Published in final edited form as:

Int J Med Inform. 2017 December ; 108: 85–91. doi:10.1016/j.ijmedinf.2017.10.003.

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}, and Chunhua Weng^c

^aDepartment of Health Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI, United States

^bBiomedical Data and Language Processing Center, University of Wisconsin Milwaukee, Milwaukee, WI, United States

^cDepartment of Biomedical Informatics, Columbia University, New York City, NY, United States

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.

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

Conflict of interests

We declare that we have no conflict of interests.

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.

Keywords

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

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. 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 heterogeneous 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](https://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](https://clinicaltrials.gov). [ClinicalTrials.gov](https://clinicaltrials.gov) is one of the largest public registries of clinical studies. We downloaded 183,000 clinical trial summaries from [ClinicalTrials.gov](https://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](https://clinicaltrials.gov) and stored the information in a local database.

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 any clinical trials. A semantic annotator [20,21] was used to map identified terms into standardized UMLS concept unique identifier (CUI) and semantic type (ST). For example, the above-mentioned hypertensive terms were uniformly mapped to “CUI: C00020538” and “ST: Disease or Syndrome”, and the Propecia drug names were mapped to concept “CUI: C0722858” and “ST: Pharmacologic Substance.”

The semantic annotator was created using a UMLS-based semantic lexicon that was created from 10,000 clinical trial patient characteristics in a prior study [20]. The lexicon contains medical terminologies and their unambiguous mapping to semantic types. We used a preference rule-based approach [22] to create mappings between the terms and semantic types. The semantic annotator was used to search the clinical trial reports and identified medical concepts that were recognizable by UMLS. The identified terms that share the same UMLS concept were grouped by the concept and assigned with a preferred name. The normalized disease names, intervention names, and sponsors were linked to the geographic locations of recruitment sites using the unique clinical trial identifiers (NCT-ID) provided by [ClinicalTrials.gov](https://clinicaltrials.gov). Therefore, a clinical investigator can use the linked information to systematically search and screen potential recruitment sites.

2.3. Geographic encoding

The trial recruitment addresses were not properly normalized and structured; therefore, the textual representation of addresses could not be used directly for integrating and comparing trial locations. To solve this problem, we developed a geographic encoding method to assign standardized geographic attitude and longitude coordinates to the recruitment locations. This geographic encoding is also called a geocoding process. Two major geocoding APIs, including Google Geocoding API and Bing Map API, were compared and tested to select a geocoding service. The test used 10,000 recruitment address strings. In the test result, it was observed that Google API had a stricter limit of query requests (2500 free requests per day), while the Bing API was able to process 12,500 requests. We manually examined 100 locations to evaluate the coding correctness. The Bing API was also slightly better at 98% than the 97% of the Google API. Based on these test results, the Bing API was selected to support the geocoding process. The extracted recruitment addresses were encoded with the geographic coordinates. After that, we assigned the mapped geographic locations to the recruitment sites. For example, trial location “Kings College Hospital Denmark Hill, London United Kingdom SE6 9RS” was encoded as: Country “United Kingdom”, City “London”, Road “Denmark Hill”, Postal code “SE6 9RS”, Latitude “51.46979”, and Longitude “-0.093550.” The geocoding process normalized recruitment site locations. This enables the comparison of recruitment sites across trials. For example, the normalized addresses can be used to examine whether two documented recruitment sites are at the same location. We can also use the geocoded information to precisely calculate the geographic distance between any given recruitment sites.

2.4. Map-based visualization

After geographic encoding, we extracted over 1.4 million recruitment locations. Since all the recruitment sites can be formally represented as geographic locations, the most intuitive way to allow users to effectively search and interact with the recruitment locations is to integrate and visualize the information on a map-based geographic platform. Therefore, the retrieved geographic coordinates were imported into the Leaflet map system for visualization. Users can interact with the map to further specify the clinical trial factors for searching recruitment sites. For example, a user can select a target recruitment condition (e.g. Hypertension, Ebola, or Asthma) or intervention (e.g. Insulin, Cefuroxime, Bupivacaine hydrochloride) to retrieve and visualize the clinical trial recruitment locations. Each individual recruitment site can be selected on the map. The user can see what trials had been conducted on the selected site. A user can also select any geographic areas to visualize the trial recruitment locations and analyze the distribution patterns.

