A query interface for clinical research with Chinese electronic health record using Natural Language Processing

Mengyang Li
College of Biomedical Engineering and Instrument Science, Zhejiang University
Hailing Cai
College of Biomedical Engineering and Instrument Science, Zhejiang University
Yani Chen
College of Biomedical Engineering and Instrument Science, Zhejiang University
Xudong Lu (✉ lvxd@zju.edu.cn)
College of Biomedical Engineering and Instrument Science, Zhejiang University
Huilong Duan
College of Biomedical Engineering and Instrument Science, Zhejiang University

Research Article

Keywords: clinical research, clinical trial, electronic health record, query interface, natural language processing

Posted Date: June 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1664677/v1

Abstract

Background: The recruitment of clinical trials is a challenging task. Traditionally, the time-consuming task is accomplished manually or assisted by form-based tools. Electronic health record (EHR) contains comprehensive information which can speed up the process. The development of natural language processing (NLP) makes it possible to reduce manual effort. To our knowledge, there are almost no query interfaces based on NLP techniques for Chinese EHR.

Methods: A query interface based on NLP was developed. Firstly, we collected and annotated eligibility criteria(EC) from the Chinese Clinical Trial Registry. Then, an information extraction(IE) system was developed to parse them including named entity recognition(NER), relation classification(RC), and concept matching(CM). Next, the extracted information was post-processed(PP) to generate formal queries. Finally, we built a natural language query interface based on the system.

Results: 4691 stroke-related EC were collected for annotation. 91.23% F1 score was achieved for the NER task. For the RC task, 92.75% F1 score was achieved. For the CM task, 90.78% accuracy was achieved and an evaluation result showed 42.42% of entities from NER results can be matched with EHR. And a natural language query interface has been implemented and applied in clinical research about stroke.

Conclusions: We build a query interface for Chinese EHR based on NLP techniques. The proposed information extraction pipeline can support medical professionals to reduce the information gap with minimal human effort when interacting with Chinese EHR.

Background

Electronic health record (EHR) contains comprehensive information. It has attracted attention from clinical research, public health, and quality improvement. Many institutions [1, 2] have established clinical data repositories (CDR) combined with query interfaces, such as i2b2 [3], to facilitate the utilization of EHR data. However, it is still a challenging task to query in EHR. Traditionally, clinical information is collected manually by looking through the patient records. Structured Query Language (SQL) is powerful enough, it is too difficult for medical experts without too much technical training. Even for users with good technical backgrounds, it is not easy because it requires a good understanding of the exact structure of databases and domain knowledge. Some query interfaces, such as keyword-based search [4], form-based based interface [5, 6], and visual query builder [3, 7] have been widely accepted, they still don't meet the requirements effectively. There exist difficulties for keywords to express complex query intent [8]. Form-based interfaces enable complex queries by a set of restrictions but require predictable queries and are limited to hard-coded logic [9]. A visual query builder can support complicated queries by providing flexible building blocks. Sometimes, they can be too restrictive or complex for users for specific requirements.

With the widespread development of NLP techniques, they have shown their promise in medical fields [10–13]. So, it provides an alternative option to improve queries in EHR. Many methods have been proposed. Specifically, some IE methods are proposed to extract medical entities within sentences [14–17]. These studies have laid the foundation for the automatic execution of queries in EHR [18, 19]. There are still many challenges to be tackled even with them. In addition, Text-to-SQL methods [20] transform queries in free text into structured and computable representation by end-to-end architectures [21–23]. They achieve automatic queries in relational databases. On the other hand, pipeline-based methods [18, 24] provide a flexible alternation. Modular components can be combined according to specific requirements. Users can revise the output of every component by manual interaction and utilize the automatic features provided by NLP techniques to a maximal extent. Most of these studies and systems mentioned above are only available in English.

Accordingly, a natural language interface for clinical research based on Chinese EHR is proposed. To be specific, patient recruitment in clinical research is treated as a representative example in our study. The query interface transforms EC into structured information. In this way, the interface provides the technology-neutral implementation to achieve a decoupling architecture where the execution of queries is transparent to specific query languages. In this paper, several sub-tasks are designed: (1) annotate EC text from clinical trials; (2) implement an IE pipeline for EC; (4) develop a natural language query interface for EHR. To our knowledge, there is barely a complete solution to transform EC into formal queries by NLP techniques automatically in Chinese.

The structure of the rest of this paper is as follows. Our method is proposed in Section 2. After EC are collected, preprocessed, and annotated, some NLP components are introduced to transform EC into a structured format. Afterward, formal queries are generated according to extracted structured information. Finally, on these bases, a natural language query interface is developed. Section 3 gives the results, including the performance of the IE pipeline and the proposed query interface. Section 4 discusses the contributions of our method, and some relevant issues and future directions. And conclusions are summarized in Section 5.

Methods

The IE system for EC was developed as shown in Fig. 1. First, clinical trials were collected from the Chinese Clinical Trial Registry(ChiCTR, http://www.chictr.org.cn/searchproj.aspx) [25] and medical text corpora were collected from Medicine Database in Library of Academy of Military Sciences in China(http://211.103.242.133:8080/Disease/Fast.aspx)[26], respectively. Then, EC in clinical trials were split, filtered, and annotated for NER and RC tasks. Meanwhile, medical text corpora were parsed and checked to generate synonyms and negative samples for the CM task. Finally, the annotated sentences were used to verify the performance of the IE system. The detailed illustrations of each component are as follows.

