TRANSFoRm

Translational Research and Patient Safety in Europe

D5.3: Query Formulation Workbench

UNIVERSITY OF BIRMINGHAM

University of Birmingham, United Kingdom

Work Package:	WP5, WT 5.3a
Type of document:	Software Deliverable Technical Description
Version:	V1.0
Date:	18 July 2012
Authors:	L. Zhao, S.N. Lim Choi Keung, J. Rossiter, T.N. Arvanitis

TRANSFoRm is partially funded by the European Commission - DG INFSO
Under the 7th Framework Programme. (FP7 247787)
7th Framework Programme http://cordis.europa.eu/fp7/ict/
European Commission http://ec.europa.eu/information_society/index_en.htm

eHealth
Better Healthcare for Europe

European Commission

Table of Contents

Executive Summary

This deliverable describes the outcomes of WT 5.2 Query and data extraction workbench, which has developed “the interfaces necessary to author, store and deploy queries of clinical data to identify subjects for clinical studies”. The work presents the design and implementation of the TRANSFoRm Query Formulation Workbench Software Tool.

The deliverable provides the list of functional and user requirements for the tool, as derived from the combined analysis of the use cases provided by deliverable D1.1 and the information elicited by conducting participatory task modelling, involving a group of expert users. This analysis is enhanced with a comparative literature review of features and functionalities of relevant state-of-the-art research study eligibility query based tools. Through this review, some go the limitations of existing solutions are highlighted, while discussed in the context of how TRANSFoRm addresses these on its query formulation workbench solution.

This is a semantically aware software tool that supports the easy authoring of distributed searches to EHR and other clinical data sources. The query authoring based on an eligibility criteria representation of the Clinical Research Information Model (CRIM), in order to achieve the identification of research subjects from EHR data sources. The use of the TRANSFoRm terminology services, in conjunction with the Clinical Data Integration Model (CDIM) allows the capturing eligibility criteria in a computable representation, based on CDIM ontology, so the criteria can be translated into executable query statements at the individual EHR data sources. In this way the workbench achieved to automatically identify ‘prevalent cases’ for research, where the searches report back counts of eligible subjects in the EHRs, flagging the subjects for recruitment and consent by the local clinical care team, in full compliance with data protection legislation and best practice.

The deliverable demonstrates the overall functionality of the query formulation workbench requirements and associated UI on an example study: the eligibility needs of the diabetes use case. The UI provides mechanisms and simple visual representations of groupings of eligibility criteria, which can be easily managed by users as a series of simple interaction steps, guiding the user throughout the process.

Integration with other TRANSFoRm components, such as the security framework, has provided a successful working functionality of the query formulation tool within the TRANSFoRm distributed infrastructure. Users can work together on study and protocol design, with rules for collaborative access, driven by the predefined user requirements. Study protocols are designed using an intuitive interface where authorisation rules are used to restrict access to a user’s individual or group permissions. Integration with the TRANSFoRm provenance service also allows for further auditing of user actions.

1. Introduction

The increased adoption of electronic health records (EHR) provides the possibility for clinical researchers to search for eligible patients on individual EHR repositories. The heterogeneity of EHR systems, however, has presented a major bottleneck to “digitally” search on multiple EHR data sources, particularly for large-scale multi-centre clinical studies. Not only are these EHR systems often implemented in different data structures with diverse access interfaces, the data itself are also encoded in different coding schemes. The TRANSFoRm project is coming to help resolve these issues. According to Annex I - "Description of Work" of the TRANSFoRm project, the research aims “to develop a common digital infrastructure to support the learning health care system, including methods, models, services, validated architectures and clinical demonstrations of software to support:

1. Epidemiological research using GP records, genomic and other databases (e.g. privacy, data linkage, data quality tools)

2. Research workflow embedded within the EHR (e.g. subject identification, alerts, functional electronic case report form)

3. Decision support for diagnosis in primary care (e.g. integration of clinical prediction rules to provide prompting/alerting within the EHR via ontology and web-services).” [1]

As part of these objectives, TRANSFoRm is developing a semantically aware query formulation workbench, in order to enable the easy authoring of distributed searches to EHR and other clinical data sources, using a controlled vocabulary service and appropriate standards-based technological solutions. The main aim of such a tool is to automatically identify ‘prevalent cases’ for research, where the searches will report back counts of eligible subjects in the EHRs, flagging the subjects for recruitment and consent by the local clinical care team, in full compliance with data protection legislation and best practice.
This deliverable describes the outcomes of WT 5.2 Query and data extraction workbench, which has developed “the interfaces necessary to author, store and deploy queries of clinical data to identify subjects for clinical studies” [1]. As part of this deliverable we discuss the way TRANSFoRm provides an interface to realise query authoring, based on an eligibility criteria representation of the Clinical Research information model (WT6.4), in order to achieve the identification of research subjects from EHR data sources. The use of the TRANSFoRm terminology services, in conjunction with the Clinical Data Integration Model CDIM (WT6.5) allows the capturing eligibility criteria in a computable representation, based on CDIM ontology, so the criteria can be translated into executable query statements at the individual EHR data sources. The deliverable also discusses how the query formulation workbench tool interacts with TRANSFoRm’s distributed infrastructure (WT7.5), provenance tool (WT5.2) and the overall security framework, as provided by WT3.3.

The deliverable is organised as follows. Section 2 provides an overview of the elicitation process (based on participatory design, use case analysis and comparative literature review) of the functional and user requirements of the query workbench and concludes with the list of these. Section 3 discusses the system architecture and its functional components, which

form the context for the design and implementation of the TRANSFoRm query formulation workbench, as defined by elicited requirements discussed in section 2. Section 4 discusses the Query Formulation Workbench User Interface design and workflow, and provides storyboarding based on an example use case. Section 5 discusses the technical implementation details of the workbench, with reference to the other components of the TRANSFoRm digital infrastructure. The document includes some concluding remarks, references and associated appendices with user interface mock-ups and implementation screenshots.

2. Overview of the Query Formulation Workbench Requirements

The design and implementation of the query formulation workbench was based on a requirements-driven approach. To gather the appropriate functional and user requirements for the tool, we followed the combined analysis of the use cases provided by deliverable D1.1 [2] and the information elicited by conducting participatory task modelling, involving a group of expert users. Participatory task modelling combines the strengths of task analysis and participatory design [3]. Task analysis comprises a wide range of research and development activities, including data collection, data analysis, and modelling of the task domain. Task analysis has mainly involved users at the data gathering stage, while the analysts are mainly involved in the data analysis and the modelling of users' tasks. Participatory design, in comparison, encourages the involvement of users with developers in the systems development activities, as highlighted [4]. As part of the analysis of the D1.1 use cases and participatory design sessions with the authors of the reports, including other domain experts available in the consortium, we produced a definitive "wishlist" of query requirements, while we enhanced this knowledge through task analysis work, based on users and existing literature. As part of latter, we conducted a literature review of existing eligibility criteria query interfaces (Section 2.2), describing the desirable features of a query tool. As part of this review, we are also highlighting, by means of comparison, some of the limitations of existing solutions that TRANSFoRm seeks to address and improve on its query formulation workbench solution. At the end of this section, we provide the list of the TRANSFoRm Query Formulation Workbench requirements list, as derived from the elicitation approach, described above.

2.1. TRANSFoRm Query Formulation Requirements Desiderata

Concerning the requirements of query formulation, data extraction and linkage, the use cases of D1.1 provide a generic "wishlist" of desiderata, as follows:

1. **The system should present options** to the researcher. This could include: **preview of available research information** (counts/specific patient characteristics).
2. **Authorize** (including present conditions and need for informed consent) data extraction (and linkage)
3. **Select** patients (study population) in different databases and/or eHR
4. **Extract** information for the selected population/patients
    a. Information to be able to determine cases and controls or to determine risk factors for a cohort design
    b. Covariates
5. **Reintegrate** this data in a new temporary TRANSFoRm-database
    a. link at the level of individual persons
    b. put information from different databases, eHR and countries in one database

6. **Present** the extracted **data** so that it is ready for further analyses

To put into perspective this generic "wishlist", we look into the details of an example use case for TRANSFoRm, as described by D1.1. The specific use case model is a genotype-phenotype diabetes study, with the main focus on the research question (RQ2): "Are well selected single nucleotide polymorphisms (SNPs) in type 2 diabetic patients associated with variations in drug response to oral antidiabetics (Sulfonylureas)? [2]. A case control design will be used to investigate the link between specific SNPs and responses to antidiabetic drugs. The selection and data extraction will be done in the following steps:

a. Select Type 2 Diabetic (T2D) patients from the genomic database.

b. Select T2D patients using specific target medications from primary care database.

c. Extract information on variables identified, as described in [2][5].

d. Link data at individual patient level.

e. Define cases and controls based on HbA1c information.