3. Result

We first evaluated the coverage of extraction and geocoding mapping. The correctness of the geocoding normalization was evaluated using randomly selected samples. To demonstrate the capacity of using the developed method to aggregate a large amount of clinical trial recruitment locations to support cross-trial analysis, we conducted an analysis to summarize the frequencies of recruitment sites on the country-level. Third, to demonstrate the ability to search and visualize clinical trial recruitment locations according to different target disease

conditions, we compared the top recruitment locations of three different target conditions, including diabetes, breast cancer, and malaria. The differences of geographic distribution of the three disease conditions were quantified using the Euclidian Distance.

3.1. Coverage and correctness of recruitment location extraction and geocoding

Over 1.4 million recruitment locations were extracted and normalized. A small set of recruitment sites (0.2%) could not be encoded due to various problems, including misspellings of addresses, incomplete inputs, or unfound addresses. Therefore, the *coverage* of geocoding for all the extracted recruitment locations was calculated as:

$$\text{Coverage} = \#(\text{sites with coordinate}) / \#(\text{total documented sites}) = (1,424,770 - 2091) / 1,424,770 = 99.8\%$$

To evaluate the correctness of the location mapping, we randomly retrieved two sets of sample recruitment locations (100 records per set) from the clinical trial data. Two investigators (JL and WC) manually examined the sample data to check whether the encoded geographical coordinates correctly pointed to the actual addresses (see Table 1). The *correctness* rate was calculated as the percentage of the retrieved sites that were correctly encoded:

$$\text{Correctness} = \#(\text{sites correctly matched}) / \#(\text{all retrieved test sites}) = 193 / 200 = 96.5\%.$$

The average correctness rate of geocoding was 96.5% for the 200 randomly sampled sites. The errors were mainly due to several reasons, including incomplete inputs of addresses, and typographical errors.

3.2. Country-level recruitment site frequency ranking

(Table 2) To demonstrate the capacity of retrieving clinical trial recruitment locations systematically for cross-trial analysis, we queried the recruitment locations across all the integrated trial data. The aggregated results of the top 30 countries are shown in Table 2. The recruitment intensity of a country is defined as the percentage of trials that recruited participants in the country. The United States was the number one country regarding recruitment intensity based on the [ClinicalTrials.gov](https://clinicaltrials.gov) data. For other countries, the reported trials were more evenly distributed. As Table 2 depicts, the developed Europe countries also held many recruitment locations. Sixteen of the top thirty countries were located in the Europe region. In recent years, eastern Europe countries (e.g. Poland and the Czech Republic) have become popular places to conduct clinical trials. Some developing countries, such as China, Brazil, and India, were the emerging nations for clinical trial research. These results demonstrate that the integrated data of clinical trial recruitment locations can be effectively used to support cross-trial recruitment pattern analysis.

3.3. Map-based visualization

A map-based visualization approach was developed to support effective browsing and selection of recruitment sites. Fig. 2 shows all the recruitment centers when an investigator searches the target condition “diabetes”. On the interactive map, a user can choose to search

one or more target disease conditions from over 6000 target conditions. The map also provides a statistical summary for the given the disease condition. The summary includes the total number of clinical trials in the area, total number of recruitments, and sponsors. When searching for information about an interested recruitment site, a user can enlarge the map and click on the recruitment site. A popup interface will show the past recruitment data of this site, such as the number of recruited patients and trial intervention methods.

