatenated afterwards.

Evaluation measures. As our main goal is to rank investigators for upcoming studies, all models were evaluated using the normalized discounted cumulative gain (NDCG).[25–26] Given a list of investigator–study pairs ranked by their enrollment scores, NDCG is calculated as $\boldsymbol{NDCG@K} = \frac{DCG@K}{IDCG@K}$, where $\boldsymbol{DCG@K} = \sum_{k=1}^{K} \frac{2^{rel_k}-1}{log_2(k+1)}$ with $\boldsymbol{rel_k}$ being a tier relevance value for the enrollment score of the k-th investigator, whereas ***IDCG@K*** represents the ideal DCG (producing the maximum possible DCG through position K). The final *NDCG* values are reported as the average of *NDCG* values over the 159 studies in the test set (see Table 1). In the conducted experiments, in addition to *NDCG@K*, *NDCG@N*$_i$ was measured as well, where N_i is the number of investigators that participated in the i-th study, and the average over all studies is reported as *NDCG@N*.

As most of the baselines directly predict investigators' enrollment scores, we also compare these models using the mean squared error (MSE) measure.

Finally, to evaluate the models' performance in a more domain-savvy manner, we measure their classification accuracy for the task of detecting the bottom/top 30% performing investigators, where the goal is to simply classify whether an investigator is within the top or bottom 30% of the performers on a given study.

RESULTS

DM and alternatives were trained using gradient descent with learning rate of 0.0001 over 20 data set iterations on NVIDIA P100 GPU machines and were evaluated such that the research questions (Q1–Q6) are answered in a clear and comprehensive manner.

Ranking investigators for new clinical trials

The goal is to rank investigators for upcoming clinical trials. To answer the first research question [Q1], we measure NDCG when ranking investigators for each upcoming test study and report the average NDCGs across all test studies. Due to the variable number of investigators across different clinical trials Table 2 reports the *NDCG@N* scores of the 6 models, where *N* is the total number of investigators for a particular study. Figure 4 shows their ranking performance in terms of *NDCG@K* when *K* is varied from 2 to 10.

Table 2. Performance of DM and alternatives w.r.t. *NDCG@N* (higher is better) for the investigators ranking task and *MSE* (lower is better) for the enrollment score prediction task. Note that *MSE** is the *MSE* calculated in terms of the unstandardized enrollment score (ie, the number of patients)

	NDCG	MSE	MSE*
ME	0.63	1.17	75
SVR	0.64	100	6423
LM	0.69	100	6423
MLP	0.70	0.62	39
DC	0.72	0.58	37
DM	**0.74**	**0.38**	**24**

Abbreviations: DC, DeepConcat; DM, DeepMatch; LM, linear model; ME, median enrollment; MLP, multilayer perceptron; MSE, mean squared error; NDCG, normalized discounted cumulative gain; SVR, support vector regression.

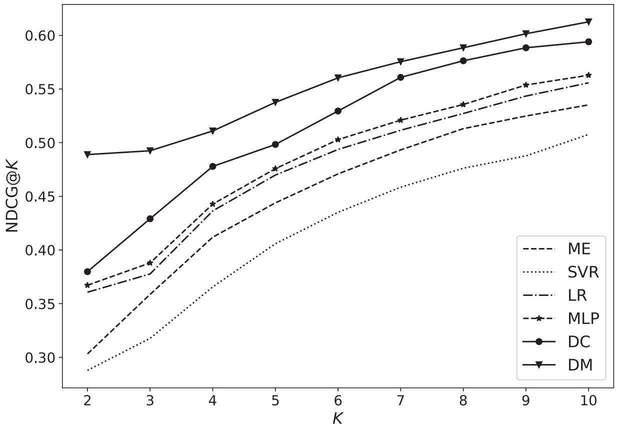

Figure 4. Average NDCG@K for DM vs 5 alternatives across 159 test studies, for K in range of 2 to 10.