Data annotation

EC incorporate the inclusion criteria and exclusion criteria. Each of them consists of multiple paragraphs which are numbered and divided by line breaks. Many sentences may occur in the same paragraph. So, these paragraphs are broken up into single sentences. Then, some characters or punctuation will be removed, such as the numbered list and the bulleted list to get normalized sentences. Not every sentence is meaningful to query in EHR. For example, recent participants in other clinical trials are excluded or informed consent is not met. So, we implement a heuristic method to classify these sentences. A series of keywords are defined to exclude these criteria.

The annotation work was performed by 3 persons including a clinician and 2 medical informatics engineers. First, the annotation models need to be defined. The entity types are summarized from EC text. Concretely, EC from 5 clinical trials were randomly selected and they were annotated by the clinician first to obtain the medical entity types as shown in Table 1. Then, semantic relations as shown in Table 2 are collected from EC conforming to demand-oriented design. Based on the annotated EC, the guideline about annotation was given to keep the consistency among the three annotators. Then, 4691 sentences are divided into 30 groups. Every group consists of nearly 150 sentences. Every annotator annotated every group independently and review the other groups annotated by other 2 annotators. The annotation process is shown in Fig. 2. The annotation work is an iterative process. After the cross annotating, every group can be annotated and reviewed many times until the occurred disagreements were eliminated.

Meanwhile, during the annotation, all entities are annotated with the longest length which enables many modifiers to be incorporated into final entities. For example, 'severe cognitive impairment' will be annotated rather than 'cognitive impairment'. All annotation work is completed in Brat [27]. Then, BIO format was employed to transform the annotated sentences.

NER task

NER is an important task of IE and a prerequisite for the success of downstream tasks. It is treated as a sequence labeling problem. In this task, 3 machine learning-based methods including Conditional Random Field (CRF), BiLSTM-CRF[28] and BiLSTM-CRF with adversarial training [29] (BiLSTM-CRF-ADV) were selected to recognize the defined entities.

Among the methods, CRF was treated as a baseline method because it has been used in Criteria2Query [18] which was a natural language interface for English. BiLSTM-CRF was widely used in clinical activities [30, 31] because it is simple and can recognize the entities effectively.

Also, adversarial training was employed as a regularization method to constrain the BiLSTM-CRF model. By artificially adding some disturbances during the training process, it can improve the robustness of the model by allowing the model to correctly identify these artificially added disturbances. Two methods, Fast Gradient Method (FGM) [32] and Projected Gradient Descent (PGD) [33] were employed to improve the performance of BiLSTM-CRF.

Table 1
Summary of entity types

Entity types	Description
Diagnosis	Indicate the existence of a disease or sign/symptom, which can be statements provided by doctors or patients.
Drug	Indicate the used medications for patients. It can be detailed drug name or names according to its indications.
Procedure	Represent the activities/behavior to support treatment and diagnosis.
Test	Represent the quantifiable concepts. It can be structured concepts from laboratory test items, or examinations, or questionnaires/scales.
Demographic	Represent the statistical characteristics of patients, including age, gender, address, etc.
Value	Represent the quantities of some measured items, such as results of laboratory tests, drug doses. Sometimes it may contain units, such as 6.7 * 10^9/L.
Negx	Represent the logic operator — negation, to exclude some medical concepts during the queries.

Table 2
Summary of relation types

Relation types	Description
has_value	Represent the relationship among some quantitative concepts such as laboratory tests and values.
has_negx	Identify the logical negation of some medical concepts.
is_a	Associate the hierarchical relationships among different entities with different granularities.
indicate	Associate some laboratory tests or imaging examination with diagnosis/sign.
cause	Identify the causal relationships, such as diseases causes signs.

RC task

After medical entities were extracted, a relation classifier will be run to identify the relationship among these entities. They can be used to provide rich semantic information and reduce the scope of the query. In this study, we built a relation classifier based on BiLSTM with the attention mechanism(BiLSTM-ATT) [34].

In a single sentence, there may exist a series of medical entities with different types. To identify relations, pairwise entities were grouped into triples with the unidirectional relationship. Two entities can have only one type of defined relations. For those triples without any relations, they were joined by "unknown" relation and treated as negative samples during the training.

The features used to capture the information about different entities consist of word features and position features. The word features were used to express the meaning of different entities. The position features were used to determine the boundary of different entities. For example, <e1 > Neurological deficit</e1 > is caused by diseases such as < e2 > brain tumor</e2>, <e3 > brain trauma</e3>, etc. <e1>, </e1>, <e2>, </e2>, <e3>, </e3 > were six positions which specify the starting and ending of the three entities. All these features were treated as an initial input to BiLSTM-ATT model.

CM task

After the medical entities were recognized, they need to match with candidate medical concepts. To align the entities, a series of synonyms were obtained from medical text corpora collected from Medicine Database in Library of Academy of Military Sciences in China(http://211.103.242.133:8080/ziyuan/CDDPdf/dis/base/□□□□\□□□□□.pdf) [26]. They were parsed by regular expressions and checked by manual.

Every parsed entity pair was labeled with '1' to indicate that they represented the same medical entities. To train classifiers for CM, negative samples were chosen from these entities randomly. To be concrete, these extracted entities were organized into a graph. And connected components were recognized from this graph. Two components were chosen from these different connected components randomly. Then a concept was chosen from every single component. The chosen concepts were combined as a negative sample.

We explored and empirically used ESIM [35] to measure semantic textual similarity. By modeling with local information, it learned the information to support global reasoning. It can help improve the prediction performance. And inference speed of the method was very fast due to its relatively simple architecture. The advantage was very important for query tasks.

PP Task

The extracted information was formed for query generation. This study aimed at providing a decoupling way for query execution. Extracted information was organized into a query-irrelevant format which can be formed into any formal queries by a profile. Every sentence in EC was treated as an independent unit. Entities in each sentence were related to their entity type and relations. They were organized into the format as shown in Fig. 3. Entities in different sentences in inclusion and exclusion criteria were joined with the “intersect” operator, respectively. Entities in the same sentence in inclusion and exclusion criteria were joined with the “union” operator, respectively. And, the final results were obtained by subtracting exclusion results from inclusion results.

