CLIPS: An interactive retrieval system for clinical trial studies with context-dependent protocol elements

Abstract

Background: Clinical trial protocol includes all the steps that should be conducted in clinical trial and all clinical trial start with development of the protocol. And the protocol is kinds of procedure manual that consists of appropriate combination of related elements, so one element selection generally affects the next element selection.

To development the protocol, researchers refer a published guides or electronic protocol templates. The detail elements of the protocol are designed based on the previous data which is retrieved separately in public database of clinical trial. Specifically, researchers retrieve for preliminary information using text based user input keywords, then they extract the elements. However, the current retrieve approaches do not provide a method to interactively select for a combination of the elements.

Objective: The purpose of our study is to provide context-dependent protocol element selection system for the development of objective and successful clinical trial protocols. In detail, we construct a database that can retrieve protocols by combined analysis of elements. Moreover, we develop a web-based interactive protocol element wise selection application using constructed database.

Methods: We have constructed a key-value type database for searching element combinations. To build the database, we have defined structure of protocol within
five factors; design, subject, variables, statistical issue and descriptions. In this study, five factors are defined as a subset of protocol elements and we have manually classified them from collected information in public database. Then, we have developed a web application to implement interactive method for retrieving selected combination of protocol element. The application in the form of a connected tree provides options to select the next element according to the decision of previous element. Therefore, a researcher can retrieve the structure of the combined protocol. Also, the application supports a function for retrieve various selected protocol structure at the same time.

**Results:** We have developed a database and search application for protocol structure retrieval. The database is built on individual protocol information extracted from previous 184,634 clinical trials and provides 13,210 integrated structural information. Furthermore, the database contains various semantic information of the protocols to filter protocols in search application. We did technical validation for evaluating of the database with ‘Cancer and Other Neoplasms’ category of clinicaltrials.gov. By comparing with the clinicaltrials.gov our method has better performance in predicting phenotypic features. Our F1 score was 0.515, while F1 score of clinicaltrials.gov was 0.377.

Finally, we developed web application, CLinical trial protocol database System (CLIPS), which provides users with search our database interactively based on protocol elements. CLIPS is available at [http://corus.kaist.edu/clips](http://corus.kaist.edu/clips).

**Conclusions:** We have developed database and application to interactively search clinical trial protocols efficiently. We believe that our system is helpful for conducting a new clinical trial. Furthermore, we expect that the database would be utilized as meta-analysis in clinical trials for various purposes.

**Keywords:** Clinical Trial; Clinical Trial Protocol; Information Retrieval;

---

**Introduction**

It is widely known that clinical trial protocols play primary role in clinical trials [1]. Well-established protocols simplify clinical procedures, help to avoid unnecessary protocol amendments, facilitate preliminary assessment of latent issues. Thereby, the protocols contribute to the success of clinical trials, not only by reducing costs but also by improving the performance of clinical trials [2, 3].

In recent years, the importance of optimized protocol design have become increasingly substantial, as the cost of clinical trials has risen and the protocols have become more complex. According to the well-acknowledged investigation conducted in 2016, Drug development cost has skyrocketed in past decades and total capitalized cost per approved new drug has reached $2,558 million. It is notable that the clinical cost per approved new drug has as well increased, reaching $1,460 million [4]. The growth rate of capitalized clinical trial cost was 8.3% per year in last decade. It is also pointed out that highly sophisticated nature of modern drug development resulted in the upward trend in the protocol complexity and work loads [5].

Despite the importance of clinical trial protocols and increasing demand to streamline for improving protocols, contemporary drug development environment
is rather inefficient. Tufts Center of the Study of Drug Development (Tufts CDSS) showed that 57% of analyzed protocols had at least one major amendment, and 45% of these amendments were caused by “avoidable” reasons originated from imperfect protocol design, such as design flaws or recruitment difficulties [6]. Moreover, the number of amendments and changes per amendment was both concentrated during phase III of the trials, which leads to larger impact cost [7].

As the demand for method to build better protocol design is evident, there have been various approaches to aid protocol design. The approaches can be categorized into three groups; (1) experts’ guidelines, (2) automated systems, and (3) database system.

Firstly, researchers can refer to experts’ guidelines for protocol design to build their own trial protocols [8]. While referring the expert’s opinion is an effective approach to guarantee the credibility of the protocol design, two obvious limitations follows; it can be applied only if the credible guideline exists in particular clinical field; also, these guidelines tend to not offer specific values for every elements of trial design. Determining these specific elements consequently relies on individual researcher’s subjective intuition.