To further demonstrate the ability to search recruitment sites based on target clinical trial conditions, Table 3 shows the results of the top 3 recruitment cities for the target disease conditions, including diabetes, breast cancer, and malaria. The most frequently used recruitment facilities and top sponsors are also listed in the third and fourth column of Table 3. For trials that were focused on breast cancer, Boston (total 783 facilities), New York (767), and Seattle (668) were the most active cities for recruitment. The most frequently used facilities in Boston were Dana-Farber Cancer Institutes, Beth Israel Deaconess Medical Center, and Massachusetts General Hospital. The top trial sponsors in Boston were Dana-Farber Cancer Institute, National Cancer Institute, and Massachusetts General Hospital. For diabetes trials, the top three city locations were San Antonio (928), Dallas (839), and Miami (776). Novo Nordisk Clinical Trial Center held many recruitment sites in these top cities. Novo Nordisk was also one of the top trial sponsors. The “local research sites” were the recruitment locations with addresses only, and their formal institutional names were not reported. The top recruitment sites of breast cancer and diabetes trials were located in the United States (US), whereas malaria trials mainly recruited in three different countries, including Bamako in Mali (47), Kisumu in Kenya (32), and Ouagadougou in Burkina Faso (31). The results suggest that the system can be used to search and find recruitment sites based on target conditions. The results also show that the geographic distribution patterns of recruitment locations could varies significantly across different target conditions.

Notice that the names of some of the recruitment sites (third column of Table 3) look similar to the names of sponsors (fourth column of Table 3). This is due to the fact that many clinical trials use the names of their sponsors to name the recruitment locations; nevertheless, in this study the full address of each of the site has been extracted and formalized. The real address and the geographic location can be retrieved from the integrated geographic information system. However, to simplify the display of the searched results, Table 3 only shows the site names, but not the full address of each of the site.

3.4. Visualization and quantitative distribution analysis of recruitment sites

In Fig. 3, the recruitment sites of the three different target disease conditions are retrieved and visualized on an interactive map. The target disease conditions are shown as: asthma (red dots), breast cancer (blue dots), and malaria (green dots). It becomes obvious that on the global scale, the geographic distribution between asthma and breast cancer is similar, and their sites are primarily located in North America, Western Europe, and Northeast Asia. These areas are more economically affluent than other developing areas. As a comparison, the recruitment sites of malaria are located mainly in Africa and Southeast Asia around the tropical Equator zone.

The geographic distribution distance of recruitment sites between any two given conditions can be quantified. We can use the Euclidean distance to calculate the distribution distance between two conditions A and B:

$$\text{Euclidean_Distance (A, B)} = \text{Dis (A, B)} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}$$

where A and B are two different diseases; and a, b represent the corresponding geographic vectors for disease A and B. For the three example diseases, if we set the geographic vectors as the number of recruitment sites per country, their pairwise distances can be quantified as:

$$\text{Dis (asthma, breast cancer)} = 0.265$$

$$\text{Dis (malaria, breast cancer)} = 1.151$$

$$\text{Dis (malaria, asthma)} = 1.149$$

The results confirm that recruitment sites of malaria have a greater Euclidean distance (malaria, asthma: 1.149; malaria, breast cancer: 1.151) to the recruitment sites of the other two diseases, whereas asthma and breast cancer have a closer geographic similarity (asthma, breast cancer: 0.265) in terms of trial recruitment. The results also suggest that there is a significant geographic variance of recruitment sites for different target disease conditions. Therefore, using data-driven analysis, we can better understand the patterns of past recruitment locations.

4. Discussion

In this study, a geographic data integration method was proposed to formalize a large amount of heterogeneous clinical trial recruitment sites and variables. The evaluation results showed that the coverage of the integrated data was 99.8% and the correctness was 96.5%. The integrated recruitment data can be used to support the query of recruitment sites using clinical trial factors, such as target disease conditions. A map-based geographic information system was also developed to support visualization and interaction with the queried results. The system allows clinical investigators to directly search potential recruitment locations using target disease conditions or interventions. The system enables systematic retrieval and analysis of geographic patterns of clinical trial recruitment sites.