The query formulation is seen to be used in the first instance for Step b above, to identify patient counts matching the criteria of T2D with targeted medications. Figure 1 shows the eligibility criteria for T2D patients in Step b, above, as shown in Deliverable D1.1 [2].

From the above example, one can identify the relevant eligibility criteria desiderata for a feasibility assessment query tool, as highlighted below:

- Intent, which categorises criteria into inclusion ("operational selection") and exclusion criteria.
- Main clinical category, which organises criteria as demographics, clinical findings, medical history, allergies, procedural or surgical history, behavioural characteristics, laboratory data, device data, vitals, prior or concomitant medications, and administrative and informed consent issues.
- Main medical topic separates eligibility criteria into disease areas (e.g. cancer) and finer clinical details specific to the topic (e.g. tumour size and stage for breast cancer topic).
- Concepts and their mapping to coding sources needs to be readily available.

A. *Concept name*
Type 2 diabetes

B. Synomyms
Diabetes type 2

C. Required operational selection criteria[2]
Coded diagnosis (including degree of certainty – definite/probable/possible)
Specific codes for type 2 diabetes:

Definite type 2 diabetes: specific T2D code and no contradictory codes

Probable type 2 diabetes: less specific diagnostic codes (e.g. maturity-onset diabetes, non-insulin dependent diabetes mellitus, NIDDM) OR presence of contradictory codes (T1D AND T2D)

Possible type 2 diabetes: codes specific for other types of diabetes as 'steroid induced diabetes', 'gestational diabetes' OR vague high level codes as 'diabetes mellitus' OR multiple contradictions in the coding.

Therapeutic data (drug therapy)
Insulin alone with a prescribed daily administration frequency < 3 OR
Insulin + oral glucose lowering agents +/- metformin OR
Oral glucose lowering agents +/- metformin OR
Metformin OR
No drug therapy

Laboratory data
Fasting blood glucose ≥7.0 mmol/l OR
(undetermined) blood glucose ≥11.1 mmol/l OR
Glycated haemoglobin (HbA_{1c}) ≥6.5%

Other data
Age ≥35 AND

Exclusion criteria
blank

Figure 1: Diabetes eligibility criteria

It is interesting to note, that Weng *et al.* [6] have conducted a systematic review of models and systems with computer-based eligibility criteria knowledge representation and have derived a similar list of desiderata for query formulation requirements, which confirms that the TRANSFoRm use cases have provided a list of query requirements that is acceptable to the wider domain of clinical research.

Furthermore, as part of the participatory design work with domain experts and the various interactions with other work tasks within the consortium, we identified the following drivers for the functional requirements of the query formulation workbench (described in detail at section 3):

1. The development of the system adopts a model-based approach, where the TRANSFoRm Clinical Research Information Model (CRIM) provides a computable information model for eligibility criteria.

2. Criteria concepts, especially clinical concepts, can be browsed and selected through the TRANSFoRm Integrated Vocabulary Service. The vocabulary service provides mappings from standard UMLS concepts to standard EHR or clinical data sources' coding schemes.

3. The eligibility criteria are captured in a computable representation, based on the Clinical Data Integration Model (CDIM) ontology; CDIM captures an extensible common representation of clinical care data, which can be submitted to the TRANSFoRm distributed infrastructure for data extraction and linkage.

To enhance our understanding on the query formulation workbench desiderata, and identify the essential list of user requirements for the TRANSFoRm query formulation workbench, we have enhanced our use case analysis and participatory work, with a comparative literature review of four existing eligibility criteria query tools. The section that follows discusses their features and services, while compares them in respect to what the TRANSFoRm tool is providing.

2.2. State-of-the-Art: Other Query Formulation Tools

Four different query tools have been studied with respect to their requirements and features. ePCRN, i2b2/SHRINE, FARSITE and VISAGE are described in the following sections. This is followed by a comparison of the tools with respect to the desiderata and proposed features of the TRANSFoRm query tool, with a summary comparative table of individual features.

2.2.1. ePCRN

The ePCRN project [7][8] built a sophisticated infrastructure to support the design and implementation of randomised clinical in the United States. The project aimed to support researchers in the patient cohort identification and recruitment processes by providing them with a set of applications, called the Study Design Workbench. The Workbench assists the researcher in creating and designing new studies, using a standard template that includes all the required fields, defined by the World Health Organisation and the US National Institute of Health's ClinicalTrials.gov service. Additionally, the Workbench helps the researcher to define the study eligibility criteria and supports the translation of eligibility criteria into actual queries and interacts with the ePCRN Practice Gateway to provide counts of potentially eligible.

Requirements

- **Identification of subjects from clinical data** – counts of potentially eligible patients at selected sites are returned to researchers, who specify eligibility criteria on a Web-based workbench, thereby protecting the privacy of patients on the data source.
- **Appropriate security and privacy controls** – to prevent unauthorized access both for users and potential external parties.
- **Collection of clinical study data** – a system that allows for a more standardized case report form for remote, electronic data collection, with in-built validation and a more robust control of meaning, especially as data is passed from one system to another.

Methods

The ePCRN workbench consists of the Study Designer that enables researchers to specify study eligibility criteria to obtain counts of potentially eligible patients. An example screenshot is shown in Figure 2. The features of the Study Designer are described below:

- Inclusion and exclusion criteria are individually specified.
- Eligibility criteria are specified as five categories: demographics (Age and gender), clinical problems, laboratory tests, vital signs, and drugs.
- The Enterprise Vocabulary Service (EVS) interface enables users to search for terms in various terminologies and coding systems, as provided by the NCI Metathesaurus [9] (Label 1 in Figure 2).
- The query logic allows for conjunctive and disjunctive relationships. (Labels 2-4 in Figure 2).
- Values relevant to particular concepts can be specified to refine the query. These are specific to the criteria type and categories.
- Data sources can be selected before the query is run, as shown in Figure 3. The results are displayed as counts in individual sites.

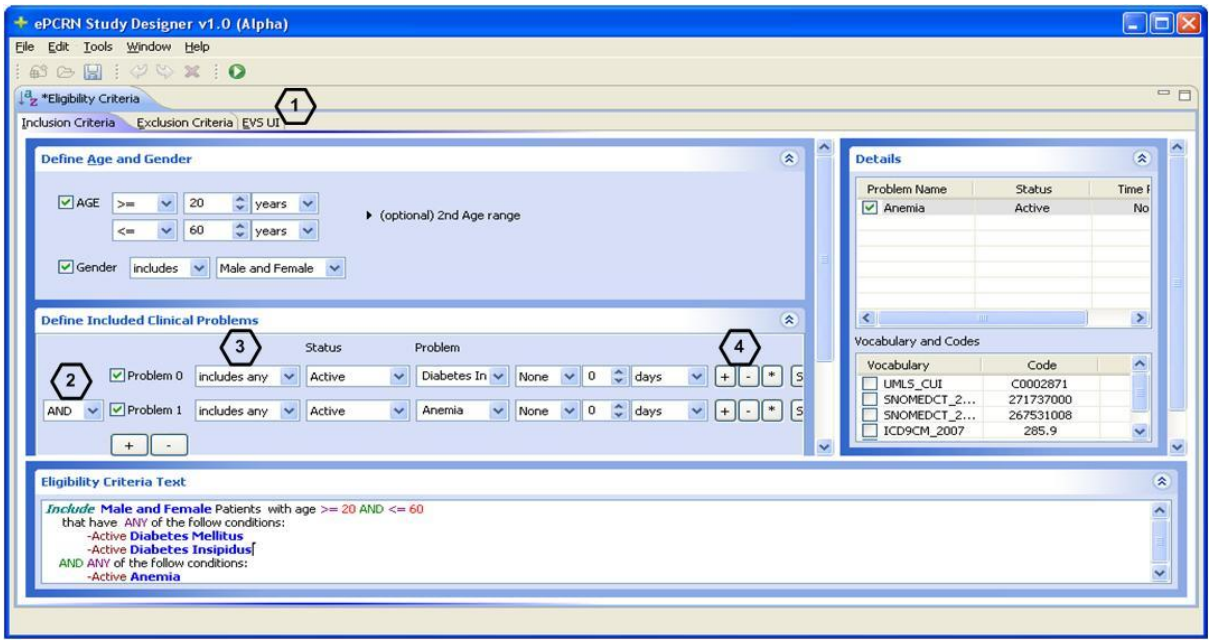

Figure 2: ePCRN Workbench

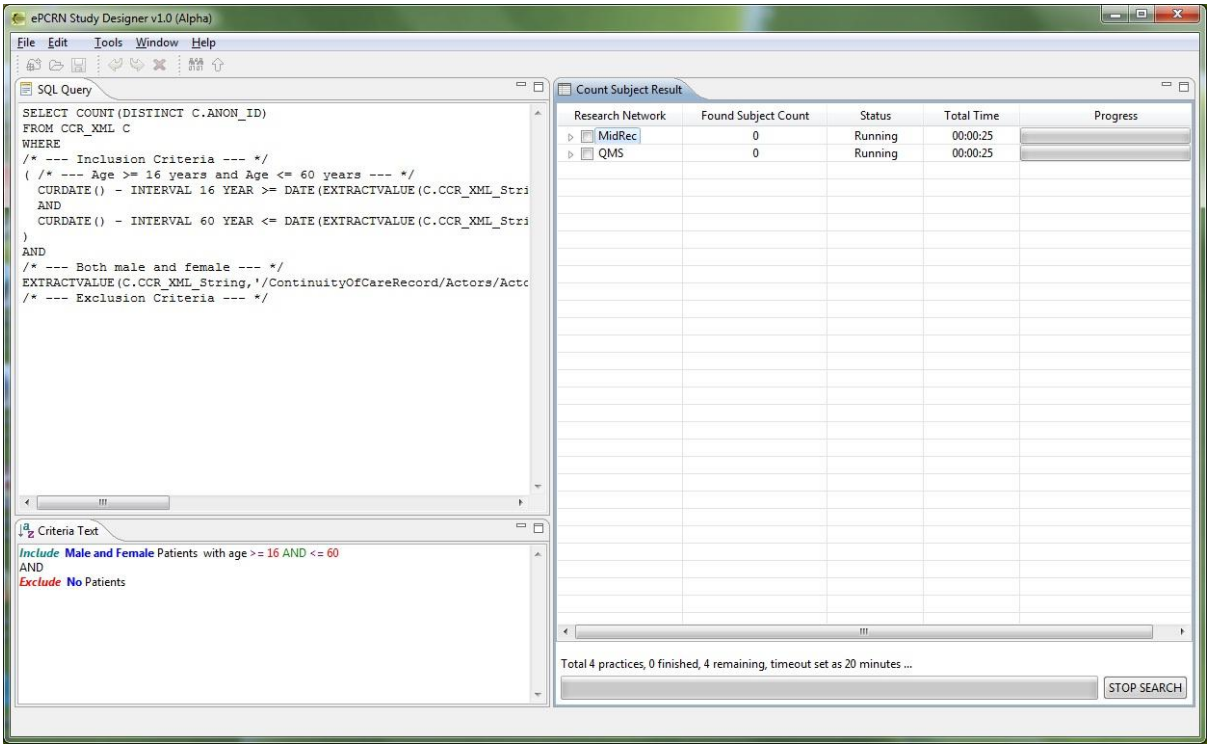

Figure 3: Data Sources Queried

2.2.2. i2b2/SHRINE

SHRINE is a prototype federated query tool for clinical data across multiple data repositories. It aimed to draw from the successes of i2b2 (Informatics for Integrating Biology and the Bedside – demonstrating the feasibility of using the data accrued during the course of healthcare for discovery research, for a single clinical data repository) and SPIN (Shared Pathology Information Network – a cross-institution data sharing system).

Requirements

The success factors for the SHRINE prototype were defined, with a development time and sponsorship of 6 months [10]:

a) A system approved by the Institutional Review Boards (IRBs) and the hospital executives.
b) The system can query in real-time, the clinical data repositories of at least three different health care centres.
c) Query interface allows searches on patient demographics and diagnoses. Other data types are less consistent in how the concepts are coded, and creating a common ontology for all SHRINE databases is time-consuming and therefore only demographics and diagnoses are included in this prototype.
d) The technical architecture has been selected based on the prototype and would need to be redesigned for scalability.
e) The prototype will return approximate aggregate counts and the minimum count threshold is 10. Counts less than 10 will appear as “less than 10” instead of the actual count.

Methods

The SHRINE user interface is divided into 4 modules as shown in Figure 4:

1) **Ontology module** that lists the hierarchical medical concepts as an expandable tree. The top two levels are demographics and diagnoses. The Find Term tab allows users to search concepts by name or code. The SHRINE demo also includes lab test and medication concepts. The i2b2 web client also contains additional concepts related to clinical trials, procedures [11].
2) **Query tool module** enables users to drag-and-drop concepts into panels to describe the population being searched. The concepts in the same panel are locally OR'ed, and the panels are themselves AND'ed. Panels can be negated by selecting the Exclude button. Date ranges can also be placed on a panel. A minimum number of occurrences can be specified for a panel – it indicates how many times a concept must appear in a patient's medical record.
3) **Query status module** displays information about the amount of time the query has been running and when the query is complete, it shows aggregate counts from each hospital.
4) **Previous queries module** lists the results of previous queries. Users can drag-and-drop items from this module to the Query tool to view the concepts used in the query.

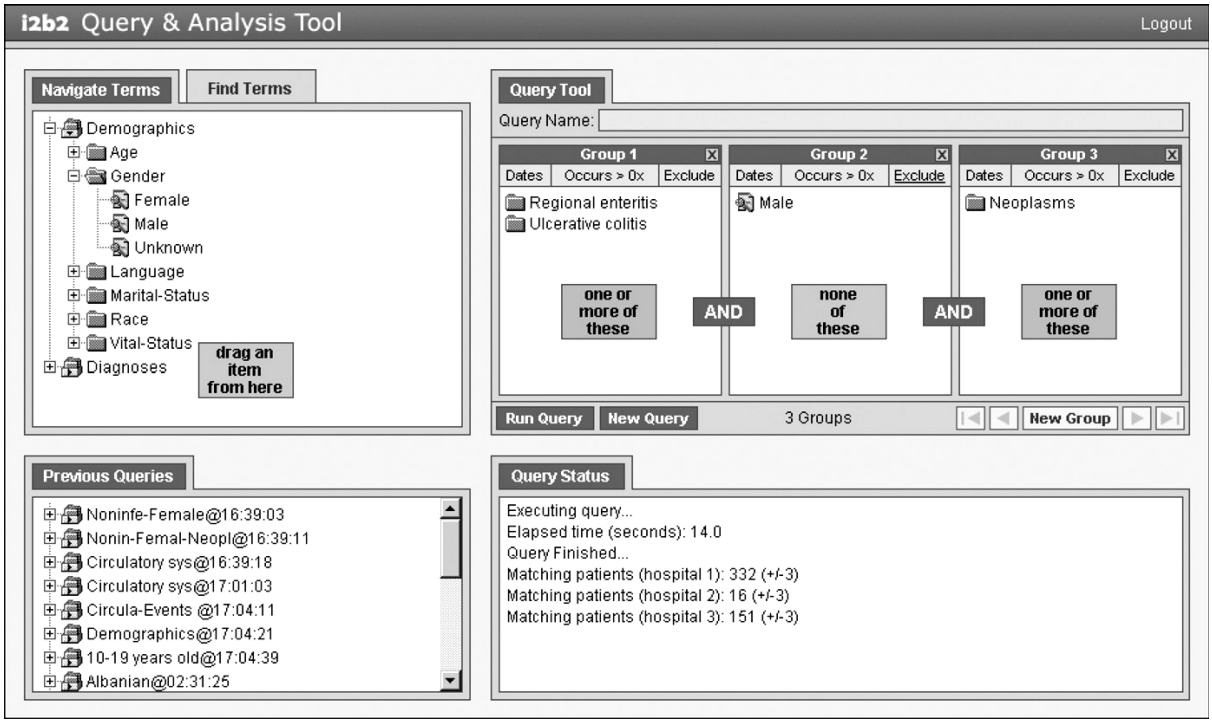

Figure 4: i2b2 Workbench [11]

The SHIRINE prototype architecture consists of a Query Aggregator and SHRINE Adapters at each hospital. The Query Aggregator has two parts: the first one is a web-based interface through which users can access the system. The second part consists of web services that broadcast the query to each Adapter and receive counts back for each hospital. The SHRINE Adapters are web services at each hospital that receive queries from the Aggregator and return patient counts. The purpose of the Adapters is to translate the query into a format that is compatible with the institution's source databases.

As the three hospitals were already planning to implement local i2b2 instances, each created an i2b2 structured database for the SHRINE prototype. This led to the design of only one Adapter that could be used at each location with minimal customizations since the Adapter would be connecting to an i2b2 database.

Values associated with laboratory tests can be specified by flag or by value, with the units specified. These are specified after the laboratory concept has been selected and dragged onto the panel. For medications however, dosage information are also concepts that need to be selected first from the Ontology module, dragged onto the panel, after which, a pop-up window allows values for the associated information to be specified. Demographics are also represented by concepts in the ontology. Age, for example, is categorized as groups, with individual age years within.

Shrine Core Ontology

The Core ontology was constructed following the recommendations of relevant government and private sector bodies and took a pragmatic approach, dependent on the availability of patient data when standards were selected. The absence of standards in particular domains required the creation of new ones [12]. The following standard terminologies provide the baseline for the Core ontology [12][13].

Table 1: SHRINE ontology and terminologies

Concept Category	Source Ontology	Concept Type	Terminology
Demographics	HITSP C32[1]	Age	
		Gender	HL7 Administrative Gender
		Language	ISO 639-1
		Marital Status	HL7 Marital Status
		Race and Ethnicity	CDC Race & Ethnicity Code Sets
		Religion	HL7 Religious Affiliation
		Vital Status	
Diagnoses	ICD-9-CM[2] Clinical Classification Software (CCS) hierarchy		
Medication	RxNorm[3] (Ingredient) National Drug File Reference Terminology (NDF-RT)[4] hierarchy		
Laboratory Test Data	LOINC codes Partners HealthCare System Hierarchy		

SHRIMP [14] is a tool developed to translate between local medical vocabularies and nationally recognized standards defined by HITSP, ICD-9 and RxNorm. It can also be used to create concept hierarchies to allow queries to occur on high level concepts, making it easier to query systems with a large number of concepts and aggregate medically related facts.

[1] HL7 CCD, http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=4&PrefixNumeric=32
[2] http://www.cdc.gov/nchs/icd.htm
[3] http://www.nlm.nih.gov/research/umls/rxnorm/
[4] http://nciterms.nci.nih.gov/ncitbrowser/pages/vocabulary.jsf?dictionary=National%20Drug%20File%20-%20Reference%20Terminology&version=February2011

Remarks

A typical query contains 2 or 3 concepts and takes 10-60 seconds on the prototype to construct. A single concept that matches few or no patients can run in as little as 5 seconds, while a query with many concepts that matches a large percentage of all patients can require up to 2 minutes [10].

SHRINE has been built around local policies, risking developing a product that works only in one institution. However, sharing clinical data for research requires both technical innovation as well as a cultural shift in attitudes towards collaboration. SHRINE has been developed to address the latter first to create greater impact and adoption of the software platform.

A more scalable model based on SPIN will be designed to allow for expansion of the SHRINE network. Plans also include the further development of the Adapter to support more complex ontologies.

2.2.3. FARSITE

FARSITE (Feasibility Assessment and Recruitment System for Improving Trial Efficiency) is an information system developed to support the evaluation of trial feasibility by providing accurate assessments if number of patients eligible for a particular trial. It also automates patient recruitment, whilst preserving consent for consent. FARSITE is being developed by North West e-Health, a collaboration between the University of Manchester, Salford Royal Foundation Trust, and Salford Primary Care Trust, UK [15].

Requirements

- Preserving consent-for-consent and patient-clinician relationship - The main requirement identified was to speed up feasibility-assessment and recruitment while preserving the patient-clinician trust relationship that is central to consent-for-consent [16]. **Consent-for-consent** is the consent to access medical records in order to identify eligible participants to approach to see if they would like to participate in a research study [17]. This applies to identifiable patient data, as anonymised or robustly pseudonymised/linked anonymised data may be used without consent [18]. However, in cases where patients are approached from their medical records, steps should be taken to make patients aware that personal information is used in research in the practice, and throughout the NHS, and that the care team includes research staff [19]. The automated feasibility assessment and recruitment is being compared to the ad hoc method, where clinical staff are asked for estimated of the number of patients with particular characteristics they are expected to see in a given period of time. The search for patients in databases and paper records is not systematic. The ideal research information system would parse the protocol, form a search query, and enable the study sponsor to assess the potential recruitment in a specific population while varying the inclusion and exclusion criteria. Therefore, the study design extends into feasibility, which might benefit the design. After the study is approved, the well-understood protocol and search mechanisms are employed in the same e-infrastructure to aid recruitment. Only at the recruitment stage is it necessary to know the identity of an individual patient [16].

- Improving research protocol design interactively - The balance of the need for tightly defined eligibility criteria against the need to get sufficient number of participants to achieve the required statistical power is important, when designing a research study

protocol. Hence, the need for a system that enables the user to progressively test and refine eligibility criteria until the correct balance is found. The definition of eligibility criteria must allow the user to select clinical codes quickly and easily. It must also allow for complex combinations of criteria using Boolean operators. Individual criteria should be able to be extracted with exact values or with a range.

- Unified process model linking trial protocol design and recruitment - Although there is a clear distinction between the actors involved in clinical trial design and clinical trial recruitment, the FARSITE authors argue that the connection between the two processes is essential to ensure that the efficiency gains do not compromise the best practice for privacy and consent.

Methods

An ideal trial protocol development and recruitment process has been captured from the subject matter experts [16] (details of their roles and experience not specified in this part). The requirements from the actors of clinical trial design are different from those of trial recruitment actors. These two sets of actors are essentially separated by a clinical care boundary, where recruitment actors (mainly attending clinician) can access patient-identifiable information, while design actors (mainly researchers or administrators) need to access only anonymised data. No patient-identifiable information should cross this boundary.

Counteracting Deductive Disclosure

In the trial protocol design process, the analyst progressively develops the protocol by issuing queries against the anonymised repository to determine how many patients will meet the specified eligibility criteria. Although these queries only return an integer count of matching patients, allowing users to issue a sequence of queries leaves the system open to deductive disclosure [20]. The FARSITE trial protocol design system will filter the results of queries to ensure that counts of less than 5 are returned as 5.

Preserving Consent-for-consent

The analyst is responsible for developing the protocol to obtain the patient counts. To preserve the consent-for-consent, it is the attending clinician who needs to run the query against the EHR system to identify patients. As the anonymised repository does not have information about the attending clinician, the query has to be transmitted across the clinical care boundary, identify the clinicians with potential recruits and rewritten so that specific queries are constructed for each clinician. The clinicians are notified of the active trial and some of their patients are eligible. The email contains a HTTP link, which when clicked, executes the query for that clinician (assuming successful authentication on the FARSITE system). The results are presented to the clinician as a form in a web browser and the clinician can select suitable patients and submit this information. The system collates the responses and emails the projected number of participants to the trial protocol designer. The trial protocol may need to be redesigned, but if the number of participants is acceptable, an email is sent to the clinicians informing them that trial recruitment can begin. The clinician can log on to FARSITE through a web browser and generate personalized letters and information sheets for each patient. When the trial is registered with the system, the system will autonomously run the query to test if new patients have become available since the last execution.

Remarks

A prototype Trail Protocol Designer tool has been developed to enable a user to construct Eligibility Criteria queries and retrieve the number of eligible subjects in an anonymised database of diabetic patients. Based on user feedback, improvements have been made to the user interface, including present the count of patients for each eligibility criteria. Plans to include a quick and easy way to find the correct clinical code, with suggestions based on an ontology, such as OpenGALEN [21], has resulted in a hierarchical display of Read Codes, grouped as Diagnoses, Procedures & Findings, Measurements & Tests, Allergies & Adverse Reactions and Other, as shown in Figure 5, a screenshot from [22].

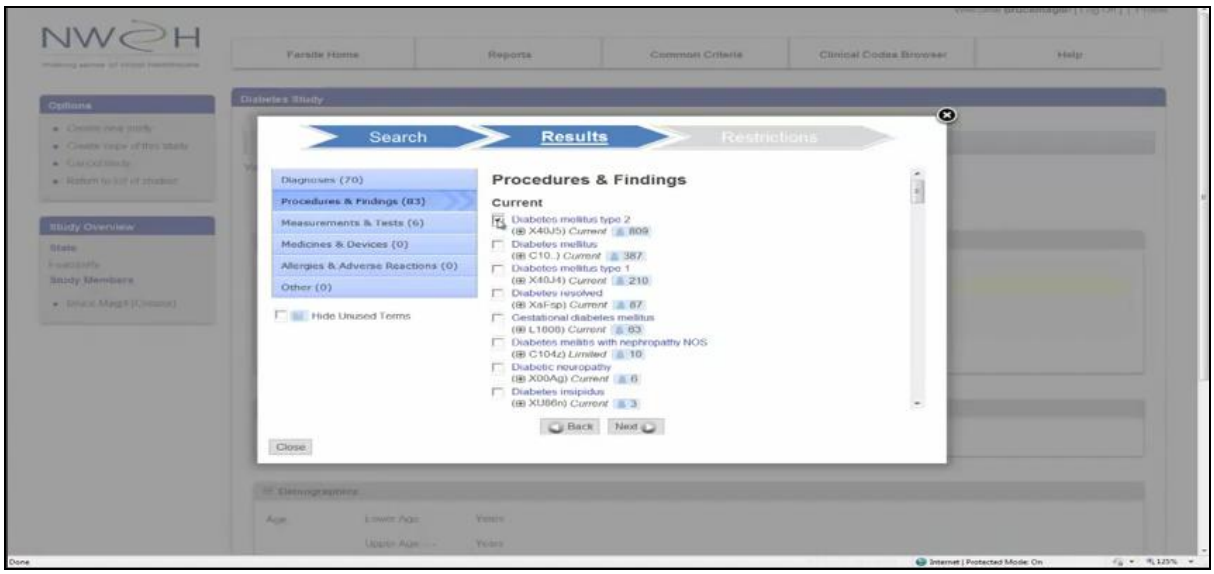

Figure 5: FARSITE terminology hierarchy - Procedures & Findings

The first deployment of FARSITE is planned in Salford in collaboration with the Greater Manchester Clinical Research Network (GMCRN) and the NHS in Salford. Salford NHS has an advanced EHR system, Salford Integrated Record (SIR), which integrates primary and secondary care data to form a single patient record. Salford is also deploying an e-Lab, a secure information system to contain the anonymised repository of patient data extracted from the SIR system. The Trial Protocol Design tool will be mounted within the Salford e-Lab for GMCRN researchers to use. The Trial Recruitment Tool will be developed as a standalone web application inside the clinical care boundary accessible only to clinicians.

Query UI Features

a) Separate inclusion and exclusion criteria sections
b) Criteria groups can be created, each with a number of individual criteria
c) Demographics is separate from the inclusion and exclusion criteria
    - Age (lower and upper value)
    - Gender
    - Ethnicity
    - Smoker status
d) For each criteria, there is a Search, Results, and Restrictions sections in the workflow
    - **Search** allows the user to start typing the term and click the Search button.

- The **Results** section then shows the results of the search, categorised into groups (Diagnoses, Procedures & Findings, Measurements & Tests, Medicines & Devices, Allergies & Adverse Reactions, Other), with the group with the most results shown. One or more concepts can be selected.
- The terminology used is SNOMED CT UK Release. The search results are limited to the first 300 terms. The description, Read code and status of the term are displayed. Additionally, the count of patients matching each term is displayed.
- The **Restrictions** section then allows the user to put constraints on the selected criteria. The constraints depend on the result/concept groups.
    i. For Procedures & Findings, and Medicines & Devices, a time period can be specified, either as having occurred before or after an exact date, or as a time period of more or less than a period of time in the past.
    ii. For Measurements & Tests, values can be entered with the operators (between, >, >=, <, <=, =) and the gender can be specified if needed. As well as the values, when the measure was taken can also be specified, as a time period, an exact date or between specific months of the year.

e) When the query is run, the count of patients is obtained, with breakdowns for each sub-criteria, as shown in Figure 6.

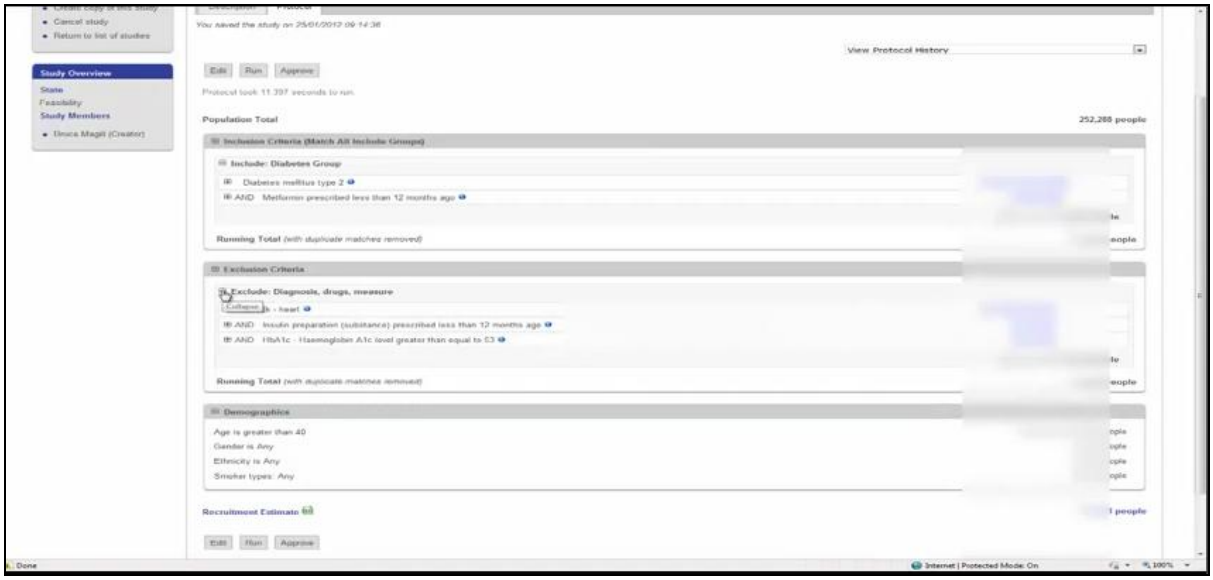

Figure 6: Breakdown of results by criteria

f) When the protocol is approved, these numbers are further broken down by GP practice with a map showing the location of the practices.

Recruitment UI Features

a) The user can select one or more GP practices and specify a date by which GPs need to respond.
b) The document manager option enables the user to attach relevant documents, e.g. patient consent form, information sheet, ethics approval letter. It also offers a document template to draft the invitation to the GPs and to send it to them with any other information, such as protocol restrictions, etc.

c) The GPs receive an email notification about the study and can then log into FARSITE and see relevant studies.
d) For each study they have been invited to participate in, information about the study, the protocol, the documents, identified patients, and an opt-out option is available.
    a. Patients can be selected for invitation and they are included for invitation letters to be printed out.
    b. The GP can also opt out patients for all future research.

The benefits of using the FARSITE application in comparison with previous modes of working have been analysed, following the deployment of the FARSITE software. A comparison of the estimates of numbers of patients eligible for the trials and the trials' actual recruitment rates shows a strong correlation between protocols with a low FARSITE recruitment estimate and trials struggling to recruit participants [23].

2.2.4. VISAGE

VISAGE (VISual AGregator and Explorer) is the query interface for the Physio-MIMI (Multi-Modality, Multi-Resource Information Integration environment), a multi-CTSA[5] site project funded by the National Center for Research Resources (NCRR), USA, designed to develop novel, flexible informatics methodologies, tools and infrastructure to facilitate the collection, management and analysis of clinical and physiological data [24][25]. Physio-MIMI has an expandable Sleep Domain Ontology (SDO); fine-grained interface for role-based data-source level access control; plug-and-play Honest-Broker adaptor to mediate data access services; and data schema to a Sleep Domain Ontology mapper that transforms a local database into a data resource that can be queried using the federated query interface VISAGE on the web.

Requirements
The uses cases include:

- The identification of a research cohort that meets specific demographic, physiological, and clinical criteria and the subsequent identification of a second similar cohort to serve as control subjects.
- A user needs to be able to specify the selection criteria and quickly obtain results of the number of records available in the selected data repositories, to save the query, and repeat the query modifying one or more criteria in order to identify the second cohort.

Additional requirements identified through the needs analysis are:

- Clinical researchers want to be able to identify clinical criteria for use in the query based on clinical and logical terminology, not technical or database schema terminology.
- The interface needs to allow for searching for available terms based on a number of synonyms.

[5] CTSA – Clinical and Translational Science Awards, http://commonfund.nih.gov/ctsa/index.aspx

- Users want to receive immediate feedback on the counts returned by the query rather than having to submit and wait each time criteria are adjusted.
- Users want to be able to direct a single query to one or more underlying sources of data without explicit knowledge of the different database structures.
- Users want to be able to save and reuse queries to avoid having to repeat the process of specifying very detailed criteria in order to change a single aspect of the query.

Methods

Query Builder Interface

The query builder interface includes the following functional areas, as shown Figure 7 and Figure 8:

- Database selector: choose the databases against which to run the query (Labels 1, 17)
- Term search bar and term selection areas, based on the Sleep Domain Ontology (Labels 2-3)
- Term display area: when a term is selected (Label 4)
- Terms can be grouped or ungrouped by selecting the terms first using the checkboxes (Label 5)
- Conjunctive and disjunctive relationships apply, with the Flip action to reverse the AND/OR operators. ANDed terms are in green and logical ORs in blue (Labels 6, 14-15)
- Terms can be rearranged with drag and drop functionality (Label 7)
- Terms can be deleted (Label 8), and groups as well (Label 13)
- Concept categories and constraints are derived from the SDO; available categorical data are indicated by checkboxes (Label 9), while continuous variables use sliders (Label 10)
- The number of records that satisfy the conditions are displayed in the Result Count Area (Label 11)
- Users can describe, save and update the query to the Query manager (Labels 12,16)

Remarks

A preliminary evaluation of the efficiency of the VISAGE query construction was performed [24]. Three common queries with increasing levels of logical complexity on patient demographics were selected and two expert users created queries in both VISAGE and i2b2 web client. The number of clicks and time needed for creating the queries were recorded (Table 2).

Table 2: Evaluation results [24]

Query	VISAGE		i2b2 Web Client	
	# of clicks	time (sec.)	clicks	time (sec.)
1	5	13	14	59
2	6	16	25	119
3	20	52	37	160

VISAGE reduced time and effort, in terms of the number of clicks, to a half or nearly a third. However, this preliminary evaluation has looked at only one specific aspect of the query interface.

VISAGE has focused on an intuitive, usable and simple interface. The query interface development used agile and user-centred methodologies and rapid prototyping was achieved through the use of various open source web development tools and frameworks, including Ruby on Rails, Prototype and script.aculo.us JavaScript libraries.

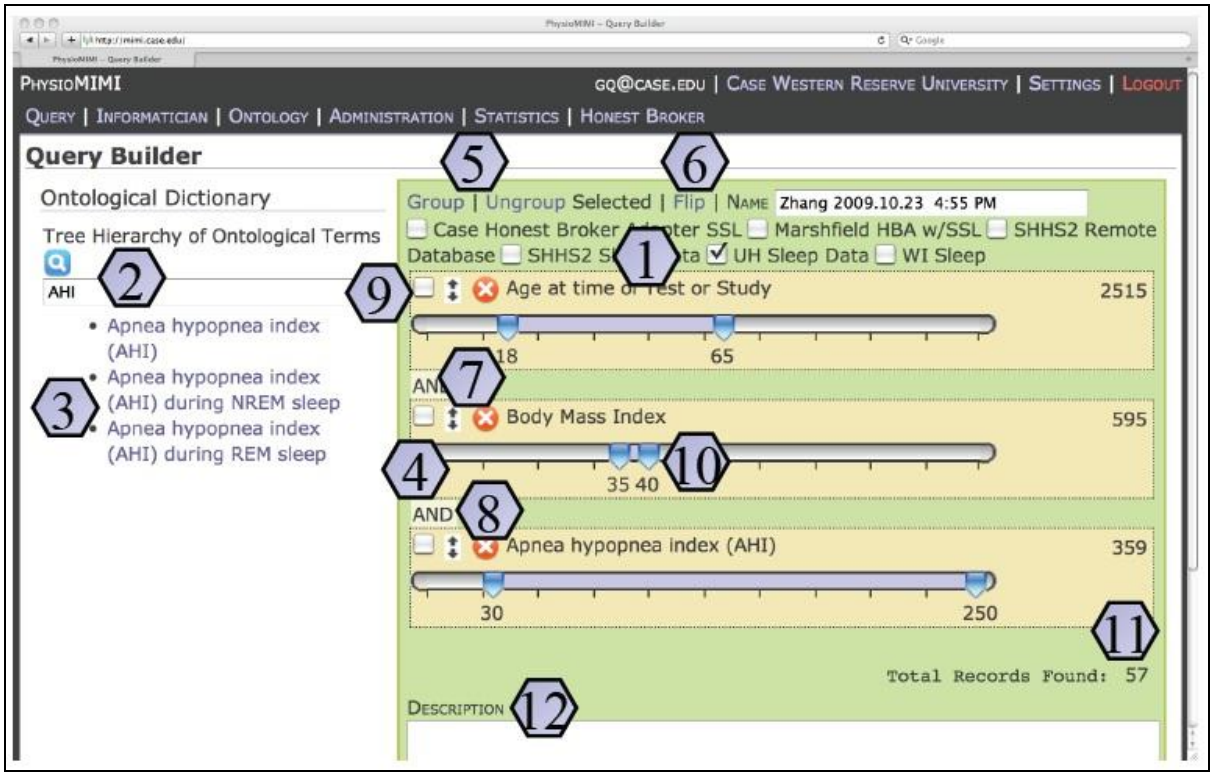

Figure 7: Query builder interface [24]

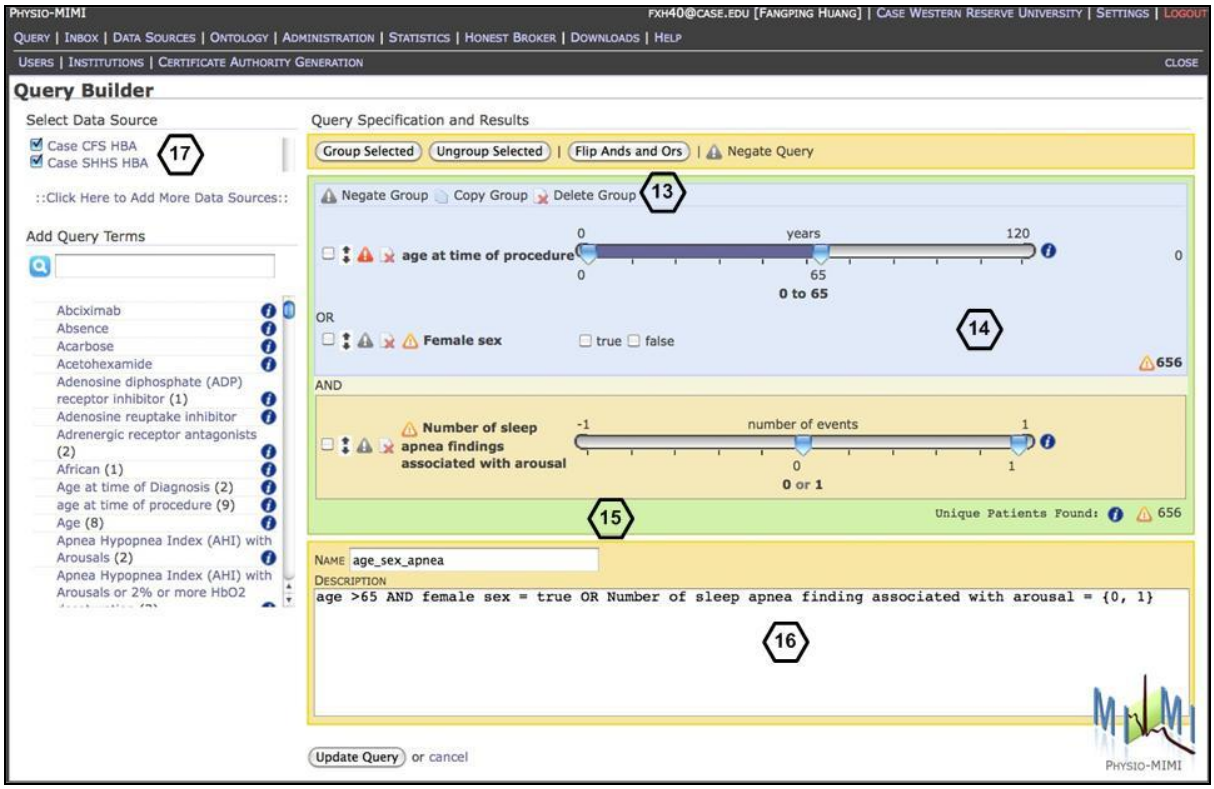

Figure 8: Query interface example [Adapted from [26] Figure 1.10]

2.2.5. Comparison of state-of-the-art query tools with the TRANSFoRm Query Formulation Workbench

In this section, we discuss, by comparison, the features and limitations of the query tools reviewed above, present the differing requirements of query formulation, data extraction and linkage in TRANSFoRm. Through this discussion we elaborate on how TRANSFoRm seeks to address the limitations of the above described state-of-the-art query tools.

Multiple Data Sources

The TRANSFoRm project aims to facilitate Europe-wide studies, and therefore needs to support multiple heterogeneous data sources, such as GPRD and NIVEL primary care data sources, without processing the data into a common format. In comparison, FARSITE uses the Salford Integrated Record, which already integrates records from different primary care system suppliers. The ePCRN infrastructure relies on data sources that are structured as continuity of care records (CCR) [27]. However, i2b2/SHRINE and VISAGE use i2b2 instances that use adapters to connect to heterogeneous databases.

Deductive Disclosure

Most of the reviewed query systems provide means to counteract deductive disclosure, to maintain patient privacy. An agreed minimum count is returned, which prevents result counts from numerous queries to potentially identify patients. In TRANSFoRm, the prevention of deductive disclosure will also be achieved and this will be handled by a thresholding mechanism for an agreed minimum count to be returned.

Terminology Usage

All the query tools reviewed use medical terminologies or ontologies to represent concepts. FARSITE uses SNOMED CT as the only source vocabulary for the regional UK primary care records. VISAGE uses the Sleep Domain Ontology, while i2b2/SHRINE caters for local terminologies as well as the Shrine core ontology. ePCRN uses the NCI Metathesaurus in the English language due to the context of the usage in the USA. TRANSFoRm requires reference to a number of different source vocabularies in different European languages. The TRANSFoRm Vocabulary Service uses the UMLS Metathesaurus as well as other vocabularies not supported by UMLS, such as Read Codes version 2, ICPC2 (version 2012), etc..

Search and browse functionality are the most common features available in query tools. The TRANSFoRm Vocabulary Service enables the searching of terms and returns results in the corresponding source vocabularies. Parent and child terms are also viewable, based on the UMLS hierarchy of concepts.

FARSITE is the only tool from the group to display the patient count matching concepts while using the terminology search and browse functionality. This feature is especially useful when searching only one source vocabulary.

Eligibility Criteria Specification

The query tools we have considered in the review enable users to define the eligibility criteria for a study protocol either in fixed groups or according to needs. For instance, FARSITE and ePCRN allow specification of inclusion and exclusion criteria and demographics separately. Additionally, ePCRN enables users to define eligibility criteria in fixed groups, such as problems, medications and laboratory tests. i2b2/SHRINE and VISAGE allow flexible addition of criteria. Similarly, the TRANSFoRm query formulation workbench will enable users to define criteria groups flexibly, whilst catering for complex queries with combinations of operators.

The attributes associated with the selected concepts help to define the criteria further. For example, laboratory measurements include the laboratory test result concepts (e.g. fasting blood glucose), the result value (e.g. 7.0) and the associated unit (e.g. mmol/l). Like many of the other query tools, the TRANSFoRm query formulation workbench will use the CDIM ontology and archetypes to determine the associated attributes for each concept. Therefore, the query formulation workbench can present the relevant options to the user while they design their study protocol, on the fly.

Query Results

The results of a query return patient counts matching the criteria defined. Results broken down by eligibility criteria give the most detailed information about the distribution of patients and enable users to update study protocols accordingly. i2b2/SHRINE displays total count per site and also has some predefined categories for results breakdown. There are plugins available for displaying demographics results breakdown. ePCRN also displays total counts per site. Similar to FARSITE and VISAGE, the TRANSFoRm query formulation workbench will display results as a total count, as well as broken down by eligibility criteria and criteria groups.

FARSITE has a map view of results to better identify locations. TRANSFoRm can also provide this feature as the information for geographical location within Europe is available for the data sources. If the user has selected the query to be run against multiple data sources, the results can also be filtered by data source.

Collaborative Work

A common feature of the existing query tools is to allow users to create, update and share protocols. In ePCRN, for instance, updates can be made to a protocol but the changes are not linked to the users who brought the changes. In TRANSFoRm, users can work collaboratively on a study and create versions of the protocol. All changes will be logged and query results will be stored.

The functionality for users to retrieve their work from the tool to use elsewhere is not often specified in the existing tools. In ePCRN, the protocols are backed up but not available offline for users. VISAGE offers statistical, metadata and sleep-related data for researchers to use. So far in TRANSFoRm, the requirement for exporting files related to protocols and results has not been expressed by users; however, this feature has been identified as useful in the literature review and discussions with expert users.

Patient Recruitment

i2b2/SHRINE and VISAGE are not recruitment tools, while FARSITE and ePCRN have this feature. The recruitment of patients follows the feasibility assessment of the study using the query tool. In TRANSFoRm, it is planned for patients matching study eligibility criteria to be flagged for consideration for study recruitment.

Table 3 below summarises the features of the reviewed state-of-the-art query tools, in comparison to the TRANSFoRm query formulation workbench.

Table 3: Comparison of Query Tools

Feature	TRANSFoRm	ePCRN	i2b2 / SHRINE	FARSITE	VISAGE
Multiple data sources	Yes - heterogeneous data sources across Europe	Yes – Data sources are research networks	Yes – enabled by SHRINE	No – use of the Salford Integrated Record	Yes
Prevention of deductive disclosure	Yes	Yes	Yes	Yes	Not specified
Use of terminology	Yes - Search; UMLS Metathesaurus and other source vocabularies	Yes – Search; English NCI Metathesaurus	Yes – Browse and search; local and standardised	Yes – Search then browse; SNOMED CT	Yes – Sleep Domain Ontology
Hierarchical view/browsing of terminology	Yes - Parent and children concepts can be viewed for UMLS Metathesaurus	Yes - Parent and children concepts can be viewed	Yes – Shrine core ontology	Yes	Yes – similar to i2b2
Term search results grouped by type	No - the concept semantic type is available in UMLS	No – the concept semantic type is available in UMLS	No – for Search, the term category can filter results	Yes	No
Separation of inclusion/exclusion criteria and demographics	No	Yes – via fixed groups	No	Yes	No
Eligibility criteria group building	Yes	Yes – via fixed groups	Yes	Yes	Yes

Feature	TRANSFoRm	ePCRN	i2b2 / SHRINE	FARSITE	VISAGE
Term/criteria values	Yes - dependent on concept types as defined in the CDIM ontology	Yes – within the criteria groups	Yes – some dependent on the term; date and frequency for all	Yes – dependent on the criteria group type	Yes – term-specific controls dependent on ontology
Patient counts per term (irrespective of criteria values)	No	No	No	Yes	No
Patient counts per eligibility criteria	Yes	No – Only total count per site	Yes – Some fixed groups and plugins for demographics breakdown	Yes	Yes
Map view of sites	Planned - Location information available	No	No	Yes	No
Collaborative work	Yes - users can collaboratively work on studies and queries	Yes – but changes are not associated with users	Yes – queries can be reused and shared	Yes – protocols can be updated, approved by users	Yes – user sharing feature within the query profile panel
Retrieval of file sets	Not required so far	Not for researcher use	Not specified	Not specified	Yes – statistical, metadata and sleep-related
Recruitment function	Planned - Flagging of eligible patients	Yes	No	Yes	No

2.3. TRANSFoRm Query Formulation Workbench Requirements List

In this section, we provide the list of functional and user requirements for the TRANSFoRm Query Formulation Workbench tool, as derived from the combined analysis of the use cases provided by deliverable D1.1 and the information elicited by conducting participatory task modelling, involving a group of expert users. In addition, this list has been enhanced by a critical and comparative literature review of relevant state-of-the-art eligibility criteria query tools.