Results

Experiment Setup

The data used in our experiment are collected from ChiCTR[25]. 389 stroke-related clinical trials are collected up to January 1, 2020, including 2178 inclusion criteria and 2513 exclusion criteria with a total of 4691 sentences. And the statistical information is summarized in Table 3. All of the data are spilt into training, validation, and testing data with the rough ratio of 8:1:1. According to the defined entity types and relations, they are summarized as shown in Table 4 and Table 5.

Table 3
Statistical information of Dataset

Statistic	Number (#)
# of clinical trials	389
# of inclusion criteria	2178
# of exclusion criteria	2513
# of total criteria	4691
Average criteria sentence length	23.54
Average concepts number in each criterion	2.13

Table 4
Data partition for Entities

Entity Types	# of medical entities	Train	Valid	Test
Diagnosis	5692	4564	535	593
Drug	225	180	23	22
Procedure	994	819	79	96
Test	1074	855	113	106
Demographic	454	376	44	34
Value	1278	1027	128	123
Negx	261	209	25	27

Table 5
Data partition for Relations

Relation Types	# of relations	Train	Valid	Test
has_value	1424	1139	142	143
has_negx	303	242	30	31
is_a	499	399	49	51
indicate	783	626	78	79
cause	393	314	39	40

To measure the performance of these tasks, we used the precision(P), recall(R), and F1 score as the evaluation metrics for NER and RC tasks. Meanwhile, we used accuracy and Area under curve (AUC) to evaluate the CM task.

We trained these models using training data and tune parameters on the validation set. Then we examined the model on the testing set. We carried out experiments five times with different random seeds to calculate the average metrics. The early stopping strategy was used to avoid overfitting.

NER results

CRF used in Criteria2Query with Macro-F1 86.78% is treated as a baseline to compare the performance of different methods. Compared with CRF, BiLSTM-CRF achieves the overall performance with a Macro-F1 90.58% according to Table 6. We can find that entity types like ‘Diagnosis’, ‘Procedure’, ‘Negx’ can get better results than theirs in CRF. However, entity types like ‘Drug’, ‘Test’, ‘Demographic’ and ‘Value’ get the degraded performances. There are different results for adversarial training strategies like PGD and FGM. PGD achieves a Macro-F1 90.84% and FGM achieves a Macro-F1 91.23% which is better than the original BiLSTM-CRF model.

Table 6
Performance results of the NER task

Entity Types	CRF			BiLSTM-CRF			BiLSTM-CRF-PGD			BiLSTM-CRF-FGM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Diagnosis	86.66	88.25	87.45	90.11	91.23	90.67	90.92	90.78	90.85	91.62	90.56	91.09
Drug	90.86	86.67	88.72	91.23	87.08	89.11	91.18	87.46	89.28	91.02	87.59	89.27
Procedure	83.61	76.77	80.04	87.12	87.50	87.31	88.89	86.92	87.89	88.94	86.87	87.89
Test	85.87	82.29	84.04	86.81	89.96	87.87	86.62	88.96	87.77	87.73	89.35	88.53
Demographic	93.59	96.89	95.21	97.67	98.59	98.13	97.78	97.58	97.68	98.54	98.41	98.47
Value	89.38	90.38	89.88	88.58	91.52	90.03	90.21	91.52	90.86	90.89	91.62	91.25
Negx	87.84	76.20	81.61	91.01	90.78	90.89	91.81	91.02	91.41	92.01	92.24	91.62
Average	88.26	85.35	86.78	90.36	90.81	90.58	91.06	90.61	90.84	91.54	90.95	91.23

To apply these methods in the query interface, experiments about recognition speed are carried out. We use fixed parameters with maximum sentence length (= 150) and batch size (= 32). Meanwhile, we test two groups in different devices as shown in Table 7. We notice the recognition time of all these methods is not more than 1 second. CRF can achieve the best recognition speed. Compared with CRF, performance degrades for the BiLSTM + CRF model. Meanwhile, FGM methods can reduce the recognition time with an improvement of 0.131 seconds compared with BiLSTM + CRF.

Table 7
Results of recognition time in the NER task

Models	Recognition time(s)
CRF	0.5491
BiLSTM + CRF	0.9928
BiLSTM + CRF + PGD	0.9820
BiLSTM + CRF + FGM	0.8618

RC results

The experiment results of the RC task are shown in Table 8. BiLSTM-ATT achieves the overall performance with a F1 92.75%. Among different relation types, the result for the “has_value” relation is greater than 95%. Other relation types are also greater than 90%. We also evaluate the classified speed. The inference time with BiLSTM-ATT is 3.931 seconds with the same experiment as the NER task.

Table 8
Performance results of the RC task

Relation Types	Precision	Recall	F1
has_value	94.59	97.90	96.22
has_negx	94.57	93.04	93.80
is_a	88.74	91.43	90.06
indicate	92.12	91.86	91.99
cause	91.39	93.58	92.47
Average	92.14	93.38	92.75

CM results

12458 positive samples are parsed into synonyms. Based on them, 12500 negative samples are chosen randomly to train classifiers. After training, the ESIM model can achieve accuracy with 90.78% and AUC with 0.9639. Also, the classifier was used in the CM task for 593 entities with the diagnosis type which were from testing data from the NER task. These entities were merged into 334 unique entities after duplicate removal. 42.42% of entities can be matched with the candidate entities in EHR.

