Development of an Open Metadata Schema for Prospective Clinical Research (openPCR) in China*

W. Xu1,2; Z. Guan3; J. Sun4; Z. Wang5; Y. Geng6

1Department of Medical Informatics, Second Military Medical University, Shanghai, China; 2Department of Library Science, Nanjing Political College, Nanjing, China; 3Department of Basic Medicine, Second Military Medical University, Shanghai, China; 4Department of Health Service, Second Military Medical University, Shanghai, China; 5Department of Medical Informatics, Shanghai Hospital, Shanghai, China; 6Shanghai Banpo Network Technologies Ltd, Shanghai, China

Keywords
Metadata, archetype, electronic health record (EHR), clinical research (CR), Dublin Core Metadata Element Set (DC Metadata Element Set)

Summary
Objectives: In China, deployment of electronic data capture (EDC) and clinical data management system (CDMS) for clinical research (CR) is in its very early stage, and about 90% of clinical studies collected and submitted clinical data manually. This work aims to build an open metadata schema for Prospective Clinical Research (openPCR) in China based on openEHR archetypes, in order to help Chinese researchers easily create specific data entry templates for registration, study design and clinical data collection.

Methods: Singapore Framework for Dublin Core Application Profiles (DCAP) is used to develop openPCR and four steps such as defining the core functional requirements and deducing the core metadata items, developing archetype models, defining metadata terms and creating archetype records, and finally developing implementation syntax are followed.

Results: The core functional requirements are divided into three categories: requirements for research registration, requirements for trial design, and requirements for case report form (CRF). 74 metadata items are identified and their Chinese authority names are created. The minimum metadata set of openPCR includes 3 documents, 6 sections, 26 top level data groups, 32 lower data groups and 74 data elements. The top level container in openPCR is composed of public document, internal document and clinical document archetypes. A hierarchical structure of openPCR is established according to Data Structure of Electronic Health Record Architecture and Data Standard of China (Chinese EHR Standard). Metadata attributes are grouped into six parts: identification, definition, representation, relation, usage guides, and administration.

Discussions and Conclusion: OpenPCR is an open metadata schema based on research registration standards, standards of the Clinical Data Interchange Standards Consortium (CDISC) and Chinese healthcare related standards, and is to be publicly available throughout China. It considers future integration of EHR and CR by adopting data structure and data terms in Chinese EHR Standard. Archetypes in openPCR are modularity models and can be separated, recombined, and reused. The authors recommend that the method to develop openPCR can be referenced by other countries when designing metadata schema of clinical research. In the next steps, openPCR should be used in a number of CR projects to test its applicability and to continuously improve its coverage. Besides, metadata schema for research protocol can be developed to structurize and standardize protocol, and syntactical interoperability of openPCR with other related standards can be considered.

1. Introduction

Prospective clinical research (PCR), especially randomized controlled trial, is of high importance for providing evidence for clinical practice and advancing medicine [1, 2]. With technological innovation, electronic data capture (EDC) has been developed and implemented [3], which has delivered benefits of reducing lag time, increasing data quality and speeding up data entry [4]. However, EDC itself is far from perfect resulting in duplication of data entry and associated costs [4].

As more and more clinical data can be extracted from electronic health record (EHR), the philosophy of integrating EHR and EDC or using EHR for clinical research (CR) attracts growing interests in the field of medical informatics [5, 6]. Some template based systems have already...
been implemented for both routine documentation and CR [7–9].

China, with a huge population, produces rich sources of diseases and cares. Carefully conducted CR from the huge sources can definitely help to find effective treatments for the improvement of health, not only in China but also in other countries of the world. But after searching the China Biomedical Literature Database, called Sino-med, we found that there were few studies on using EDC or clinical data management system (CDMS) for CR. About 90% of clinical studies collected and submitted clinical data manually in China [10], and only one CR related standard was found: Drug Clinical Trial and the Quality Control Standard of China (DCTQCS) [11]. China is lagging far behind in this area.

One Challenge for China to do CR is the lack of registration awareness. Registration of CR will lead to a better overview of which trials are currently running or have stopped. This prevents identical trials from the very beginning and leads to access to basic information about all ongoing clinical trials, thus helping clinicians to use results for the benefit of patients [12]. However, registration of CR in China is still at a low level. According to a survey, in China, funded clinical trials are registered in 23 out of 755 colleges or universities, and 23 out of 6,100 hospitals at or above the county level [13]. Besides, many research results without registration can not be accessed or published worldwide, because some are never published or published in domestic journals and can not be indexed in international databases.

Another challenge is the lack of data standards and related research in this area, and EDC and CDMS have not been efficiently used. Although a web-based clinical data management system called ResMan has already been developed in China, and all the clinical data can be recorded on the web and sent to the central database [14], it charges for fund-supported CR. Through telephone interview with the manager, we are informed that there are only about 40 clinical trials using this ResMan up to now, and the data fields are individualized for each individual trial. Data exchange or interoperability is impossible.

Facing this awkward situation, our objective is to build an open metadata schema for Prospective Clinical Research (openPCR) in China based on openEHR archetypes so that researchers can easily create specific data entry templates for research registration, study design, and clinical data collection.

Metadata is the data about data, and it has different connotations in various domains, fields, or even different studies. In healthcare section, metadata describes and defines the hierarchical structure of health information [15, 16]. Metadata schema defined in ISO 23081 [17], “is logical plan

![Figure 1](methods-online.com)
showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality (obligation level) of values”. In the field of medical informatics, openEHR Archetypes can be considered as a kind of metadata model.

OpenEHR provides an EHR-oriented architecture using two-modeling approach with the reference information model as its first level and archetype and template based form as the second [18]. Archetypes represent knowledge-level concept definitions, and “denote a model defining some domain concept, expressed using constraints on instance structures of an underlying reference model” [19]. According to some experts, openEHR archetypes are reusable and structured models, therefore they can well support CR [1, 20].

2. Methods

We used Singapore Framework for Dublin Core Application Profiles (DCAP), a generic construct for designing metadata records [21], to develop openPCR. The framework holds the design and documentation components of specific metadata applications which can be grouped into four processes for metadata schema development: defining functional requirements, developing a domain model, defining metadata terms and designing a metadata record, and finally developing implementation syntax. We followed the four steps in modeling openPCR: 1. Define the core functional requirements and deduce the core metadata items; 2. Develop archetype models; 3. Define metadata terms and develop archetype records; 4. Define implementation syntax. See supplementary web material for the methodological details (Web appendix A).

3. Results

3.1 Results of Step 1

The top level objective of openPCR was to develop a metadata schema for research registration, trial design and CRF. Using business process method, we created the core functional requirements under three categories: requirements for research registration, requirements for trial design, and requirements for CRF. And the more specific business process of each category formed a core requirements framework (Figure 1).

There are three stages of registration. The minimum protocol items and the consent forms should be registered prior to enrollment of trial participants. Protocol amendments should be dated and registered after participants enrollment begins. Trial results should be registered once the analyses are completed and verified [22]. Development of openPCR is in the process of preparing a protocol before participants enrollment. At this registration stage, we determined four essential metadata categories: agent, business process, record content, and mandates based on ISO 23081-2: Information and documentation-Managing metadata for records-Part 2: Conceptual and implementation issues [23]. The ISO 23081 series describes metadata for records. Some of the purposes of registration such as easily accessing to valid...
information about trials and facilitating the ability to understand trials conform to benefits of metadata for records in ISO 23081. Thus 5 core requirements of Registration and 22 deduced metadata items were created (Supplementary Table 2).

The process in designing a research includes allocating study identification, and determining study methods, enrollment rules, visits, and outcomes. Core requirements of Trial design can be grouped into study identification, study methods, enrollment, visits, and outcomes. 12 core requirements and 22 deduced metadata items were created (Supplementary Table 3).

CRF is a subject related data collection tool for visits. In CR, each subject is assigned with a subject identifier, and entered into visit process. Visit process can be roughly grouped into four periods: screen, baseline, treatment, and follow up. Some requirements are specific to certain visit periods and some are generic to all the visit periods. 14 core requirements and 37 deduced metadata items were created (Supplementary Table 4).

3.2 Results of Step 2

Eighty-one core metadata items deduced from core functional requirements were analyzed and metadata items of the same semantic meaning and content were combined. Contents of metadata items of Intervention(s), Study type, Primary outcome(s), and Key secondary outcomes in core requirements of Registration corresponded to those of items of Primary endpoints, Secondary endpoints, Study type, Method of allocation, Blinding, and Intervention, Dosage in core requirements of Trial design. We combined them and determined that metadata items of Intervention, Study type, Primary outcome(s), and Key secondary outcomes can be reused by both public document archetype for registration and internal document archetype for trial design. Thus 74 metadata items were identified and their Chinese authority names were created.

By adopting concepts of document, section, data group, and data element in Chinese EHR Standard, we constructed a hierarchical structure of openPCR. The top level container in openPCR was composed of public document archetype, internal document archetype and clinical document archetype.

There were 22 data elements in public document archetype. We used DC Metadata Element Set to structure and represent the metadata elements in public document. DC Metadata Element Set could be grouped into three sections: Content, Intellectual property, and Identification. In addition to these three sections, we extended another section containing elements related mainly to administration of CR (Figure 2).

Internal document can not only help to model the structured protocol, but also provide value options for clinical metadata. Internal document archetype was categorized into six data groups: Study identification, Study description, Trial arms, Trial elements, Trial visits, and Trial inclusion/exclusion. In internal document, 19 data elements were structured hierarchically by the six data groups (Figure 3).

Unlike study-level metadata containers as public document and internal document, clinical document is subject-level metadata and was broken up into a header section and a body section. The header section was composed of patient identifier, arm code, visit number, element code, and demographics related information. The body section included clinical data for CRF corresponding to each study visit (Figure 4).

The body section included data groups of Inclusion and exclusion criteria, Adverse event, Medical history, Substance use, Exposure, Physical examination, Prior and
concomitant medications, Vital signs, Laboratory test results, ECG test results, andDisposition. Some of the data groups in body section could be further categorized into smaller data groups, and we divided these data groups according to Chinese EHR Standard. Core data elements under these data groups could be reused by the smaller data groups. We created a generic data group containing the core data elements as a slot. Data group archetype of Physical examination can be expressed by Figure 5.

### 3.3 Results of Step 3

Three documents, six sections, 26 top level data groups, 32 lower data groups, three generic data groups and 74 data elements for minimum metadata set of openPCR were created. All the documents, sections, data groups, and data elements in openPCR were described by basic metadata attributes, and the complete archetype record was developed. We took data element of Date of Collection in the data group of laboratory test results as an example (Supplementary Table 5).

We used the multi-axial identifier in metadata attribute of metadata identifier analogous to identifier specified in the openEHR Support [24]. Multi-axial identifier is a kind of hierarchical identifier. Each identifier instance denotes a single metadata within a versioned 3-dimensional space, with the dimensions being: model name plus document name, data group name, data element name, and version. Each part has been divided by “.”. Multi-axial identifier can express the hierarchical level of metadata, and support mnemonics well.

### 3.4 Results of Step 4

We represented openPCR archetypes by UML and developed openPCR XML schema. We took the data group: skin as an example. In the body section of clinical document, the top level data group: Physical examination has lower data group: Skin, which can reuse data elements in data group: generic: physical examination. The XML instance of data group: Skin can be shown in Supplementary Figure 2.

### 4. Discussions and Conclusion

#### 4.1 Related Work

CR data are important for health care improvement, and are accumulating increasingly. There are several ongoing efforts to represent and structure CR.
Ontology of Clinical Research (OCRe) is a modular ontology of clinical investigation which can represent structure of human study and associated entities, provide informational entities and terms, and bind to external standards. Based on conceptualization of life cycle of human studies, OCRe is composed of three core modules (clinical, research, and study design) and one extension module (study protocol) applicable to any clinical domain. OCRe focuses on design and analysis of human studies aiming to query and annotate various human studies [25]. Both OCRe and openPCR aim to standardize metadata of CR in a way, and are being developed by using the general business process cycle of conducting CR.