4.1. Related studies

There are many applications of geographic data retrieval and visualization in the clinical research and healthcare domain [23–25]. For example, in epidemiology research, a recent study analyzed the global distribution of cervical cancer [26]. The burden of cervical cancer in every country around the world was analyzed and the results were reported using geographic visualization. The study discovered significant variety among different

geographic areas and identified areas that were impacted the most by cervical cancer. In another study [27], the investigators used a geographic map to uncover the distribution patterns and risk factors of lung cancer in China. The study discovered two most significant factors contributing to diagnoses of lung cancer, including the poor quality of air and the associated low annual precipitation. This has been a hot social topic with a lot of attention in China recently. Using geographic analysis to examine access of healthcare, Chan et al. [28] found that rural residents have to travel a longer distance to access healthcare services when compared to their urban counterparts, and this could lead to a lower quality of care for rural residents. By combining the analysis of socioeconomic and geographic information, Cutter et al. [29] created the Social Vulnerability Index (SoVI) which has become a widely accepted reference to evaluate environmental health hazards. These study cases demonstrate the importance of using geographic information to uncover important patterns related clinical research and healthcare. As shown in our study results, clinical trials also have strong geographic patterns. However, there is a lack of effective approaches to allow clinical investigators and trial regulators to analyze clinical trial geographic. This exploratory study is an important step to address this gap.

On the technology aspect, this study focused on developing novel methods for retrieving, transforming, and visualizing clinical trial data. It addressed the challenge of systematic query and analysis of clinical trial recruitment sites. We developed an interactive map using the Leaflet geographic map service to visualize recruitment information. On the application level, the map-based system is similar to the Center for Disease Control (CDC) geographic visualization system [30] and the Agency for Healthcare Research and Quality (AHRQ) geographic information system [31]. The CDC system focuses on providing data and visualization to analyze environmental risk factors. The AHRQ system provides geographic visualization to assist the analysis of healthcare quality and delivery. Our platform was developed to support the search and analysis of clinical trial recruitment sites. The approach developed in this study is an important step toward building a geographic information system to support clinical trial planning and analysis.

4.2. Limitation and future work

The main limitations of this study are associated with the choice of the data source. Currently, the data source is extracted from clinical trial reports at [ClinicalTrials.gov](https://clinicaltrials.gov). [ClinicalTrials.gov](https://clinicaltrials.gov) is one of the largest public clinical trial registries around the world. The US Food and Drug Administration (FDA) mandatorily requires all pharmaceutical products in the United States to register their trials on [ClinicalTrials.gov](https://clinicaltrials.gov). The United States currently is the biggest pharmaceutical market of the world; therefore, we would expect to see only a small portion of trials that are not registered on [ClinicalTrials.gov](https://clinicaltrials.gov). Furthermore, the International Committee of Medical Journal Editors (ICMJE) also mandatorily requires clinical trials to register on one of the two endorsed registry platforms, the WHO Registry Platform or [ClinicalTrials.gov](https://clinicaltrials.gov). This is a prerequisite for publication on ICMJE journals. This requirement applies to both US and non-US trials. The ICMJE publication policy is also widely-accepted by many biomedical journals. Therefore, the trials registered on [ClinicalTrials.gov](https://clinicaltrials.gov) are actually diversified and representative, covering different countries around the world. The data in this study now contains 180,000 trials from 180 countries.

Due to the fact that [ClinicalTrials.gov](https://clinicaltrials.gov) is hosted in the United States, there could be a higher percentage of US trials registered on [ClinicalTrials.gov](https://clinicaltrials.gov) than other countries. This could potentially lead to representativeness bias of trials. However, at the moment of this study, [ClinicalTrials.gov](https://clinicaltrials.gov) provides the best quality and quantity of recruitment data that we can find. To analyze recruitment locations, we need to collect detailed recruitment addresses (e.g. country, city, street, postal code, facility name). At this moment, many other registries do not provide data in such details. For example, the WHO registry currently only collects country or state level recruitment locations. Some country-specific registries, such as the Chinese Clinical Trial Registry ([ChiCTR.org.cn](https://www.chictr.org.cn)), also provide detailed recruitment addresses. However, these registries are significantly smaller than [ClinicalTrials.gov](https://clinicaltrials.gov). For example, [ChiCTR.org.cn](https://www.chictr.org.cn) is about 8% of the size of [ClinicalTrials.gov](https://clinicaltrials.gov). Furthermore, many trials on [ChiCTR.org.cn](https://www.chictr.org.cn) are also registered in duplication on [ClinicalTrials.gov](https://clinicaltrials.gov) because of the ICMJE requirements and FDA regulations. In the future, we can collect data from country-specific trial registries to enrich the [ClinicalTrials.gov](https://clinicaltrials.gov) data source.

