Modeling Participant-Related Clinical Research Events Using Conceptual Knowledge Acquisition Techniques

Philip R.O. Payne, Ph.D.¹; Eneida A. Mendonca, M.D. Ph.D.²; Justin B. Starren, M.D. Ph.D.³

¹The Ohio State University, Department of Biomedical Informatics, Columbus, OH
²Columbia University, Department of Biomedical Informatics, New York, NY
³Marshfield Clinic Research Foundation, Marshfield, WI

Abstract
The active phase of a clinical trial is defined by a protocol schema consisting of participant-related events organized into multiple visits. Current efforts to model protocol schemas in a computable format have focused on high-level abstractions, such as the temporal relationships between visits. However, such approaches do not address the need for a more granular computational model of the individual events that comprise each visit. To address the preceding gap in knowledge, this paper will describe a study in which conceptual knowledge acquisition (CKA) techniques were applied to a corpus of 32 clinical trials protocol documents in order to develop a knowledge collection of common participant-related clinical research events. These techniques identified 7 high-level concepts that could be used as organizing principles in the resulting knowledge collection. Such results confirm the utility of CKA methods in the clinical research domain.

Introduction
The modern conduct of clinical research involves multiple workflows and activities, such as protocol development, participant recruitment, participant treatment and evaluation (i.e., active phase of a protocol), participant tracking, data collection and analysis, data quality assurance and monitoring, and regulatory compliance [1, 2]. Many of these are defined and organized via protocol schemas, which describe participant-related visits and the events that comprise them, as well as the temporal relationships between such visits. A small number of studies have demonstrated that such protocol schemas can be represented at high levels of abstraction in computable formats. Examples include protocol authoring templates [3] that describe major classes of schema concepts, and temporal constraint models that define the relationships between schema visits [4]. Ultimately, the availability of computable protocol schemas will make it possible to apply information technology (IT) to support clinical research. Such IT usage has the potential to increase productivity and the quality of study results, while decreasing the resources required to conduct clinical studies [1, 2, 5, 6]. Examples of computational systems that can utilize computable protocol schemas to increase research productivity are study calendar and participant tracking tools. However, within the relevant literature, a frequently cited gap in knowledge is the absence of systematic models for describing the participant-related events that comprise protocol schema visits, such as a laboratory test carried out in order to collect pharmacokinetics data for an investigational therapy. The study described in this paper attempts to address the preceding gap in knowledge by applying conceptual knowledge acquisition (CKA) techniques [7] in order to abstract and formalize a prototypical, computational knowledge collection of participant-related clinical research events from a corpus of clinical trial protocol documents.

Background
As described in the Introduction, a limited number of studies have examined high level abstraction models for computable clinical trial schemas. A common theme spanning work in this area is the development of knowledge collections containing domain-specific concepts used to define templates or computational workflow models. Such approaches focus on the improvement of human workflow via computational tools capable of replicating expert performance in targeted tasks [7]. In the following section, we will review two such studies that we believe are representative of the current state of knowledge in this domain, and complementary to the computational model of participant-related clinical research events that we have developed.

Protocol Authoring Templates
Nguyen et al. [3] have described the use of protocol design patterns to support authoring environments intended to generate computer executable protocols. These authoring environments use a top-down approach for refining pre-existing libraries of domain-specific design constructs that occur frequently within Phase I-III protocols, and are intended to allow domain experts to express computable protocols in an intuitive manner. The specific design patterns described by Nguyen et al.
were divided into four tiers: 1) scientific design patterns, 2) protocol schema patterns, 3) event patterns, and 4) task patterns. Specific protocol design patterns in their study were abstracted and formalized from a sample of twenty Phase I-III protocols. As part of this abstraction process, 93 protocol-related descriptors or primitives were discovered, of which 39 were deemed to be specific to protocol design. A library of protocol design patterns defined in terms of these 39 primitives was then constructed by empirically re-analyzing the initial twenty protocols. The resulting library of design patterns was described by Nguyen et al. as a “generic, partially specified solution to the problem of modeling common protocol designs”. One major limitation of this work was the lack of a common set of descriptors for protocol-related tasks and events, which are needed to define event and task patterns that comprised their model.

Temporal Abstractions of Protocol Schemas
Weng et al. [4] have reported on the development of a temporal knowledge representation model for clinical trial task scheduling. Their work asserts that the complexities of the temporal models required to formulate a computable protocol schema preclude the use of existing guideline representation methods. Therefore, they developed a temporal abstraction model based upon an ontology of temporal event descriptors specific to the clinical research domain. The resulting ontology was organized around two high-level concepts: ‘time object’ and ‘time offset’. The latter concept is specifically designed to allow for the representation of temporal interdependency between protocol related visits. Using this ontology, Weng and colleagues proposed a three-stage process for coding and representing protocol schema:
1) Define “anchor”, or absolute time points upon which relative other time points are dependent.
2) Define time patterns or sequences of interrelated time points defined by the preceding “anchor points” and time intervals.
3) Associate protocol-related visits with the preceding time patterns.

In their study, Weng et al. validated their proposed approach to the temporal coding of protocol schemas by accurately (determined by expert evaluation) representing a random sample of twelve Phase I-III clinical protocols. However, similar to the preceding work of Nguyen et al., Weng and colleagues also cite as one limitation to be the lack of a systematic model for describing the specific events to which their temporal model is intended to be applied.

Methods
In order to address the previously introduced gap in knowledge concerning the systematic computational representation of clinical research participant-related events, we have employed a CKA-based technique for abstracting such commonly occurring concepts from a corpus of clinical trial protocol documents and subsequently formalizing them as a computable knowledge collection. The specific three-stage methodology is summarized in Figure 1.

Figure 1: Overview of the three-stage process used to create conceptual knowledge collection consisting of common clinical trial participant-related events.

I. Concept Extraction and Terminology Mapping
A convenience sample of Phase I-III therapeutic clinical trial protocol documents targeting multiple treatment areas were drawn from both the Columbia University Clinical Trials Network (CTN) and Chronic Lymphocytic Leukemia Research Consortium (CLLRC). Each protocol contained a schema represented as a temporal grid. The tasks or event which were included in these temporal grid representations and also satisfied the following two criteria were abstracted by a subject matter expert:

a) The task or event was participant-centric (i.e., described an activity which pertained or was applied to the research participant), and would yield one or more elements of either quantitative or qualitative data.
b) The task or event occurred during the active phase of the protocol, which was defined as the time frame after initial screening and eligibility assessment (i.e., after completing all therapeutic interventions), but prior to long-term follow-up.

Each event or task abstracted from the protocol documents was mapped to a unique UMLS concept using the free-text search engine available via the UMLS Knowledge Source (UMLSKS) server (http://umlsks.nlm.nih.gov). For those tasks or
events that did not result in an exact match using this approach, one of the following strategies was used to assign an adequate UMLS concept:

a) For compound concepts (e.g., “height and weight measurement”), the text was decomposed to the smallest semantically significant units (e.g., “height”, “weight”) and the free-text matching algorithm was then applied to each component. The UMLS concepts found via this process were then subject to post-coordination.

b) Any possible synonyms provided via the free-text matching algorithm were explored, and if a suitable semantic match (as determined by expert opinion) was found, then that concept was selected as the matching UMLS concept.

The resulting collection of unique UMLS concepts, with associated occurrence frequencies at the protocol, treatment group and corpus levels were descending rank-ordered using a composite support metric (Equation 1). A treatment group in this context was defined as a group of protocols targeting a specific disease (e.g., chronic lymphocytic leukemia) or closely related group of diseases (e.g., vascular disease), that was distinct from other diseases or groups of disease in the corpus. Those concepts that composed 95% of the distribution of total concept instances given the preceding rank-ordered by the composite support were selected for subsequent inclusion in a prototype clinical trial participant-related event knowledge collection. The selection of such a threshold was necessary to enhance the generalizability of the results by ensuring that concepts included in our knowledge collection were broadly representative of commonly occurring clinical research participant-related events, rather than characteristic of a single protocol or small number of protocols where the concept may have occurred multiple times. It was derived based upon an iterative, heuristic evaluation of the data set during its analysis as performed by a group of three subject matter experts.

$$S_A = \left( \frac{n}{t_n} \right) + \left( \frac{p}{t_p} \right) + \left( \frac{g}{t_g} \right)$$  \hspace{1cm} \text{Equation 1.}

Composite support ($S_A$) for a given concept ($A$), where $n$ is the total number of occurrences of concept $A$, $t_n$ is the total number of all concept occurrences, $p$ is the number of protocols in which concept $A$ occurs, $t_p$ is the total number of protocols, $g$ is the number of treatment groups in which concept $A$ occurs, and $t_g$ is the total number of treatment groups.

2. Categorical Sorting

Subjects with backgrounds in the conduct of clinical research (e.g., physicians, nurses, study coordinators/managers) were recruited from the Columbia University Medical Center. Each subject performed an “all-in-one” categorical sort [8] of the selected concepts using a Web-based application (www.websort.net). They viewed a list of the concepts, placed them into groups based upon the similarity of their meanings or any other sorter-selected criteria, and provided descriptive names for each group created during the sorting process. The subjects were not given any pre-existing categories to be used or constraints on the number or size of groups to be created during this exercise. The results of the categorical sort were represented using a symmetric agreement matrix where each cell was assigned a numerical score indicating the number of sorters who placed the two concepts indicated by the column and row indices together in a group. Agreement statistics were then calculated to determine how many sorters agreed on each possible pair-wise grouping of a single concept with all remaining concepts.

3. Formalization

Hierarchical cluster analysis was performed, using an average linkage algorithm as implemented in the JMP 5.0.1 statistics package to generate “consensus clusters” of the sorted concepts [9]. Thematic analysis was performed to assess the high-level group names assigned to the concepts that comprised each “consensus cluster”. To enable this analysis and provide a consistent nomenclature for comparison across sorters, each group name assigned by the sorters was manually mapped to a semantically similar UMLS concept. The thematic analysis results were used to organize the subsumed concepts into a basic taxonomy using parent-child relationships.

Results

Our convenience sample of protocols yielded 32 Phase I-III therapeutic protocol documents, which could be generally classified as belonging to one of six major treatment groups: Oncology (54%), Gastrointestinal (19%), Endocrine (9%), Neurology (9%), Vascular Disease (6%), and Hypertension (3%). From these 32 protocols, 522 participant-related event concepts were abstracted and mapped to 93 unique UMLS concepts. All of the abstracted concepts were successfully mapped to UMLS concept and used in subsequent analyses. As described earlier, a composite support metric ($S_A$) was calculated for each unique concept, with values ranging from 0.19 to 2 (SD=0.4). Those unique concepts that comprised 95% of the distribution of concepts when descending rank-ordered by the
composite support metric were selected for subsequent analysis, resulting in a set of 67 concepts (Table 1). Following the selection of the preceding concept set, five subjects were recruited to participate in an “all-in-one” categorical sorting exercise using those concepts. The four female and one male subjects ranged in age from 30 to 57 years (average = 42 years). All of the subjects had significant experience in the area of clinical research (average = 18 years), serving as either research staff (e.g., study coordinator, data manager, research nurse) or clinical investigators. The subjects created 47 unique groups of concepts, ranging in size from one to 33, with an average group size of 6.9 concepts (SD=7). The average pair-wise agreement was 80%. In comparison, a computational simulation of comparable random sorting behavior demonstrated an average pair-wise agreement of 5% [10]. The magnitude of difference between the observed and simulated pair-wise agreement was on average 5 standard deviations. Hierarchical cluster analysis generated 26 “consensus clusters” with an average cluster size of 3.5 concepts.

Table 1. Example of top 10 clinical research participant-related events selected using a composite support metric ($S_4$).

<table>
<thead>
<tr>
<th>UMLS Concept Name</th>
<th>$S_4$</th>
<th>% of Total Concept Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Examination</td>
<td>2.00</td>
<td>6.47</td>
</tr>
<tr>
<td>Medical History</td>
<td>1.92</td>
<td>8.24</td>
</tr>
<tr>
<td>Electrocardiogram</td>
<td>1.50</td>
<td>2.94</td>
</tr>
<tr>
<td>Hematology</td>
<td>1.33</td>
<td>3.92</td>
</tr>
<tr>
<td>Blood Chemical Analysis</td>
<td>1.30</td>
<td>4.51</td>
</tr>
<tr>
<td>Adverse Effects</td>
<td>1.27</td>
<td>2.94</td>
</tr>
<tr>
<td>Inclusion and Exclusion</td>
<td>1.26</td>
<td>2.55</td>
</tr>
<tr>
<td>Dispensing Medication</td>
<td>1.17</td>
<td>2.16</td>
</tr>
<tr>
<td>Obtain or Verify Patient’s Informed Consent</td>
<td>1.10</td>
<td>2.75</td>
</tr>
<tr>
<td>Laboratory Procedures</td>
<td>1.09</td>
<td>5.29</td>
</tr>
</tbody>
</table>

When thematic analyses of the group names associated with the concepts in each “consensus cluster” was performed, it was found that each cluster had an average of seven names associated with it. The seven most frequently occurring thematically unique group names (corresponding to the average number of themes per “consensus cluster”) were selected for use in organizing the knowledge collection.

These theme names and their occurrence frequencies are enumerated in Table 2. Given the preceding “consensus clusters” and associated thematic analysis results, a taxonomy was constructed using parent-child relationships. These relationships were instantiated by assigning the role of “parent” to the seven unique group name concepts selected via thematic analysis, and all of the other subsumed concepts from the initial concept set were assigned the role of “child”. Multiple-hierarchies, cases where a child has more than one parent, were allowed in the taxonomy. In addition, no constraints were applied to the number of categories a concept could belong to. The resulting knowledge collection was represented using an ancestor-descendant table (available at: www.bmi.osu.edu/~payne/).

Table 2. Seven most common thematically unique group names, and number of occurrences associated with "consensus clusters" generated via hierarchical cluster analysis.

<table>
<thead>
<tr>
<th>Theme Name</th>
<th>Occurrences (% of “Consensus Clusters”)</th>
<th>Subsumed Concepts (% of Initial Concept Set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory Procedures</td>
<td>11 (42%)</td>
<td>44 (66%)</td>
</tr>
<tr>
<td>Research Administrative Procedures</td>
<td>10 (38%)</td>
<td>28 (42%)</td>
</tr>
<tr>
<td>Procedures</td>
<td>8 (31%)</td>
<td>35 (52%)</td>
</tr>
<tr>
<td>Screening Procedure</td>
<td>5 (19%)</td>
<td>15 (22%)</td>
</tr>
<tr>
<td>Diagnostic Radiologic Examination</td>
<td>3 (12%)</td>
<td>10 (15%)</td>
</tr>
<tr>
<td>Measurement</td>
<td>3 (12%)</td>
<td>15 (22%)</td>
</tr>
<tr>
<td>Specimen</td>
<td>1 (4%)</td>
<td>10 (15%)</td>
</tr>
</tbody>
</table>

Discussion
As stated at the outset of this paper, the current lack of a common model for describing clinical research participant-related events is a significant barrier to realizing the benefits of computable protocol schemas. The study we have described demonstrates a CKA-based approach to generating a prototype knowledge collection intended to address the preceding gap. Our initial results have shown a systematic means of abstracting clinical research participant-related events from a corpus of protocol documents, and subsequently organizing them into a taxonomy based upon the knowledge of multiple domain experts. Such an approach maximizes multiple lines of reasoning and is widely held in the knowledge engineering literature to generate superior quality knowledge collections [7]. While outside the immediate scope of this manuscript, it important to
note in the context of the preceding results that a subsequent study in which a human-computer interaction model was designed using the Presentation Discovery methodology [11] in conjunction with the knowledge collection generated during this study showed statistically significant improvements in the ability of clinical research staff to perform common clinical trials management tasks via a computer-based tool [12]. Such results indicate that the participant-related tasks and events identified in this study were both recognizable to multiple subject matter experts, and corresponded to common tasks that those individuals performed on a regular basis. We have made our initial results available to the clinical research informatics community in the hopes that other groups may contribute to what could become a valuable resource to support the design of computable clinical trial schemas.

There are several limitations of this work which must be mentioned. The first is our reliance on domain experts for the purposes of organizing our knowledge collection, which assumes the ability to recruit appropriate individuals to participate in this type of study. Second, the semi-automated abstraction and thematic analysis techniques employed to determine participant-related event concepts within protocol documents and interpret categorical sorting results are both resource-intensive and subject to certain amounts of investigator bias, which could limit the reproducibility and scalability of our results. Third, a larger and more broadly representative sample of protocol documents, in contrast to our convenience sample, may be able to produce more generalizable results. Finally, greater validation of the generalizability of our results would be afforded by performing a comparative evaluation using an additional test set of protocol documents, a step the authors intend to pursue as an extension to the initial results reported in this manuscript.

**Conclusion**

Despite the potential limitations described above, we believe that the initial work described in this paper has the potential to contribute significantly to our collective ability to design and implement computable clinical trial protocol schemas. By emphasizing systematic and consensus-based CKA-based approaches in our study, our methods are able to yield a knowledge collection representative of the best available expert knowledge concerning the clinical research domain. Furthermore, by providing the framework to our approach, as well as our initial results to the clinical research informatics community, we believe that there is a potential for significant benefits in terms of enabling the design of computable protocol schemas as our knowledge collection matures, ultimately yielding an improved capability to conduct high-quality clinical studies.

**Acknowledgements**

The authors wish to acknowledge the contributions made to this work by J. Thomas Bigger, James Deitzer, and Stephen Johnson (CU), as well as Andrew Greaves (UCSD). This work was supported in part by NLM Training Grant 5-T15-LM007079-13.

**References**