Query Interface

A query interface is designed and implemented based on the work mentioned above. The interface is shown in Fig. 4. It is used for patient screening for clinical research at present. For a clinical trial, there exist inclusion and exclusion criteria about the enrolled research objects and they are grouped in our interface. The recognized medical entities are highlighted in each query sentence to help users confirm their query intention more clearly. If the identified entities are wrong or they need to be further revised, users can change the query text through the interactive components. The returned results can be customized according to actual requirements. Furthermore, the sorting of results depends on the number of medical concepts. The more medical entities occur, the higher the interface is likely to rank the patient.

Discussion

We implement a query interface for Chinese EHR using NLP techniques. To our best knowledge, it is the first natural language query interface for Chinese EHR by transforming EC into a formal representation. Therefore, our contribution can be summarized as follows:

First, we generate an annotated corpora for Chinese EC and implement an IE pipeline. Compared with studies in English, the related studies are much less about clinical trials recruitment, patient matching in Chinese even there exist some tasks [36, 37]. From the usability perspective, there is almost no comprehensive work before our studies. Therefore, it can be treated as a baseline for further studies. And the pipeline can be used in multiple applications including patient recruitment mentioned in this paper. It can also support the construction of medical knowledge base in clinical trials [38].Compared with Criteria2Query in English, we explore deep learning-based methods empirically and get better performance. Meanwhile, it is implemented based on OMOP CDM which constrains that it can only be used in relational databases. Scalability can be a considerable challenge.

Second, the query interface provides a representative solution for clinical research with EHR data. One of the most challenging tasks of the utilization of EHR data is data access. Nevertheless, the query interface enables researchers, clinicians, or nurses to query EHR data autonomously without having to master detailed information techniques. This makes it easier for users to access medical data and avoids time-consuming manual work. Meanwhile, the query interface can be packaged as APIs to embed in multiple medical applications to promote the utilization of EHR data.

Limitations And Future Work

Inevitably, there are limitations to our study, although it can improve the user's experience from the perspective of usability. According to these limitations, we put forward the corresponding solutions.

Firstly, it is necessary to improve annotated corpora in terms of types and quantity. In this paper, we use 4691 sentences as corpora. They are mainly extracted from cardiovascular diseases. There exist challenges to be generalized to other diseases. Currently, we have implemented several methods to address the issue. One is to allow users to revise recognition results. They can be treated as a golden standard to further train models. Rule-based methods are used to build a dictionary with string similarity matching. The revised results are imported into the dictionary to help subsequent tasks. Also, we are planning to use active learning methods to expand the training dataset which can reduce manual annotation efforts.

Second, better methods for NER, RC, and CM tasks are required. In this study, we explore empirically deep-learning methods to accomplish these tasks. Meanwhile, we also use pretrained models BERT[39] to evaluate them and achieve an F-1 score of 94.26%, 97.96% for NER and RC tasks, respectively. However, it also brings 64.7574 seconds and 152.7334 seconds inference time. The results hinder its application in the query interface. Therefore, different aspects including accuracy and speed should be taken into account and more methods are required for further evaluation.

Finally, we have not conducted a large-scale user evaluation for the query interface. It is used in a clinical research platform now. Clinical researchers use it to find potential patients. To improve the query interface and analyze the importance of different components, formal and rigorous statistical study is necessary.

Conclusion

The query interface demonstrates how to transform Chinese EC into computer-executable queries. The proposed IE pipeline reduces the information gap among clinical researchers and computers. It can support medical professionals to access Chinese EHR more flexibly with no need to learn underlying information techniques.

Abbreviations

EHR: Electronic health record

NLP: Natural language processing

EC: Eligibility criteria

IE: Information extraction

NER: Named entity recognition

RC: Relation classification

CM: concept matching

PP: Post-processed

CDR: Clinical data repository

SQL: Structured query language

FGM: Fast gradient method

PGD: Projected gradient descent

ChiCTR: Chinese Clinical Trial Registry

AUC : Area under curve

CRF: Conditional random field

Declarations

Ethics approval and consent to participate

All experiments were performed in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due because the datasets are related to other ongoing studies and we do not have consent from them but are available from the corresponding author on reasonable request.

Competing interests

The author(s) declare(s) that they have no competing interests.

Funding

This study was funded by the Chinese National Science and Technology Major Project (grant number 2020YFC2003401) and National Natural Science Foundation of China(grant number 82172070).

Authors’ contributions

Mengyang Li designed the research and conducted the experiments. Hailing Cai and Mengyang Li developed the query software. Mengyang Li wrote the manuscript. Yani Chen reviewed the manuscript. Xudong Lu and Huilong Duan supervised the study and reviewed the manuscript.

Acknowledgements

Not applicable

References

1. Waitman, L. R., Warren, J. J., Manos, E. L., & Connolly, D. W. (2011). Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. In AMIA Annual Symposium Proceedings (Vol. 2011, p. 1454). American Medical Informatics Association.
2. Abend, A., Housman, D., & Johnson, B. (2009). Integrating clinical data into the i2b2 repository. Summit on translational bioinformatics, 2009, 1.
3. Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association, 17(2), 124-130.
4. Wu, H., Toti, G., Morley, K. I., Ibrahim, Z. M., Folarin, A., Jackson, R., et al. (2018). SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*. Journal of the American Medical Informatics Association, 25(5), 530–537. http://doi.org/10.1093/jamia/ocx160
5. J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories," in Proceedings of the IEEE Symposium on Visual Analytics Science And Technology (VAST), pp. 167–174, Baltimore, MD, 2006.
6. Plaisant, C., Lam, S., Shneiderman, B., Smith, M. S., Roseman, D., Marchand, G., ... & Rappaport, H. (2008). Searching electronic health records for temporal patterns in patient histories: A case study with microsoft amalga. In AMIA annual symposium proceedings (Vol. 2008, p. 601). American Medical Informatics Association.
7. Gall, P. S. G. D. W. D. W. (1999). A retrieval system for the selection and statistical analysis of clinical data. Medical Informatics and the Internet in Medicine, 24(3), 201–212.
8. Wei, X., & Zeng, D. D. (2016). ExNa: an efficient search pattern for semantic search engines. Concurrency and Computation: Practice and Experience, 28(15), 4107–4124.
9. Wagholikar, K. B., Ainsworth, L., Vernekar, V. P., Pathak, A., Glynn, C., Zelle, D., et al. (2019). Extending i2b2 into a framework for semantic abstraction of EHR to facilitate rapid development and portability of Health IT applications. AMIA Jt Summits Transl Sci Proc, 2019, 370–378.
10. Murff HJ, FitzHenry F, Matheny ME, et al (2011) Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 306:848–855 . doi: 10.1001/jama.2011.1204
11. Maddox TM, Albert NM, Borden WB, et al (2017) The learning healthcare system and cardiovascular care: A scientific statement from the American Heart Association. Circulation 135:e826–e857 . doi: 10.1161/CIR.0000000000000480
12. Wu P-Y, Cheng C-W, Kaddi CD, et al (2017) –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng 64:263–273. doi:10.1109/TBME.2016.2573285
13. Wang Y, Wang L, Rastegar-Mojarad M, et al (2017) Clinical Information Extraction Applications: A Literature Review. J Biomed Inform 77:34–49. doi:10.1016/j.jbi.2017.11.011
14. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011 Dec;18 Suppl 1:i116-i124
15. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. J Am Med Inform Assoc 2017 Nov 01;24(6):1062-1071
16. Tseo Y, Salkola MI, Mohamed A, Kumar A, Abnousi F. Information extraction of clinical trial eligibility criteria. arXiv preprint arXiv:2006.07296 2020 Jun 12:1-4.
17. Chen M, Du F, Lan G, Lobanov VS. Using pre-trained transformer deep learning models to identify named entities and syntactic relations for clinical protocol analysis. 2020 Presented at: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1); March 25, 2020; Palo Alto, CA p. 1-8.
18. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc 2019 Apr 01;26(4):294-305
19. Xiong Y, Shi X, Chen S, Jiang D, Tang B, Wang X, et al. Cohort selection for clinical trials using hierarchical neural network. J Am Med Inform Assoc 2019 Nov 01;26(11):1203-1208

20. Kalajdjieski, J., Toshevska, M., & Stojanovska, F. (2020). Recent Advances in SQL Query Generation: A Survey. arXiv preprint arXiv:2005.07667.
21. Wang P, Shi T, Reddy CK. Text-to-SQL generation for question answering on electronic medical records. 2020 Presented at: Proceedings of The Web Conference 2020; April 20-24, 2020; Taipei, Taiwan p. 350-361
22. Zhang, X., Xiao, C., Glass, L. M., & Sun, J. (2020, January 23). DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction. arXiv.org.
23. Gao, J., Xiao, C., Glass, L. M., & Sun, J. (2020, June 16). COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching. arXiv.org.
24. EMR2vec: Bridging the gap between patient data and clinical trial. (2021). EMR2vec: Bridging the gap between patient data and clinical trial. Computers & Industrial Engineering, 156, 107236. http://doi.org/10.1016/j.cie.2021.107236
25. Chinese Clinical Trial Registry. Available at: http://www.chictr.org.cn/index.aspx. Accessed 21 November 2021.
26. Medicine Database. Available at: http://211.103.242.133:8080/Index.aspx. Accessed 21 November 2021.
27. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 102-107).
28. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.
29. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
30. Qin, Y., & Zeng, Y. (2018). Research of clinical named entity recognition based on bi-lstm-crf. Journal of Shanghai Jiaotong University (Science), 23(3), 392-397.
31. Huali, Z. H. A. N. G., Xiaodong, K. A. N. G., Bo, L. I., Yage, W. A. N. G., Hanqing, L. I. U., & Fang, B. A. I. (2020). Medical name entity recognition based on Bi-LSTM-CRF and attention mechanism. Journal of Computer Applications, 0.
32. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
33. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
34. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016, August). Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 207-212).
35. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., & Inkpen, D. (2016). Enhanced lstm for natural language inference. arXiv preprint arXiv:1609.06038.
36. Zeng, K. (2020). An Ensemble Learning Strategy for Eligibility Criteria Text Classification for Clinical Trial Recruitment: Algorithm Development and Validation. JMIR Medical Informatics, 8(7), e17832. http://doi.org/10.2196/17832
37. CHIP2019. http://www.cips-chip.org.cn:8000/evaluation. Accessed 21 November 2021.
38. Liu, H., Chi, Y., Butler, A., Sun, Y., & Weng, C. (2021). A knowledge base of clinical trial eligibility criteria. Journal of Biomedical Informatics, 117, 103771.
39. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Figures