There has been alternative approaches to build automated system for protocol design; A context-aware architecture for clinical trial protocol design composed of decision support module and semantic search engine [9]. While the idea of constructing automated system was innovative approach, the system had shown limited performances, and the web-based service has become unavailable now.

In the best of our knowledge, the most promising alternative approach is utilizing database systems for clinical trial protocols. Current databases contains extensive number of conducted clinical trials that cover wide range of clinical fields [10, 11]. Researchers can search specific clinical trials on the database according to their purpose. However, current clinical trial databases offer only limited support to search clinical trial protocol. The current databases focus on only providing clinical trial information with simple text search, not offering structured data capable of query-based protocol search to apply context-dependent protocol elements.
To solve the limitations of current database system, we suggest our Clinical Trial Protocol Database System (CLIPS). We developed database enables semantic search for core contents of clinical trial protocols, along with semantic filterable features and frame structures for the protocols. Furthermore, our system based on the database efficiently finds clinical trial protocols by query refinement method. To resolve difficulty to find specific protocols from the database of complex structures, we developed graph based querying system based on the database (Figure 1).

**Methods**

**Clinical Protocol Database**

A clinical trial protocol contains a structure of a clinical trial. The protocol is composed of various elements and the elements of the protocol are able to cluster into key factors [12]. Consequently, we defined key factors as five clusters on previous baseline research [13]; design, subject, variables, statistical issues and descriptions. Design is fundamental decisions about how the trial is structured and modelling to measure data in the trial for evaluating effects. Subject is recruitment decisions about who is eligible to participate the trial and how they are treated to represent generalizability on target population. Variables are about what measurement will be made to verify efficacy or safety. Statistical issues are methods about how the clinical trial will be analyzed and it specify sampling procedure or statistical significance for variables. Finally descriptions about practice organization, a step of phase, additional explanation of the protocol or trial itself.

| Table 1. Key factors and its elements included in CLIPS |
| --- | --- | --- |
| **Key factors** | **Categorical Type Elements** | **Value Type Elements** |
| Design | type, model, allocation, time perspective, masking | number of groups, design group, group label, |
We selected and clustered elements from Aggregate Analysis of ClinicalTrials.gov (AACT) [11] which was released at March 27, 2015. We downloaded dump file of the AACT database and completely overhauled the loaded database on local database. The total number of columns and tables are 270 and 42, respectively. We conducted to classify data types of columns into four types; categorical type, value type, description type and not union type (N/U). Categorical type contains nominal data and data in the type would be substituted for composing frame structure of a protocol. Value type included interval data, ratio data which would be important values in key factors. Description type is additional explanation text, numeric values, abbreviated words or dates for description factor. N/U type is primary keys, foreign keys and database management values. Based on this type classification, we selected categorical type and value type, then we cluster them into design, subject, variable and statistical issue factor on above criteria (Table 1).

Table 2. Table schema for clinical trial protocol

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinicalTrialID</td>
<td>Char(10)</td>
</tr>
<tr>
<td>Design</td>
<td>JSON</td>
</tr>
<tr>
<td>Subject</td>
<td>JSON</td>
</tr>
<tr>
<td>Variable</td>
<td>JSON</td>
</tr>
<tr>
<td>Statistical Issue</td>
<td>JSON</td>
</tr>
<tr>
<td>Description</td>
<td>JSON</td>
</tr>
</tbody>
</table>
We designed table schema and make data compilation progress. N/U is eliminated because we constructed relational table with key-value attribute and it is not necessary on the attribute. Utilization of key-value attribute make possible to search skeleton structure of a clinical trial protocol efficiently on contained dependent elements and it can effectively manage data storage for deploying inconsistent data [14]. We designed table schema with this aspects (Table 2). The next step was to data compilation. We organized values of elements by resolving typos, reflecting dependency structure, removing some control characters and type conversion. Furthermore, we removed ambiguous design types which are null, expanded access and observational(patient registry) to search specific clinical trial protocol clearly. As a result, we could collect 184,634 clinical trial protocols, more detailed information of the protocol table is described in Table 2. The result of the works is, in addition, used to optimize query refinement to retrieve protocols.