There are also significant differences. OpenPCR is a kind of conventional information models focusing more on research registration, trial design and clinical data collection, while OCRe is an ontology primarily for study design and analysis, and does not provide detailed data elements at the execution phase of CR. OpenPCR is designed for CR data acquisition. OCRe, in contrast, is deployed mainly for query and annotation of CR. Another difference between openPCR and OCRe is the languages used. OCRe is formalized in OWL, which supports automated inferences by logical axioms and formal semantics. In comparison, openPCR adopts XML to represent its syntax.

The Linked Clinical Trials (linkedCT) published clinical data according to the rules of publishing linked data. The core dataset of LinkedCT is derived from the clinical trials registry, ClinicalTrials.gov. LinkedCT provides HTTP Uniform Resource Identifier (URI) to each entity, transforms the data into RDF, delivers a SPARQL query service, and supports types of links to other relevant information sources [26]. OpenPCR also provides metadata schema about research registration.

LinkedCT aims to enhance the discovery of clinical trials on the web, and the main data is from ClinicalTrials.gov. On the other hand, openPCR aims to streamline CR practice by designing metadata schema for research registration, study design, and clinical data collection. The difference also lies in language using. LinkedCT transforms XML files of ClinicalTrials.gov into RDF format, while openPCR uses XML syntax.

Open Study Data Management System (openSDMS) is developed based on openEHR to facilitate reuse of EHR data. OpenSDMS is a prototypical system facilitating entering medical data for clinical trials and recurring health care data in the EHR module. OpenSDMS defines archetypes of clinical trial, trial arm, and trial visit. Data elements representing concepts of clinical trials are identified from German Clinical Trials Register and Coordination Center for Clinical Trials Heidelberg [1]. Both openSDMS and openPCR developed metadata of clinical trial based on openEHR. OpenSDMS defined archetypes for specific clinical trial metadata in addition to existing archetypes for medical data. Although most archetypes in clinical document had corresponding ones in the openEHR Clinical Knowledge Manager such as adverse event, physical examination, vital signs, and etc. [27], openPCR created new archetypes for all the concepts, because many clinical data for research are different from those in EHR. For example, adverse event archetype in openPCR has metadata elements quite different from data in adverse reaction archetype in openEHR. Metadata elements of Serious adverse event, Relationship of adverse event to study, and Action taken to adverse event with study in CR have no counterparts in openEHR and vice versa. The next step will include data mappings of the same concepts between openPCR and openEHR.

Archetype approach in openEHR is a hierarchical, combinational, and building block method. It is especially useful to express complex data content and data structure, such as medical information. Simple archetypes at any level can be created easily, and complex archetypes can be created by combining the simple archetypes. OpenEHR is a publicly available specification based on the formal European standard for EHR communication, and is also suitable for representing data from clinical trial.

OpenPCR is a prototypical schema also based on openEHR and just core functional requirements and minimum metadata elements have been developed. Users can define custom metadata items with self-defined or already existed archetypes. It is an open metadata schema based on research registration standards, standards of the Clinical Data Interchange Standards Consortium (CDISC) and Chinese healthcare related standards, and plans to be publicly available throughout China. We recommend that the method to develop openPCR can also be used to design metadata schema of CR at the national level in other countries with well-developed EHR standard. It includes not only data collection metadata, but registration metadata and trial design metadata aiming to streamline CR practice.

Although openPCR is a metadata schema for CR, it considers future integration of EHR and CR from three aspects:

1. Information model of openPCR adopts Data Structure in Chinese EHR standard. Data Structure in Chinese EHR standard serves the same purposes as EHR information model in openEHR. EHR information model defines a stable logical EHR information architecture and consists of a relatively small number of non-volatile classes. It significantly reduces the dependency of deployed systems and data on variable content definitions. By adopting Data Structure of Chinese EHR standard, openPCR can be mapped to Chinese EHR standard based EHR system model at a high level.