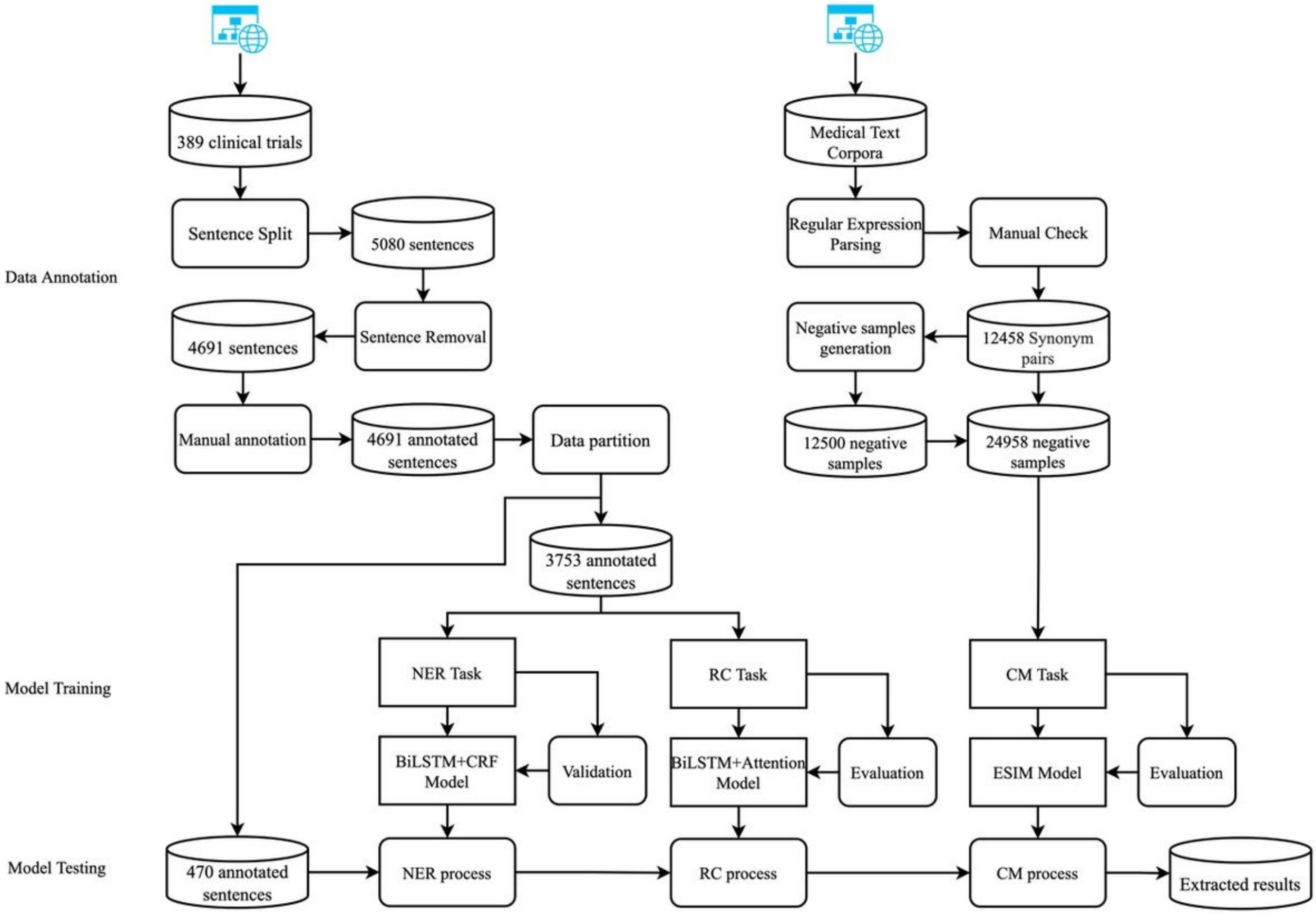


Figure 1

Development process of the IE system for EC

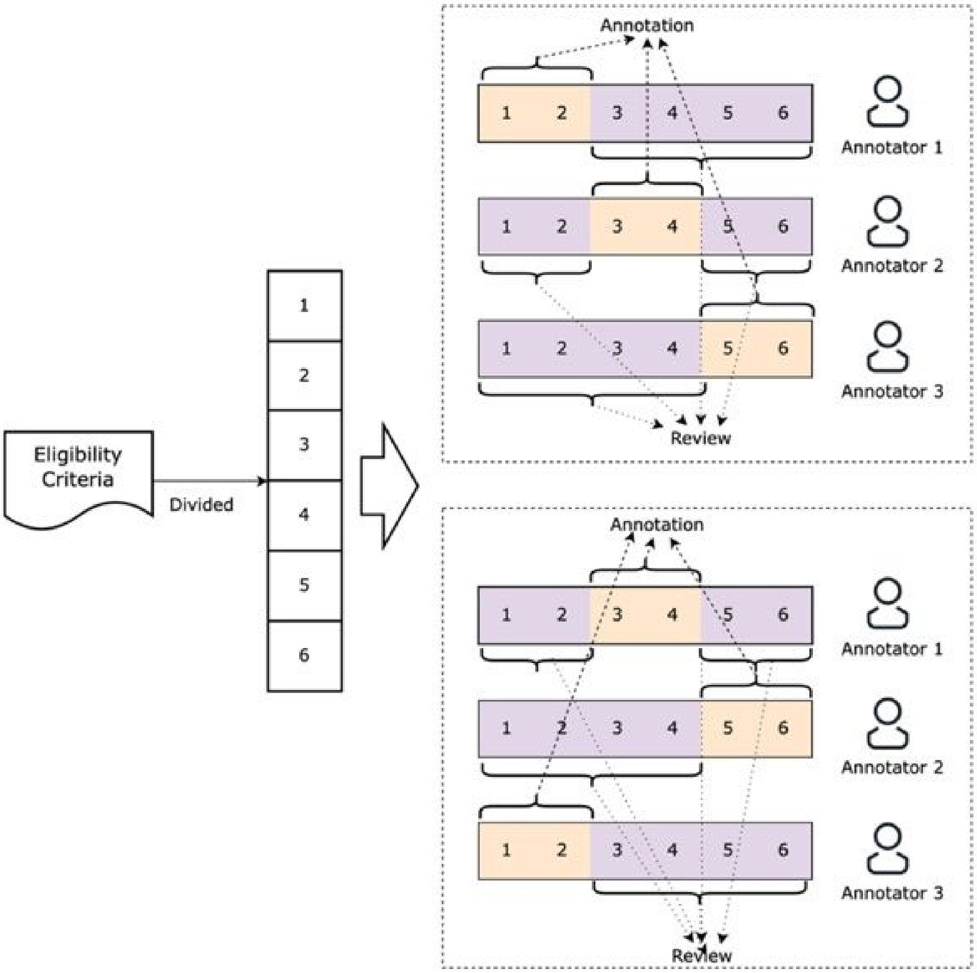
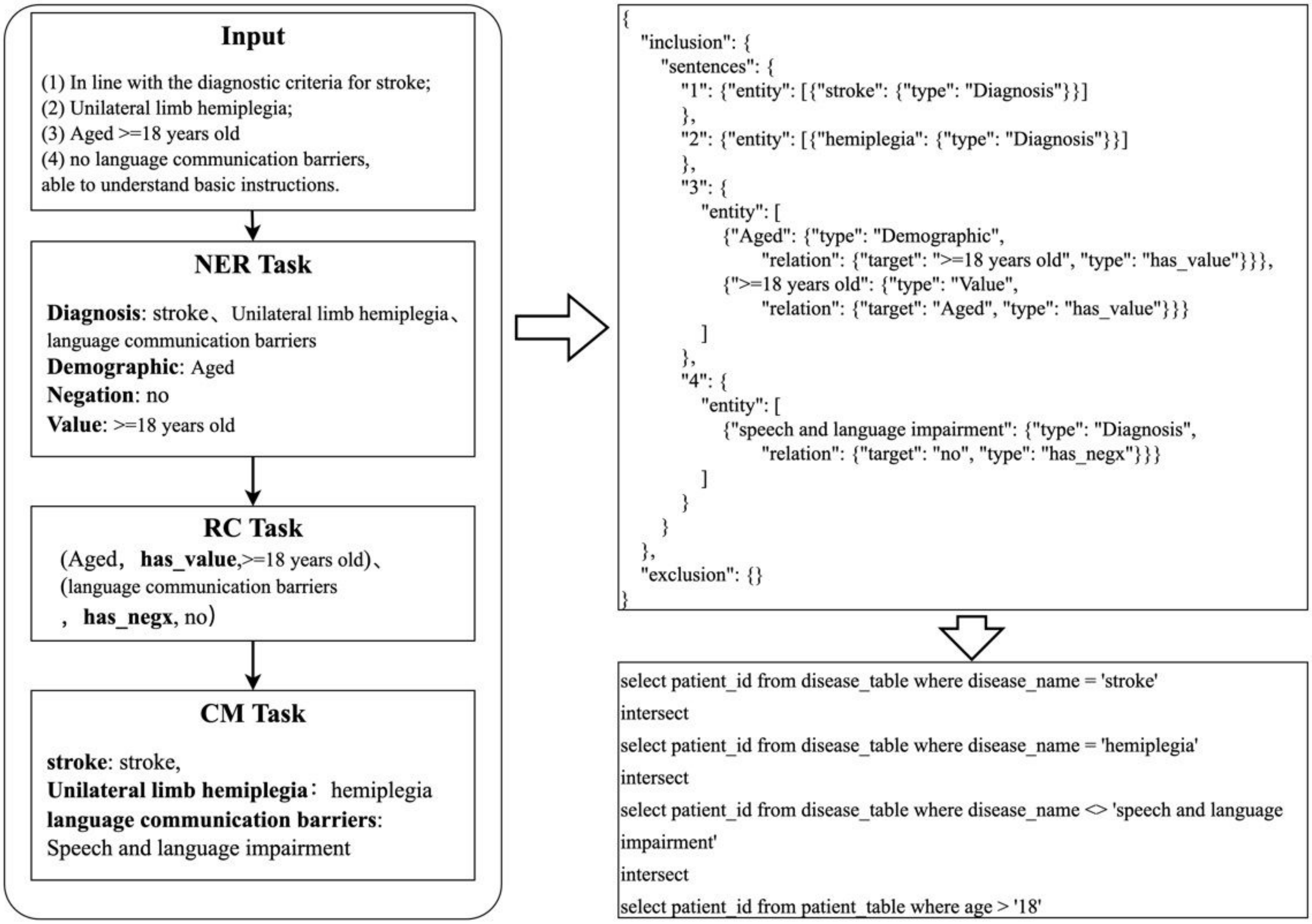