**Semantic Filtering Feature Generation**

In spite of, we developed clinical trial protocol database with frame structure of a protocol, we cannot not claim that the detailed contents of trials are similar because the frame structures are similar. Medical Subject Heading (MeSH) can be a candidate solution to solve the problem. MeSH is used for indexing and cataloging of clinical trials in ClinicalTrials.gov/AACT [11, 15]. MeSH has, however, coverage limitation that is not fulfilled across the spectrum of various biomedical terminologies [16]. Consequently, we proposed to generate various biomedical semantic features to find or filter similar clinical trials in the searched structure.

We generated semantic filterable features related to the conditions, interventions which are considered as significances in clinical trials, for the subdivided semantic similarity search. Condition is being studied in clinical trial and condition includes disease, disorder, symptoms which are the set of observable characteristics as disease specific phenotype. Drugs commonly refer to intervention that is the focus of a clinical trial. Drugs involve chemical compounds [17]. And searching or filtering similar clinical trials can be conducted through each of the corresponding elements. In addition, gene similarity is promising method to search similarities on chemical compounds and the phenotypes [18, 19]. As a result, we did named entity recognition (NER) of phenotype, chemical compound, gene for enabling semantic search as semantic filters on description elements; brief title, official title, brief summary, detailed description, keywords, conditions.

**Phenotype**

We generated semantic features to represent disease specific phenotype words. Unified medical language system (UMLS) is a repository of integrated biomedical terminologies and we used UMLS2015AB version to process phenotype words [16]. To NER on descriptive values, we applied Metamap 2016 and cTakes 3.2.2 to the values [20, 21]. We combined each results and removed duplicated results with above tools to synthesize the advantages [22], then we selected 11 semantic types which are considered as disease phenotypic types and we removed other types from the result (Table 3). As a result, we generated disease phenotypic features by Concept unique IDs per clinical trial.
### Table 3. Selected phenotypic types from UMLS

<table>
<thead>
<tr>
<th>Entity Type ID (TUI)</th>
<th>Entity Type Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>T038</td>
<td>Biologic Function</td>
</tr>
<tr>
<td>T039</td>
<td>Physiologic Function</td>
</tr>
<tr>
<td>T041</td>
<td>Mental Process</td>
</tr>
<tr>
<td>T019</td>
<td>Congenital Abnormality</td>
</tr>
<tr>
<td>T020</td>
<td>Acquired Abnormality</td>
</tr>
<tr>
<td>T033</td>
<td>Finding</td>
</tr>
<tr>
<td>T034</td>
<td>Laboratory or Test Result</td>
</tr>
<tr>
<td>T046</td>
<td>Pathologic Function</td>
</tr>
<tr>
<td>T047</td>
<td>Disease or Syndrome</td>
</tr>
<tr>
<td>T048</td>
<td>Mental or Behavioral Dysfunction</td>
</tr>
<tr>
<td>T049</td>
<td>Cell or Molecular Dysfunction</td>
</tr>
<tr>
<td>T184</td>
<td>Sign or Symptom</td>
</tr>
<tr>
<td>T190</td>
<td>Anatomical Abnormality</td>
</tr>
<tr>
<td>T037</td>
<td>Injury or Poisoning</td>
</tr>
</tbody>
</table>

**Chemical Compound**

We did NER for chemical compound entities from descriptions with ChemSpot [23]. ChemSpot provides Chemical Abstract Service (CAS) Id and International chemical identifier (InChI), however, it does not provide standard InChIKey. The InChIKey is compacted version of InChI and standard InChIKey is a stable identifier for reflecting in the identifier version designation [24]. Moreover, standard InChIKey is considered as large description what the community believes to be equivalence between compounds in drug discovery [25]. To include the advantage of standard InChIKey to chemical compound entities, we conducted original words of NER processed entities with ChemSpider [26]. Simple search web application programming interface (API) of ChemSpider is exploited, then we could generate chemical compound entities with standard InChIKey, InCh and simplified molecular-input line-entry system (SMILES) notation.

**Gene**

We appended gene entities to semantic filter elements. Moara [27] is gene annotation tool was used for gene NER, considering that Moara is capable of performing both recognition and normalization of gene entities; Moara recognizes entities and their position in the input text, then links the entities to gene IDs of know gene database. Moara provides various pre-constructed machine learning based model for various organism species. We have adopted human oriented model for our task. For gene normalization procedure, we have obtained lists of gene IDs corresponding to each gene entities. A gene ID with highest score was selected and mapped to the recognized gene terms.
Web application development for query refinement

A clinical trial protocol has complex structure and these complexities are increasing [28]. In addition, major issue to pharmaceutical and biotechnology companies is to optimize the complexities [29]. Optimizing is needed to many trials and errors, however, it is unable to proceed many trials due to high cost on clinical trials [30]. Consequently, most of clinicians refer to previous conducted clinical trials to design a protocol and are having difficulty to search protocols since complexities of protocols are increasing. On these circumstance and from query refinement standpoint, we developed CLIPS web application to provide graph querying interface for finding most reliable clinical trial protocols rather than text querying interface that cannot visualize dependency of prior elements affecting protocol structure [31, 32].

We defined categorical type elements as frame structure of protocols. Although we provide default orders of the elements, a user can freely choose an order of the elements. Once a user makes a decision about the order, the user can find varying combinations of protocol elements on graph based search interface. Dependent elements are retrieved from database in real-time, the user can confirm the number of exist protocols on selected elements for reference. The user also search another protocol frame structures after clipping selected structure to user clip pane. To search whole information of selected protocol structures, the user should click clipped protocol, then user can scrutinize a detail of selected protocol and trial information. Furthermore, the user can add semantic filters when searching whole information to reduce or focus on certain areas of clinical trials. Data-flow example of the interface is described in Figure 2.
The backend of the interface is developed on Node.js [33] and the visualization of the interface is manipulated by d3.js [34]. We developed custom functions on d3.js to show each title of elements and protocol counts on user's selection. To implement semantic filter function on user's free-text input, backend engine is connected to representational state transfer (REST) NER API (CORUS) and the API provides NER processed entities with its types for searching relevant entity in database.

Results
As a result, we collected 184,634 clinical trial protocols, 13,210 frame structures of clinical trial protocols, generated 5,765,054 phenotype, 1,151,053 chemical compound, 222,966 gene features for semantic filtering (Table 4). Then, we developed web application system to use the database based on visually searchable clinical trial protocol. We publicly opened data and source code to utilize our result to other clinical researchers. In conclusion, we achieved database system to find efficiently clinical trial protocols for conducting a new clinical trial.

<table>
<thead>
<tr>
<th>Entity Type ID (TUI)</th>
<th>Entity Type Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>T038</td>
<td>Biologic Function</td>
</tr>
<tr>
<td>T039</td>
<td>Physiologic Function</td>
</tr>
<tr>
<td>T041</td>
<td>Mental Process</td>
</tr>
<tr>
<td>T019</td>
<td>Congenital Abnormality</td>
</tr>
</tbody>
</table>

Technical validation
We used the trial information categorized by disease conditions which is collected from Clinicaltrials.gov for evaluation. Since the information is manually curated, not originated from Clinicaltrial database itself, it can be used as a golden standard to evaluate AACT and phenotypic features from CLIPS. We collected the disease conditions and corresponding trial case information from ‘Cancer and Other Neoplasms’ category of Clinicaltrials.gov to construct gold standard dataset. 82,584 trials from 520 conditions were collected. Using the gold standard, we have evaluated the condition information of AACT and phenotypic features from CLIPS (Figure 3). F1-scores of CLIPS phenotypic features marked significantly higher score (f1=0.515). This shows that CLIPS database is technically well-established.
Conclusions
Clinical trial protocols are crucial factor for clinical trials to achieve primary purpose. However, clinical researchers have designed clinical trial using researcher’s individual expertise. It could cause consistency and objectivity problems on clinical trial protocols. To solve these problems, retrieval system for clinical trial protocol is needed.

In this study, we developed clinical trial protocol database system for all clinician to search clinical trial protocol conveniently, and they can download whole database. ([http://corus.kaist.edu/clips](http://corus.kaist.edu/clips)). For instance, they can use our system as following scenarios; (1) Define key components and ordering components (2) Search each dependent components its prior selected components (3) User can clip selected clinical trial protocol and she or he also can select more clinical trial protocol with clipped protocol (4) Search for relevant clinical trial protocol including all information of clinical trial protocol (5) Download the result. We believe that many clinicians can utilize our system to design more reliable clinical trial protocol.

Acknowledgements
This work was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT and Future Planning through the National Research Foundation. Specially, thanks to David Sharpe for cooperation of using ChemSpider APIs and thanks to Seyol Yoon for helping to select phenotypic TUIs from UMLS.

Conflicts of Interest
We declare no competing financial interests
References


