Cross-Trial Query System for Cancer Clinical Trials

Radu Calinescu, Steve Harris, Jeremy Gibbons and Jim Davies
Computing Laboratory, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, UK

Abstract **Data sharing represents one of the key objectives and major challenges of today's cancer research. CancerGrid, a consortium of clinicians, cancer researchers, computational biologists and software engineers from leading UK institutions, is developing open-standards cancer informatics addressing this challenge. The CancerGrid solution involves the representation of a widely accepted clinical trials model in controlled vocabulary and common data elements (CDEs) as the enabling factor for cancer data sharing. This paper describes a cancer data query system that supports data sharing across CancerGrid-compliant clinical trial boundaries. The formal specification of the query system allows the model-driven development of a flexible, web-based interface that cancer researchers with limited IT experience can use to identify and query common data across multiple clinical trials.**

I. Introduction

CancerGrid [1] is a project that brings together clinicians, cancer researchers, bioinformaticians and software engineers from leading UK institutions. The project uses a generic clinical trials model [2] based on controlled vocabulary and common data elements (CDEs) [3] for the design, execution and analysis of cancer clinical trials. The advantage of this approach is twofold. Firstly, it makes possible the model-based development of open-standards IT systems for clinical trial execution [4]. Secondly, the CancerGrid approach enables data sharing across cancer clinical trial boundaries. The latter capability is demonstrated by the cross-trial query system presented in this paper.

The remainder of the paper is organised as follows. After a review of related work in the next section, Section III uses Z notation [5] to formally define common data elements. Section IV shows how CDE sets are associated to trial events to create a *trial design*, i.e., a full specification of a cancer clinical trial. Section V describes the cross-trial query model, and its use of the CDEs associated with the same execution stage (e.g., patient registration or follow-up) of multiple trials.

A proof-of-concept system that implements the cross-trial query model is presented in Section VI. The system allows cancer researchers with limited IT experience (e.g., clinicians and statisticians) to easily identify common data across multiple clinical trials, and to build queries targeted at these data using a familiar interface.

Section VII discusses two possible extensions of the query model. First, the grouping of CDEs associated with different trial execution stages is considered as a way of making queries less restrictive. Second, a solution to the generation of queries compliant with the security constraints specific to clinical trials is investigated. Section VIII concludes the paper with an overview of the query system, and an analysis of future work directions.

II. Related Work

The query of data from multiple sources has been an important research topic for the last two decades. Generic approaches for querying multiple information sources were proposed [6, 7, 8] that use a model of a problem domain to devise global query systems. The approach in [6] requires the user to build a semantic *domain model* as well as a model of each database and knowledge base used as an information source. Therefore, this solution is appropriate only for users with expertise in both data modelling and the target problem domain. Similar approaches are described in [7, 8], where sophisticated techniques are used to create a "metadatabase" [7] or a "reference data model" [8] that are then employed to generate the global query. Unlike these approaches that address the query of heterogeneous data sources, our query system takes advantage of the homogeneity of data across cancer clinical trials to hide most of the complexity of a cross-trial query. Implementations of this system can therefore be used directly by cancer researchers with limited data modelling expertise.

In the cancer research area, the US cancer Biomedical Informatics Grid (caBIG) project [9] models clinical trials [10] and has cancer data sharing as one of its primary objectives. Their caCORE software development kit [11] provides building blocks for many software components employed in cancer research. The inclusion of multiple data source querying in a proprietary language (i.e., the caBIG Query Language) is planned for the next release of the kit. While this will provide the functionality required to implement a system for querying multiple cancer data sources, the query system presented in this paper allows the automatic generation of a complete query form ready for immediate use by clinicians and statisticians.

Other medical projects such as VOTES [12] and PRATA [13] are concerned with the integration of data from multiple, distributed databases. The VOTES system [12] is concerned with the integration of distributed medical data pertaining to the same patient, so candidate patients for new clinical trials can be identified easily. The query forms used by the VOTES portal resemble those from the prototype implementation of the query system introduced in this paper, however they are encoded manually by software developers familiar with the internal structure of the data sources. The PRATA system [13] addresses the XML integration of data extracted from multiple, distributed databases. The integration and visualisation of the data is based on a user-specified XML schema that requires inside knowledge of the data sources. On the contrary, the CancerGrid query forms are model-based, and provide information to guide user querying rather than relying on the users for this knowledge.

III. Common Data Elements

The consistent use of a controlled vocabulary (i.e., a list of explicitly enumerated terms managed by a vocabulary

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 385–390.

registration authority) is key to sharing data between projects in any field of research. This is particularly relevant to cancer research, where tremendous human and financial resources are often employed for the generation of relatively small amounts of data [1]. The ability to analyse these data across multiple clinical trials is crucial to reaching statistically relevant conclusions.