5. Conclusion

In this study, we proposed a new method to extract and normalize a large amount of heterogeneous clinical trial recruitment data using geocoding and data integration techniques. A map-based geographic information system was developed to query and visualize the integrated recruitment site data. More than 1.4 million clinical trial recruitment sites were extracted and formalized using the data of 183,000 clinical trial summaries. The data of over 6000 clinical trial target conditions and 4000 interventions was also integrated and linked with the recruitment locations. The geocoded locations covered 99.8% of the extracted recruitment sites, and the correctness was 96.5% for the 200 randomly evaluated sites. We also demonstrated that the recruitment locations could be effectively queried using the key clinical trials factors. Recruitment sites of three different target disease conditions were queried and compared. The results showed that recruitment locations could vary significantly across different target diseases. In summary, this study provides an effective geographic data processing approach for clinical investigators and clinical trial regulators to retrieve, analyze and visualize clinical trial recruitment sites.

Acknowledgments

We thank College of Health Sciences at the University of Wisconsin Milwaukee (UWM) for providing a SEED grant for this project, and UWM Research Foundation for supporting the Center for Biomedical Data and Language Processing.

References

1. Adams CP, Brantner VV. Spending on new drug development. *Health Econ.* 2010; 19(2):130–141. [PubMed: 19247981]
2. NIH. [cited 2015 March 1] National Institute of Health Funding Facts. <http://www.report.nih.gov/fun-dingfacts/fundingfacts.aspx>
3. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J. Health Econ.* 2003; 22(2):151–185. [PubMed: 12606142]
4. Dickson M, Gagnon JP. The cost of new drug discovery and development. *Discov. Med.* 2009; 4(22):172–179.

5. Pierre, C. Recruitment and Retention in Clinical Trials: What Works, What Doesn't and Why. Drug Information Association Annual Summit; Philadelphia, PA: 2006.
6. Califf RM, et al. Characteristics of clinical trials registered in clinicaltrials.gov: 2007–2010. JAMA. 2012; 307(17):1838–1847. [PubMed: 22550198]
7. Thiers FA, Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. Nat. Rev. Drug Discov. 2008; 7(1):13–14.
8. Shah S. Globalization of clinical research by the pharmaceutical industry. Int. J. Health Serv. 2003; 33(1):29–36. [PubMed: 12641261]
9. CTTI. Clinical Trial Transformation Initiative, Strategic Recruitment Planning. 2017. Available from: <https://www.ctti-clinicaltrials.org/projects/recruitment>
10. Sullivan J. Subject recruitment and retention: barriers to success. Appl. Clin. Trials. Apr.2004 2014
11. Carlson GA, et al. Methodological issues and controversies in clinical trials with child and adolescent patients with bipolar disorder: report of a consensus conference. J. Child Adolesc. Psychopharmacol. 2003; 13(1):13–27. [PubMed: 12804123]
12. Potter JS, et al. Site selection in community-based clinical trials for substance use disorders: strategies for effective site selection. Am. J. Drug Alcohol Abuse. 2011; 37(5):400–407. [PubMed: 21854283]
13. EMA. European Medicines Agency: Release of data from clinical trials. 2013. Available from: http://www.ema.europa.eu/ema/pages/special_topics/general/general_content_000555.jsp
14. GSK. GSK Clinical Study Requests. 2016. Available from: <http://www.clinicalstudydatarequest.com/>
15. NLM. About ClinicalTrials.gov. <http://www.clinicaltrials.gov>. 2016; Available from: <http://www.webcitation.org/6jT37msHF>
16. Zarin DA, Keselman A. Registering a clinical trial in ClinicalTrials. gov. Chest J. 2007; 131(3): 909–912.
17. Crockford D. The application/json media type for javascript object notation (json). 2006
18. Luo, Z., Zhang, G-Q., Xu, R. Mining patterns among adverse events in clinical trials. AMIA Joint Summit on Translational Science; San Francisco: 2013.
19. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32(Suppl 1):D267–D270. [PubMed: 14681409]
20. Luo, Z., et al. Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS. AMIA Joint Summit of Translational Informatics; San Francisco: 2010. p. 26–31.
21. Luo Z, Miotto R, Weng C. A human–computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. J. Biomed. Inf. 2012; 44(1):33–39.
22. Johnson SB. A semantic lexicon for medical language processing. J. Am. Med. Inf. Assoc. 1999; 6(3):205–218.
23. Yusuf S, et al. Global burden of cardiovascular diseases part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. Circulation. 2001; 104(23):2855–2864. [PubMed: 11733407]
24. Dusheiko G, et al. Hepatitis C virus genotypes: an investigation of type-specific differences in geographic origin and disease. Hepatology. 1994; 19(1):13–18. [PubMed: 8276349]
25. Trifonov V, Khiabanian H, Rabadan R. Geographic dependence: surveillance, and origins of the 2009 influenza A (H1N1) virus. New Engl. J. Med. 2009; 361(2):115–119. [PubMed: 19474418]
26. Arbyn M, et al. Worldwide burden of cervical cancer in 2008. Ann. Oncol. 2011; 22(12):2675–2686. [PubMed: 21471563]
27. Lin X-L, et al. Geographic distribution and epidemiology of lung cancer during 2011 in Zhejiang province of China. Asian Pac. J. Cancer Prev. APJCP. 2014; 15(13):5299. [PubMed: 25040992]
28. Chan L, Hart LG, Goodman DC. Geographic access to health care for rural medicare beneficiaries. J. Rural Health. 2006; 22(2):140–146. [PubMed: 16606425]
29. Cutter SL, Boruff BJ, Shirley WL. Social vulnerability to environmental hazards. Soc. Sci. Q. 2003; 84(2):242–261.

30. Croner CM, Sperling J, Broome FR. Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Stat. Med.* 1996; 15(18):1961–1977. [PubMed: 8888488]
31. Ricketts, TC. Agency for Health Care Policy and Research. Dept. of Health and Human Services, US Public Health Service; 1997. *Using Geographic Methods to Understand Health Issues.*

Summary points

Already known

Clinical trial studies are the foundation of medical advances but are very expensive and time consuming.

Research shows that over 60% of the planned trial sites recruited less than 100 patients and 15% of the sites never recruited anyone.

Effective methods for supporting recruitment site selection for clinical trials are scarce.

Geographic data retrieval and visualization can be an effective approach to support clinical research and healthcare service applications.

This study has added

We design a method for geographic information retrieval and visualization to enable clinical investigators to search and query geographic distribution of clinical trial recruitment.

We developed a data-driven informatics approach that integrated over 1.4 million clinical trial recruitment sites for over 183,000 trials.

We evaluated that the geography location extraction covered 99.8% of the recruitment locations, and the geocoding correctness is about 96.5% in 200 randomly selected locations.

We demonstrated that clinical trial recruitment sites show distinctive geographic patterns, which is critical for trial planning.

Fig. 1.
Data processing framework.

Fig. 2.
Example of diabetes recruitment centers visualization.

Fig. 3.
The global recruitment sites enrolled patients with asthma, breast cancer, and malaria patients.

Table 1

Recruitment location evaluation.

	Correct	Incorrect	Not-found	Correctness
Set #1	98/100	1/100	1/100	98.0%
Set #2	95/100	5/100	0/100	95.0%
Total	193/200	6/200	1/200	96.5%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Top 30 countries ranked by their recruitment site counts.