2. Terms of metadata elements are adopted from ChiCTR registry, Chinese EHR standard and NHDD. Defining terms of metadata elements from Chinese standards facilitates consistency and stable representation. Directly using terms in Chinese EHR standard to represent metadata elements with same meanings will help future data exchange and sharing between data of EHR and CR.

3. Data elements mappings are created between openPCR and Chinese EHR standard. In relationship category of metadata attributes, attribute of mappings is created to express the corresponding metadata elements in other Chinese and international standards. This also provides candidates for mapping.
We used DC Metadata Element Set to represent metadata elements of public document. The DC Metadata Element Set is a simple and effective element set for describing online resources and has been widely supported [28]. It can also represent and retrieve health information well [29, 30]. Besides registering CR, researchers can also publish CR on their websites. And using DC Metadata Element Set facilitates information retrieval by Internet search engines [29].

4.2 Evaluation

OpenPCR developed here is still a draft version, and openPCR platform is still in progress. To enable openPCR usability, we will bring it into a critical, public review process. The strengths and weaknesses of openPCR will be evaluated in a real but manageable clinical trial setting. The following evaluation aspects are planned for evaluation:

1. Metadata quality: the National Information Standards Organization (NISO) provides six metadata principles for building good digital collections [31]. We believe they are also applicable to control the metadata quality in openPCR. These principles include that good metadata should conform to community standards, support interoperability, use authority control and content standards, include a clear statement of the conditions and terms of use, support the long-term curation and preservation, and have the qualities of good objects, including authority, authenticity, archivality, persistency, and unique identification.

2. Metadata creation: Since openPCR is a metadata schema for minimum dataset, users should create many metadata items themselves for a specific department or research. This evaluation factor is mainly to test the time efficiency of metadata creation and metadata’s conformance to openPCR usage guidelines.

3. Metadata use: Ease of use is the ultimate goal of openPCR. We will assess whether openPCR will streamline the CR practice and whether it will be easy for researchers to register their clinical research, submit the data, and make statistics.

4.3 Limitations and Future Work

We developed openPCR based on Chinese EHR Standard. The function of openPCR for EHR and CR will largely depend on the quality and implementation of the EHR standard. Since Chinese EHR Standard is still in its draft version and has not been widely used by EHR systems in China, the implementation of function for EHR and CR will not be realized for a long time.

OpenPCR is still a prototypical schema, and only minimum data set has been built. Although we developed an openPCR platform, it is still in its progress. In the next step, a number of CR projects should be carried out by openPCR to test its applicability and to improve its coverage continuously. We will also publish openPCR platform on the web for an open review and open use for Chinese researchers.

At this time, we did not consider the metadata schema for CR protocol. Protocol is the soul of CR. It not only helps to manage and audit conduct throughout CR process, but also provides information for trial registration and CRF creation. In a next step, metadata schema for CR protocol will be developed to structurize and standardize protocol and streamline the process of protocol development, trial registration, and CRF creation.

We developed XML format of openPCR, but we did not take into account syntactical interoperability of openPCR with other healthcare or CR standards such as ODM and SDM-XMI in the field of CR, and HL7 CDA in the field of healthcare. In the future work, we will consider the XML mappings between openPCR and other standards to enhance the ability of data exchanging and sharing.

In conclusion, we successfully developed an open metadata schema for Prospective Clinical Research (openPCR) in China. The method may well be also used to design metadata schema of CR at the national level in other countries with well-developed EHR standard. It makes it easier for researchers to register research information, design research, collect clinical data, and takes into account future integration of EHR and CR system in China.

Acknowledgments

This study was sponsored by National Nature Science Foundation of China (No. 30972549) and Shanghai Pujiang Program. We sincerely thank Prof. Zhao Liang, Zhang Zhengqiang and Dr. Wu Jianzhong for valuable advice in the research.

Wei Xu and Jinhai Sun (correspondence to: Jinhai Sun, Department of Medical Service, Second Military Medical University, 800, Xiangyang Road, Yangpu District, Shanghai, China; E-mail: sunjinhai2003@sina.cn) contributed equally to this work.

References

27. http://www.openehr.org/cckm/