Figure 2

Data annotation process for EC

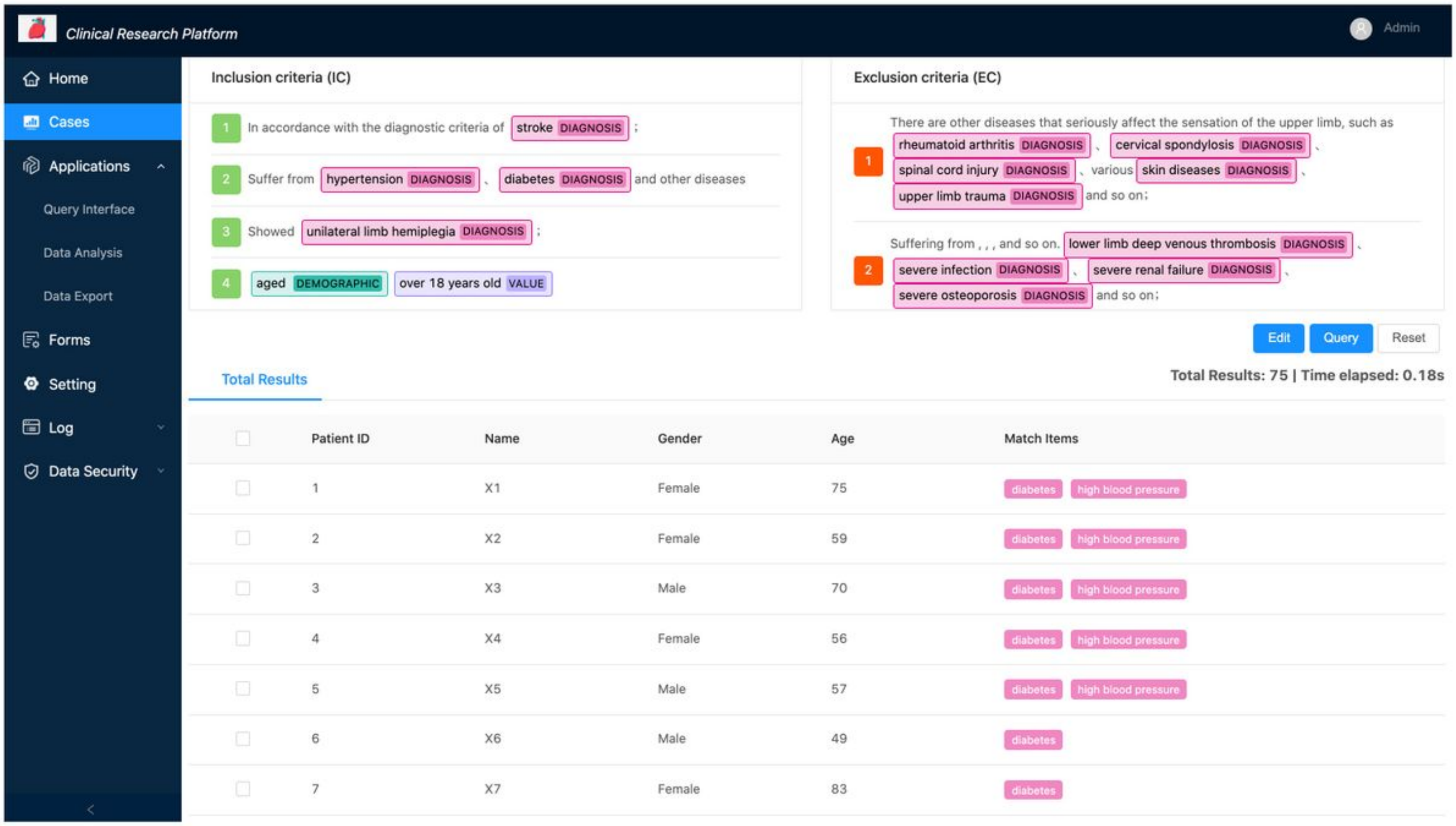

Input
(1) In line with the diagnostic criteria for stroke;
(2) Unilateral limb hemiplegia;
(3) Aged >=18 years old
(4) no language communication barriers,
able to understand basic instructions.
NER Task
Diagnosis: stroke、Unilateral limb hemiplegia、language communication barriers
Demographic: Aged
Negation: no
Value: >=18 years old
RC Task
(Aged，has_value,>=18 years old)、
(language communication barriers
，has_negx, no)
CM Task
stroke: stroke,
Unilateral limb hemiplegia：hemiplegia
language communication barriers:
Speech and language impairment

```
{
  "inclusion": {
    "sentences": {
      "1": {"entity": [{"stroke": {"type": "Diagnosis"}}]
      },
      "2": {"entity": [{"hemiplegia": {"type": "Diagnosis"}}]
      },
      "3": {
        "entity": [
          {"Aged": {"type": "Demographic",
                "relation": {"target": ">=18 years old", "type": "has_value"}}},
          {">=18 years old": {"type": "Value",
                "relation": {"target": "Aged", "type": "has_value"}}}
        ]
      },
      "4": {
        "entity": [
          {"speech and language impairment": {"type": "Diagnosis",
                "relation": {"target": "no", "type": "has_negx"}}}
        ]
      }
    }
  },
  "exclusion": {}
}
```

```
select patient_id from disease_table where disease_name = 'stroke'
intersect
select patient_id from disease_table where disease_name = 'hemiplegia'
intersect
select patient_id from disease_table where disease_name <> 'speech and language
impairment'
intersect
select patient_id from patient_table where age > '18'
```

Figure 3

PP Task



Inclusion criteria (IC)

1 In accordance with the diagnostic criteria of stroke DIAGNOSIS ;

2 Suffer from hypertension DIAGNOSIS 、 diabetes DIAGNOSIS and other diseases

3 Showed unilateral limb hemiplegia DIAGNOSIS ;

4 aged DEMOGRAPHIC over 18 years old VALUE

Exclusion criteria (EC)

1 There are other diseases that seriously affect the sensation of the upper limb, such as rheumatoid arthritis DIAGNOSIS 、 cervical spondylosis DIAGNOSIS 、 spinal cord injury DIAGNOSIS 、 various skin diseases DIAGNOSIS 、 upper limb trauma DIAGNOSIS and so on;

2 Suffering from , , , and so on. lower limb deep venous thrombosis DIAGNOSIS 、 severe infection DIAGNOSIS 、 severe renal failure DIAGNOSIS 、 severe osteoporosis DIAGNOSIS and so on;

Total Results: 75 | Time elapsed: 0.18s

□	Patient ID	Name	Gender	Age	Match Items
□	1	X1	Female	75	diabetes high blood pressure
□	2	X2	Female	59	diabetes high blood pressure
□	3	X3	Male	70	diabetes high blood pressure
□	4	X4	Female	56	diabetes high blood pressure
□	5	X5	Male	57	diabetes high blood pressure
□	6	X6	Male	49	diabetes
□	7	X7	Female	83	diabetes

Figure 4

The query interface used in clinical research