Rank	Country	#Trial	# Sites	%Trial	Rank	Country	#Trials	# Sites	%Trial
1	United States	81148	681987	31.6%	16	Poland	3817	20383	1.5%
2	Canada	13084	49937	5.1%	17	Taiwan	3767	8130	1.5%
3	Germany	12222	92552	4.8%	18	Switzerland	3715	7585	1.4%
4	France	11370	81351	4.4%	19	Sweden	3706	10278	1.4%
5	United Kingdom	9760	38695	3.8%	20	Japan	3406	44942	1.3%
6	Italy	7253	37682	2.8%	21	Austria	3282	8538	1.3%
7	Spain	6760	36253	2.6%	22	Russian Federation	2862	20181	1.1%
8	Korea, Republic of	5794	17474	2.3%	23	Czech Republic	2662	12809	1.0%
9	Netherlands	5738	17276	2.2%	24	India	2539	11635	1.0%
10	China	5503	20013	2.1%	25	Norway	2431	5221	0.9%
11	Belgium	5496	19537	2.1%	26	Hungary	2417	11672	0.9%
12	Israel	4944	10622	1.9%	27	Mexico	2220	7973	0.9%
13	Denmark	4619	9329	1.8%	28	South Africa	1942	7879	0.8%
14	Australia	4399	18181	1.7%	29	Finland	1935	5115	0.8%
15	Brazil	4148	12519	0.016	30	Argentina	1822	9014	0.007

Table 3

Top cities, sites, and sponsors that recruited trial participants.

Condition	Top Cities that Hosted Recruitments (Total Hosted Facility Counts)	Most Frequently Used Sites in the City (Site Frequency)	Top Sponsors (Sponsor Frequency)
Breast Cancer	Boston, MA, US (783)	Dana-Farber Cancer Institute (126)	Dana-Farber Cancer Institute (60)
		Beth Israel Deaconess Medical Center (95)	National Cancer Institute (NCI) (39)
		Massachusetts General Hospital (90)	Massachusetts General Hospital (25)
	New York, NY, US (767)	Memorial Sloan Kettering Cancer Center (116)	Memorial Sloan Kettering Cancer Center (145)
		Columbia University Medical Center (38)	National Cancer Institute (NCI) (34)
		Mount Sinai Medical Center (28)	New York University School of Medicine (23)
	Seattle, WA, US (668)	Fred Hutchinson Cancer Research Center (53)	University of Washington (43)
		Group Health Central Hospital (51)	Southwest Oncology Group (24)
		University Cancer Center at University of Washington Medical Center (48)	NSABP Foundation Inc. (18)
Diabetes	San Antonio, TX, US (928)	Novo Nordisk Clinical Trial Center (96)	AstraZeneca (75)
		GSK Investigational Site (78)	Novo Nordisk (59)
		Local Research Site (55)	Pfizer (39)
	Dallas, TX, US (839)	Novo Nordisk Clinical Trial Center (169)	Novo Nordisk (72)
		GSK Investigational Site (52)	Eli Lilly and Company (44)
		Local Research Site (35)	AstraZeneca (41)
	Miami, FL, US (776)	Novo Nordisk Clinical Trial Center (101)	Novo Nordisk (48)
		GSK Investigational Site (38)	AstraZeneca (45)
		Local Research Site (33)	Merck Sharp & Dohme Corp (40)
Malaria	Bamako, Mali (47)	Malaria Research and Training Center (17)	National Institute of Allergy and Infectious Diseases (NIAID) (22)
		University of Bamako (12)	Pfizer (3)
		Pfizer Investigational Site (4)	University of California San Francisco (2)
	Kisumu, Kenya (32)	GSK Investigational Site (11)	Centers for Disease Control and Prevention (6)
		Kenya Medical Research Institute (5)	GlaxoSmithKline (5)
		CDC KEMRI Research Institute (2)	Pfizer (2)
	Ouagadougou, Burkina Faso (31)	Centre National de Recherche et de Formation sur le Paludisme (7)	GlaxoSmithKline (6)
		GSK Investigational Site (5)	London School of Hygiene and Tropical Medicine (5)
		Pfizer Investigational Site (2)	Gates Malaria Partnership (3)