The best performing model for the task of ranking investigators for new studies was the proposed DM. It brings in an average accuracy improvement of 2.5% in NDCG over the next best performing DC model (a statistically significant improvement according to the Wilcoxon signed-rank test, where the corresponding *P* value of 1.61e^{-11} is measured on *NDCG@2* through *NDCG@30* along with *NDCG@N*) and even an ~11% improvement over the industry standard ME.

Even though our business task relies on a ranking problem, it is approached through regression; therefore, we report MSE in Table 2. According to these results, DM outperformed all alternatives on the enrollment score prediction task as well. It was more accurate, on average, than the industry standard and the best-performing baseline by approximately 4 and 2 recruited patients (calculated as difference of squared roots of MSE*), respectively. The average number of recruited patients being 4 supports the significance of these improvements (Table 1).

Deeper architectures work better. Another conclusion drawn from these results is that the models relying on deeper architectures seem to capture certain latent patterns among the investigator–study pairs, thus leading to greater generalization performance compared to the shallow alternatives. More precisely, the generalization performance when ranking investigators ameliorates as additional layers are introduced: the worst accuracy is obtained by a linear model, it is improved by using 2-layer MLP, and was the most accurate for DC and DM, which answers research question [Q2].

DM captures complex relations between investigators and trials through the matching layer. The statistically significant improvement introduced by DM's matching layer over DC's simple concatenation layer shows that DM can capture complex relations between investigators and trials while learning a joint representation, which is useful for the overall goal of predicting enrollment scores. This observation unfolds the answer to [Q3].

Trial-related text data improve the performance of DM. To answer question [Q4], we evaluated DM's performance using only internal data to determine whether the effort of crawling and processing the public study data brings any improvement. The experimental results indicate that using the partially observed public study data (available for ~75% of the studies in the internal database) brings an improvement of ~1% in average NDCG over all studies and ~3% of improvement measured on new investigators only, thus justifying the engineering effort.

Handcrafting features from text data is a tedious task even with well-nuanced and clearly written documents of clinical trials, primarily due to nonconsistency in clinical trial terminology (eg, some companies use the term *clinical trial*, while others use *clinical study*) and writing styles. Therefore, we learned text representations on original clinical trial texts (without text preprocessing) as keeping even a single medical abbreviation (eg, "rct" or "t2dm") is of great importance when using word2vec$_{21}$ to initialize the embedding layers of DM.

The largest accuracy improvement was observed for investigators with no previous clinical trial participation history. In the test set, ~25% of the cases (534 out of 1841) are investigators with no previous experience in clinical trials. All methods based on historical enrollment, including ME, are incapable of making predictions for such cases. To answer [Q5], in Table 3 we show the *NDCG@N* obtained by DM and all alternative models when measured (1) on all test cases, (2) only for investigators with previous experience, and (3) only for new investigators (not previously observed in the training data set).

DM showed the largest improvement of 5% over the next best baseline model when predicting scores of investigators with no previous experience. Contrary to ME, the proposed framework is capable of handling cold-start cases by exploiting the available context of each investigator (his patient visits or his reported specialties), and drawing benefits from the matching layer and the information available in the study data view. Although the proposed model has shown smaller improvements (such as 2%) on (1) all cases, and (2) previously known investigators, it still yields the best performance, with the highest gain being achieved against the current industry standard.

Note that the values from different columns in Table 3 are not comparable in this form as *N* is different for (1), (2), and (3) per study. However, when *K* was fixed to a smaller number in our further analysis, simpler methods (ME, LM) obtained the best results on known investigators (2), while deeper models were able to learn informative features that enabled them to perform well on all 3 groups.

Table 3. *NDCG@N* for DM and alternative models measured (1) on all test cases, (2) only for investigators with previous experience, and (3) only for new investigators (not previously observed in the training data set)

	(1) All	(2) Known Only	(3) Unknown
ME	0.63	0.67	0.61
SVR	0.64	0.68	0.71
LM	0.69	0.72	0.73
MLP	0.70	0.73	0.74
DC	0.72	0.74	0.75
DM	**0.74**	**0.76**	**0.80**

Abbreviations: DC, DeepConcat; DM, DeepMatch; LM, linear model; ME, median enrollment; MLP, multilayer perceptron; SVR, support vector regression.

Table 4. Classification performance in terms of Recall, ROC AUC, Accuracy and F1 score on the task of detecting bottom and top 30% performers

	Bottom 30				Top 30			
	Recall	AUC	Accuracy	F1	Recall	AUC	Accuracy	F1
ME	0.64	0.72	0.75	0.64	0.53	0.64	0.68	0.53
SVR	0.50	0.60	0.64	0.50	0.42	0.55	0.59	0.42
LM	0.63	0.71	0.74	0.63	0.48	0.60	0.65	0.48
MLP	0.65	0.72	0.75	0.65	0.54	0.65	0.69	0.54
DC	0.65	0.73	0.76	0.65	0.60	0.69	0.73	0.60
DM	**0.67**	**0.75**	**0.78**	**0.67**	**0.65**	**0.74**	**0.77**	**0.65**

Abbreviations: DC, DeepConcat; DM, DeepMatch; LM, linear model; ME, median enrollment; MLP, multilayer perceptron; SVR, support vector regression.

Detecting bottom 30% and top 30% performers

In addition to ranking investigators for upcoming studies, an important task [Q6] is to identify which investigators will be the low-performing and exclude the bottom 30% in the business process of selecting sites for clinical trials. The 30% of top enrollees are identified as well, and the performance in both cases in terms of Recall, ROC AUC, Accuracy, and F1 score are shown in Table 4.

The current industry standard to site performance prediction is to take an average (or median) of the investigators' historical performance within the therapeutic area of interest. If a low-enrolling investigator is defined as 1 ranked in the bottom 30% of a given study, this approach can accurately identify low enrollees 39% of the time (as measured using internal data). However, we report the performance of ME, a variant of the industry standard, that predicts the enrollment score of an investigator as the median over all his or her past enrollment scores instead of over the investigators' scores within the therapeutic area of interest, since such a modification has shown to be more accurate in our additional analysis. This modified ME predicts the median of an investigator's historical enrollment scores, or it predicts zero for an investigator with no previous enrollment history.

The results provide evidence that the proposed model was able to identify top 30% and bottom 30% performers more accurately than alternatives, bringing large business value to the company. As mentioned earlier, for a typical Phase 2 or Phase 3 clinical trial for which enrolling patients can take years, even a minor improvement in accuracy at predicting the top and bottom enrolling sites is worth tens of millions of dollars.

CONCLUSION

DM, a novel deep learning model, is proposed to rank investigators for clinical trials along with an outline of a system that utilizes this model for optimizing site selection. The matching capabilities of DM were assessed for the tasks of (1) ranking investigators, and (2) identifying bottom/top performers for upcoming clinical trials. Compared against several alternative models, and under a variety of scenarios, DM outperformed its alternatives. Moreover, advances presented in this study are applicable beyond the presented case. The proposed methods can be utilized in any task where one needs to discover a match among instances from multiple sources of structured and unstructured data. Prominent examples of such tasks are online recommender systems and professional networking services. Due to the high impact of the problem at hand, the authors believe that related research efforts will shift the paradigm of how investigators are selected for clinical trials and pave the road to clinical trial optimization and automation, reduce drug development costs, and ultimately expedite delivery of new therapies.

FUNDING

This study was supported by the IQVIA grant Big Data Analytics for Clinical Trials Optimization.

CONTRIBUTORS

J.G. and Dj.G. designed and implemented the method. J.G., Dj.G. and M.P. conducted all the experiments. All authors were involved in developing the ideas and writing the paper.

COMPETING INTERESTS

None.

REFERENCES

1. Friedman, Lawrence M., Curt Furberg, David L. DeMets, David M. Reboussin, and Christopher B. Granger. *Fundamentals of clinical trials*. Vol. 4. New York: Springer, 2010.
2. Martin L, Hutchens M, Hawkins C, *et al*. How much do clinical trials cost? *Nat Rev Drug Discov* 2017; 16 (6): 381–82.
3. Hurtado-Chong A, Joeris A, Hess D, *et al*. Improving site selection in clinical studies: a standardised, objective, multistep method and first experience results. *BMJ Open* 2017; 7: 7.
4. Mullard A. 2016 FDA drug approvals. *Nat Rev Drug Discov* 2017; 16 (2): 73–6.
5. How Do CROs Choose Sites and Investigators for Their Clinical Trials? https://www.theclinicaltrialsguru.com/blog1/how-do-cros-choose-sites-and-investigators-for-their-clinical-trials. Accessed January 2019.
6. Choi E, Taha Bahadori M, Schuetz A, *et al*. Doctor ai: predicting clinical events via recurrent neural networks. In: *Machine Learning for Healthcare Conference*; 2016: 301–18.
7. Stojanovic J, Gligorijevic D, Radosavljevic V, *et al*. Modeling healthcare quality via compact representations of electronic health records. *IEEE/ACM Trans Comput Biol Bioinform PP* 2016; 99: 1–10.
8. Che Z, Kale D, Li W, *et al*. Deep computational phenotyping. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 10–13 Aug, Sydney, NSW, Australia. 2015: 507–16.
9. Gligorijevic D, Stojanovic J, Satz W, *et al*. Deep attention model for triage of emergency department patients. In: Proceedings of the 2018 SIAM International Conference on Data Mining. 3–5 May. San Diego, California, USA. SIAM; 2018: 297–305.
10. Chapman W, Nadkarni P, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011; 18: 540–3.
11. Nguyen A, Lawley M, Hansen D, *et al*. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc* 2010; 17 (4): 440–5.
12. Coden A, Savova G, Sominsky I, *et al*. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J Biomed Inform* 2009; 42 (5): 937–49.
13. Schuster M, Paliwal K. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 1997; 45 (11): 2673–81.
14. Jagannatha A, H Yu. Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the Association for Computational Linguistics Conference; 12–17 Jun, 2016; 473–482. San Diego, California, USA: NIH Public Access.
15. Gligorijevic D, Stojanovic J, Djuric N, *et al*. Large-scale discovery of disease-disease and disease-gene associations. *Sci Rep* 2016; 6: 1–12.
16. Choi E, Taha Bahadori M, Searles E, *et al*. Multi-layer representation learning for medical concepts. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining;13 – 17 Aug, San Francisco, California, USA: ACM 2016: 1495–504.
17. LeCun Y, Bengio Y, *et al*. Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks* 1995; 3361: 10.
18. Gligorijevic, J., Gligorijevic, D., Stojkovic, I., et al. Deeply supervised model for click-through rate prediction in sponsored search. *Data Mining and Knowledge Discovery*, 2019: 1–22.
19. Huang P, He X, Gao J, *et al*. Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, October 27 - November 01; 2013: 2333–2338. San Francisco, California, USA: ACM.
20. Edizel B, Mantrach A, Bai X. Deep character-level clickthrough rate prediction for sponsored search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval; 07 – 11 Aug, 2017: 305–314. Shinjuku, Tokyo, Japan: ACM.
21. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems* 2013: 3111–9. Lake Tahoe, Nevada:Curran Associates Inc.
22. Hu B, Lu Z, Li H, *et al*. Convolutional neural network architectures for matching natural language sentences. In: *Advances in Neural Information Processing Systems*; 2014: 2042–50. Cambridge, MA, USA: MIT Press.
23. Ingo S, Andreas C. *Support Vector Machines*. New York: Springer Science & Business Media; 2008.
24. Giffin, R.B., Lebovitz, Y. English RA, eds. *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary*. Washington, DC: National Academies Press; 2010.
25. Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst (TOIS)* 2002; 20 (4): 422–46.
26. Wang Y, Wang L, Li Y, *et al*. A theoretical analysis of NDCG ranking measures. In: Proceedings of the 26th Annual Conference on Learning Theory; 12–14 Jun,2013; 30: 25–54. Princeton, NJ, USA:PMLR.