The CancerGrid project is addressing this important requirement by basing its clinical trials model [2] on the use of thesauri—collections of controlled vocabulary terms and their relationships, and common data elements—controlled sets of cancer concepts and measurements. A common data element [3] is defined in terms of several basic types:

- *CdeID*, the set of all common data element identifiers. CDE identifiers are used to uniquely refer to specific CDEs.
- *CdeType*, the set of all types that CDE values may have. Typically, any XML schema simple type is allowed.
- *CdeInfo*, the actual details of the CDE, including a name and a description.

These basic types are summarised below using Z notation [5]:

$$[CdeID, CdeType, CdeInfo], \tag{1}$$

and the common data element type can be specified as:

$$\begin{array}{|l} Cde \\ \hline id : CdeID \\ valueDomain : CdeType \\ info : CdeInfo \\ \hline \end{array} \tag{2}$$

Common data elements used to model data in a specific research field are maintained in a CDE (or metadata) repository for that area of research:

$$\begin{array}{|l} CdeRepository \\ \hline cdeSet : \mathbb{P}\ Cde \\ \hline \forall x, y : cdeSet \bullet x.id = y.id \Rightarrow x = y \\ \hline \end{array} \tag{3}$$

IV. Clinical Trial Events and Trial Designs

Clinical trial data are generated during the execution of a trial as a result of a number of trial events, each of which corresponds to a stage in the execution of the clinical trial. For instance, clinical and personal patient data are collected during the *registration* stage, treatments are allocated in the *randomisation* stage, and periodical *follow-up* data collection is performed to assess response to treatment. The complete set of trial events in the CancerGrid trial model is given below:

$$\begin{aligned} TrialEvent ::= {} & registration \mid eligibility \mid randomisation \mid \\ & onStudy \mid treatment \mid offStudy \mid response \mid \\ & followUp \mid adverseEffect \end{aligned} \tag{4}$$

Clinicians gather the data corresponding to the trial events by filling in case report forms that comprise CDEs drawn from the cancer CDE repository [3],

$$\mid cancerCdeRep : CdeRepository. \tag{5}$$

For the purpose of data analysis, a clinical trial is composed of a set of trial events, each of which is associated with a set of common data elements [2]. This is defined by the *TrialDesign* specification below:

$$\begin{array}{|l} TrialDesign \\ \hline events : \mathbb{P}\ TrialEvent \\ eventCdeSet : TrialEvent \rightarrow \mathbb{P}\ cancerCdeRep.cdeSet \\ \hline \mathrm{dom}\ eventCdeSet = events \\ \hline \end{array} \tag{6}$$

To give an example of a clinical trial design, consider the following common data elements:

$$\begin{array}{|l} NodalStatus, AdjuvantRadiotherapy, Her2Level, \\ ECOGStatus, InvasiveCarcinoma, TumorResectionStatus, \\ DiseaseStage, AdjuvantChemotherapyIndication, \\ PatientFitness, BoneMarrowHepaticRenalFunction, \\ InformedConsent, NoPreviousTherapy, KnownRadiotherapy, \\ LastSurgeryDate, NoPreviousMalignancy, \\ NotPregnantLactating, PatientNameInitials, \\ PatientBirthDate, TissueSubstudyConsent, \\ QualityOfLifeSubstudyConsent, ProgesteroneReceptorStatus, \\ OestrogenReceptorStatus : cancerCdeRep.cdeSet \end{array} \tag{7}$$

that are associated with three of the trial events for the tAnGo clinical trial [14]:

$$\begin{array}{|l} tAnGo : TrialDesign \\ \hline \{registration, eligibility, randomisation\} \subset tAnGo.events \\ tAnGo.eventCdeSet\ registration = \\ \quad \{TissueSubstudyConsent, QualityOfLifeSubstudyConsent, \\ \quad ProgesteroneReceptorStatus, OestrogenReceptorStatus\}. \\ tAnGo.eventCdeSet\ eligibility = \\ \quad \{InvasiveCarcinoma, TumorResectionStatus, \\ \quad DiseaseStage, AdjuvantChemotherapyIndication, \\ \quad PatientFitness, BoneMarrowHepaticRenalFunction, \\ \quad InformedConsent, NoPreviousTherapy, \\ \quad KnownRadiotherapy, LastSurgeryDate, \\ \quad NoPreviousMalignancy, NotPregnantLactating\} \\ tAnGo.eventCdeSet\ randomisation = \\ \quad \{NodalStatus, AdjuvantRadiotherapy, Her2Level, \\ \quad ECOGStatus\} \end{array} \tag{8}$$

The common data elements used to register *tAnGo* participants, to establish their eligibility, and to stratify the allocation of treatments for the eligible participants (i.e., the trial *randomisation* [15]) are explicitly specified in the *tAnGo* trial design. Note that the complete trial design for tAnGo comprises all of the trial events defined in (4), however for the sake of brevity only three of these are presented above.

V. Clinical Trial Queries

Having introduced the data components of a CancerGrid clinical trial in the previous section, we will now define the cross-trial queries for sharing data among clinical trials using the same CDE repository. A number of comparison operators are used to build the query: