International Journal of Medical Informatics 108 (2017) 85–91

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Systematic data ingratiation of clinical trial recruitment locations for geographic-based query and visualization

Jake Luo[a,b,*], Weiheng Chen[a,b], Min Wu[a,b], Chunhua Weng[c]

[a] Department of Health Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI,United States
[b] Biomedical Data and Language Processing Center, University of Wisconsin Milwaukee, Milwaukee, WI, United States
[c] Department of Biomedical Informatics, Columbia University, New York City, NY, United States

ARTICLE INFO

Keywords:
Geographic visualization
Clinical research informatics
Clinical trial
Geocoding
Interactive map
Geographic data extraction
Data integration

ABSTRACT

Background: Prior studies of clinical trial planning indicate that it is crucial to search and screen recruitment sites before starting to enroll participants. However, currently there is no systematic method developed to support clinical investigators to search candidate recruitment sites according to their interested clinical trial factors.

Objective: In this study, we aim at developing a new approach to integrating the location data of over one million heterogeneous recruitment sites that are stored in clinical trial documents. The integrated recruitment location data can be searched and visualized using a map-based information retrieval method. The method enables systematic search and analysis of recruitment sites across a large amount of clinical trials.

Methods: The location data of more than 1.4 million recruitment sites of over 183,000 clinical trials was normalized and integrated using a geocoding method. The integrated data can be used to support geographic information retrieval of recruitment sites. Additionally, the information of over 6000 clinical trial target disease conditions and close to 4000 interventions was also integrated into the system and linked to the recruitment locations. Such data integration enabled the construction of a novel map-based query system. The system will allow clinical investigators to search and visualize candidate recruitment sites for clinical trials based on target conditions and interventions.

Results: The evaluation results showed that the coverage of the geographic location mapping for the 1.4 million recruitment sites was 99.8%. The evaluation of 200 randomly retrieved recruitment sites showed that the correctness of geographic information mapping was 96.5%. The recruitment intensities of the top 30 countries were also retrieved and analyzed. The data analysis results indicated that the recruitment intensity varied significantly across different countries and geographic areas.

Conclusion: This study contributed a new data processing framework to extract and integrate the location data of heterogeneous recruitment sites from clinical trial documents. The developed system can support effective retrieval and analysis of potential recruitment sites using target clinical trial factors.

1. Background

Clinical trials are considered the gold standard for validating the efficacy and effectiveness of health care treatment, but unfortunately clinical trials are expensive and time consuming. The total expenditure on clinical trials in the United States was estimated at over $35 billion per year [1]. It was also estimated that clinical research accounted for at least one-third of the expenditure of the NIH, and a large portion of the budget was spent on clinical trial studies [2]. Expenditures for the development of new treatments has continued to grow in the United States [1,3]; however, new drug developments have not keep pace with the rising expenditures [3]. Some studies [4,5] argued that the slow drug development could be related to the increasing cost of trial recruitment, low participant rate, and insufficient enrollment. To enroll more patients, many research agencies and companies carried out multi-center clinical trials to expand recruitment and participation. There has been a trend of carrying out more and more multi-center trials in multiple countries. However, a study of trial recruitment for the time period of 2007–2010 showed that over 60% of the planned recruitment sites enrolled could not enroll more than one hundred patients, and close to 15% of the sites could not recruit a single patient [6]. The results of these studies indicate that there is a significant waste of resources and time for setting up recruitment sites. Therefore, there is a strong need for evidence-based recruitment sites planning.

* Corresponding author at: Department of Health Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI, United States.

http://dx.doi.org/10.1016/j.ijmedinf.2017.10.003
Received 15 June 2016; Received in revised form 27 September 2017; Accepted 2 October 2017

However, currently there is still a lack of systematic methods for gathering information and evidence that can support early-stage decision making for clinical trial site planning.

Due to the expansion of international multi-center clinical trials and the demand of improving health care research in developing countries, the number of international clinical trials has been increasing steadily. Developed areas, such as North America and Western Europe, continue to conduct many clinical trials [7,8]. The significant expansion of international clinical trial creates the needs of global clinical trial monitoring and management. Therefore, it is desirable to develop effective methods to facilitate the retrieval of clinical trial information to support decision making for policymakers and clinical investigators.

Prior studies discussed the challenges of finding suitable recruitment sites during the planning stage of clinical trials. The Clinical Trials Transformation Initiative (CTTI) [9] is a large public-private partnership that aims to develop novel practices to improve the efficiency of clinical trials. CTTI identified key strategies for clinical trial planning. One of the key strategies of CTTI is to develop novel methods to support recruitment site selection [9]. In a study that analyzed potential factors affecting subject enrollment [10], the investigator confirmed that poor choice of study site was one of the major barriers to patient recruitment and retention. In another study that discussed issues related to recruiting young patients for clinical trials [11], the consensus of the investigator team was that site selection was one of the top five issues associated with subject recruitment. For community-based clinical studies, Potter et al. [12] discussed the challenges of site selection within the National Drug Abuse Treatment Clinical Trials Network (CTN). The investigators argued that past recruitment performance and recruitment site location were two of the most important factors for finding potential recruitment sites. However, currently the recruitment location and performance data are not always easily accessible by clinical investigators when they start planning patient recruitment. Therefore, in this study we aim to address this gap by integrating and formalizing the heterogenous data of 1.4 million clinical trial recruitment sites to construct a map-based geographic information system to support effective search and retrieval of potential sites.

Recently, there has been a significant national and global trend for releasing clinical trial data for public use [13,14]. For example on ClinicalTrials.gov [15,16], the number of registered clinical studies increases from 3968 in 2000 to 254,982 in 2017. The publication of clinical trial data not only improves the transparency of clinical studies, but also provides new opportunities to further enhance the efficiency of clinical research. In this study, we propose a novel approach to integrate a large amount of heterogeneous data of recruitment sites as well as clinical trial factors that have been documented in clinical trial protocols. The integrated data is used to develop a geographic information system to enhance search and visualization of potential recruitment sites. As far as we know, no other studies addressed this need. The outcomes of this study include: 1) Systematically integrating 1.4 million recruitment location data of 183,000 trials; 2) Formalizing clinical trial data elements to enable search of past recruitment sites according to their research focuses, including target conditions and interventions; 3) Visualizing the integrated recruitment data on a map-based geographic information system.

2. Methods

The modularized framework (Fig. 1) shows the process of extracting recruitment locations and clinical trial factors from clinical trial summaries.

2.1. Clinical trial summaries extraction

The clinical trial data used in this study was extracted from ClinicalTrials.gov. ClinicalTrials.gov is one of the largest public registries of clinical studies. We downloaded 183,000 clinical trial summaries from ClinicalTrials.gov in the XML format. A parser was developed to read the data elements from the clinical trial documents. The downloaded data was transformed into the JSON [17] format for cross-trial data integration and analysis. We also extracted several key clinical trial factors from the trial documents, such as trial title, target disease condition, intervention method, and recruitment locations. Target disease conditions are the names of diseases or conditions studied in a clinical trial. Intervention methods are the names of drugs, medical devices, procedures, vaccines, and other medical products studied. Interventions also include noninvasive study approaches, such as surveys, education, and interviews. The target disease conditions and intervention methods are the two key factors for searching potential recruitment sites. Therefore, the extracted trial factors were linked with the recruitment locations during the data integration process. We retrieved all the published data on ClinicalTrials.gov and stored the information in a local database.

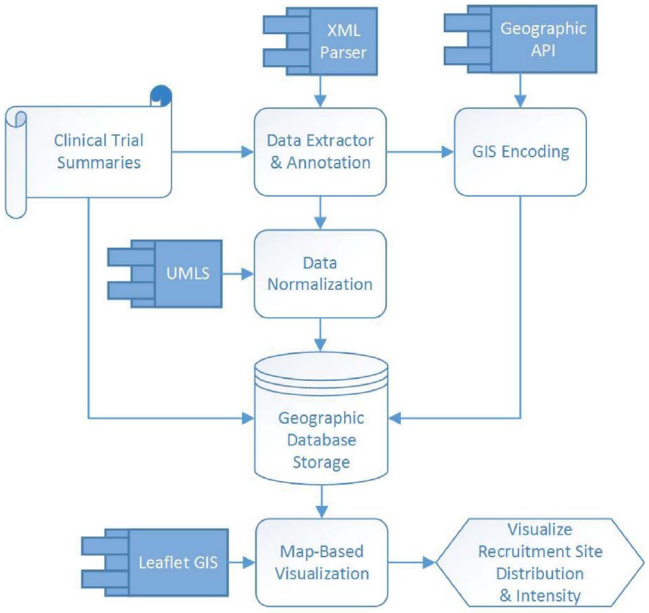

Fig. 1. Data processing framework.

2.2. Data element normalization

Because most clinical trials do not have a uniformly agreed standardized terminology to encode the reported information, there is still a significant heterogeneity gap for the integrating data across trials. This poses a challenge for cross-trial data analysis. For example, different trials could use different terms to describe hypertensive patients, such as “Hypertension”, “Hypertensive disorder”, and “High Blood Pressure”. Another example of this type of disparity can be found in drug names, for example “Propecia” is also known under the name of “Finasteride”. We also found that more than 6 different terminology standards were used in the adverse event reports of trials [18]. Such data disparity creates a barrier for data processing and analysis.

To enable data retrieval and analysis across different clinical trial studies, we developed a data normalization method to synthesize the extracted clinical trial summaries. The Unified Medical Language System (UMLS) [19] was used as the terminology standard for data normalization in this study. UMLS is one of the largest standardized biomedical terminology sources. The Metathesaurus in UMLS contains over one million medical concepts from over 150 controlled terminology sources. Terms in the clinical trial reports were semantically mapped to the UMLS concept unique identifier (CUI). We focused on normalized two clinical trial factors: the target disease and the intervention. These two parameters are the two most important factors of