Users and study query formulation

- Every user will have a unique login and password to the system.
- Each study created in the system can have multiple queries (query versioning).
- Information about the study includes name, reference numbers and description.
- Multiple users can work collaboratively on a study and its associated queries. Only one user can modify and run a query at any one time. User activity will be logged.
- Queries can be saved and edited at a later time.

Available data source properties

- Users registered to the system have access to the properties of the available data sources depending on their permissions. A mechanism will be in place for users to request those permissions (at least a contact).
- Users can view the data source properties: name, organization, country, region, coding schemes, population size, data extract date. The issue of vicarious identification and data source/search granularity needs to be considered.

Eligibility criteria design

- Eligibility criteria groups can be created and combined to represent the study protocol.
- Each group represents an inclusion or exclusion criteria group.
- The eligibility criteria group will have a name and a number of individual criteria added.
- The attributes the user can specify about the criteria depend on the type of criteria being entered. For example, if it is a medication criterion, the dose, unit, duration, and method, for instance, are the attributes that can be specified to put restrictions to the search. Whereas, for a diagnosis, a temporal attribute to indicate when this diagnosis was made is available to the user. As concept attributes are linked to data type, the user will need to specify a CDIM concept first, and then specialize to a specific concept using the vocabulary service.
- The CDIM concept will inform the concept attributes that are associated with the concept type.
- It is assumed that only one set of attributes can be specified for a concept. For example, for blood glucose laboratory test result, the user can define values >= 7 mmol/l. This value can be equivalent to >= 126 mg/dl, but since unit conversion will be available, the user does not need to specify more than one value/unit pair.

Medical vocabulary selection

- When formulating a criterion, the user first chooses the kind of concept he is looking for from the CDIM concepts. The CDIM concept helps the query formulation workbench define what attributes are associated with the concept for display. For instance, a Laboratory Test concept can have temporal and value attributes.
- The user then refines the concept by using the vocabulary service to access the clinical terms and their corresponding codes in different coding schemes.
- The user will type in a keyword and search for available terms.
- The results of the vocabulary search will be displayed in result groups (concept types) and hierarchy of concepts enabled for the results returned to view child concepts. As the user would have specified the CDIM concept previously, a mapping solution to the vocabulary/UMLS concept can allow the restriction of vocabulary search results by concept type.

Results display

- The count of patients is returned matching the eligibility criteria as a total.
- The count of patients is broken down by eligibility criteria group.
- The count of patients is broken down by individual criterion. The same task can be achieved by running queries with and without individual criteria and returning data per match may extend the time taken to run queries.
- The patient count can be filtered further, by country, region, organization, where applicable, with consideration of vicarious results.

The following sections discuss how this list of requirements is supported by the overall system architecture of the query formulation workbench (section 3); how these requirements are translated into an interface design and workflow (section 4); and finally how these requirements are implemented technically into a software tool (section 5).

3. TRANSFoRm Query Formulation Workbench Architecture

This section gives a brief description of the system architecture and its functional components, which form the context for the design and implementation of the TRANSFoRm query formulation workbench, as defined by elicited requirements discussed in section 2.

3.1. Conceptual Architecture

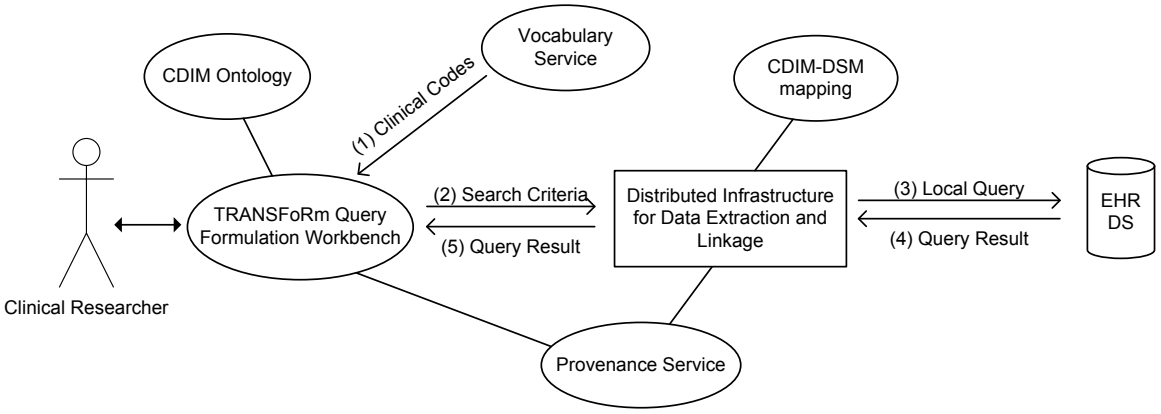

Figure 9: Conceptual Architecture

TRANSFoRm aims to support the whole clinical research cycle: from research protocol design, feasibility assessment to subject recruitment and study data collection. An essential part of this workflow focuses on supporting clinical researchers in the design of appropriate eligibility criteria for identifying subjects to be recruited in research studies. Such criteria are formulated as computational queries across distributed and/or federated EHR data sources. As described in Figure 9, each EHR data source is controlled and administered by a distinct organisation, namely a *data controller*. Clinical researchers design eligibility criteria using the user interface (UI) of the TRANSFoRm query formulation workbench. A specially designed vocabulary service, the TRANSFoRm Integrated Vocabulary Service (see section 3.3), allows researchers to browse and select standard clinical concepts. The vocabulary service provides mappings from standard UMLS concepts to standard EHR coding schemes[6].

The eligibility criteria are captured in a computable representation, based on the CDIM ontology; CDIM captures an extensible common representation of clinical care data and, as such, is generic and not dependent to the local EHR data source implementation schemata

[6] The TRANSFoRm Integrated Vocabulary Service can be browsed on the web at the following URL: http://www.eutransformvs.bham.ac.uk:8080/evs/

(see section 3.3). The computable representation of the eligibility criteria is then submitted to the distributed infrastructure for data extraction and linkage. The distributed infrastructure securely distributes the various search requests to individual data sources and coordinates their execution.

With the help of CDIM-DSM (Data Source Model) mappings, the criteria are translated into executable query statements, which can be run on local EHR systems. Results are pushed back to the infrastructure, which then encrypts them and subsequently delivers them to the query formulation workbench. Finally, the workbench aggregates individual results and presents them to the user.

Provenance information of the entire query workflow is captured by the distributed provenance system (Deliverable 5.2 TRANSFoRm Provenance Tool [28]), deployed alongside the federated query infrastructure. This conceptual architecture enables loose-coupling between data sources and the query formulation interface, while it allows for flexible implementation options on how the query workbench will work with individual EHR data sources. The following sections discuss in some further detail some of the important functional components of the overall conceptual architecture.

3.2. Query Formulation Workbench

The query formulation workbench brings together information from diverse data sources and provides a central interface for users to access a variety of TRANSFoRm applications and services. Specifically, the workbench provides a user interface for clinical researchers to create clinical studies, design eligibility criteria, initiate distributed queries, monitor query progress, and report query results. Through the workbench, clinical researchers can browse the TRANSFoRm Integrated Vocabulary Service, search for appropriate criteria concepts and bind clinical codes in these criteria concepts. In addition, researchers can use the federated query infrastructure to remotely search EHR data sources, document the findings and the overall workflow in the provenance system for auditing purposes.

The development of the query formulation workbench system is driven by the TRANSFoRm Clinical Research Information Model (CRIM). Therefore, a model-based approach was adopted for the development of the system. CRIM provides a standard, computable information model of clinical studies, conducted in primary care settings, and represents the baseline reference model for real-world design and implementation of systems that can be used to support all aspects of clinical trial design, execution, analysis, and reporting in the complex environment of primary care [29]. The query formulation workbench implements the CRIM eligibility criteria model, which specifies the information constraints on the formulation of inclusion/exclusion criteria. A detail discussion on how CRIM is implemented in the current version of the system is available in section 5.4. The query formulation workbench UI is implemented as a web application, using modern, standards-compliant, JavaScript and Ajax technologies, in order to deliver a rich user interface, readily accessible from diverse client platforms (further detailed discussion available in section 5.2).

3.3. CDIM Ontology and Vocabulary Service

TRANSFoRm employs an ontology-driven mechanism to address the data model heterogeneity issues and facilitate data interoperability among EHR data sources. At the core of this mechanism is the CDIM ontology. CDIM is an extensible and common representation of clinical care data, which abstracts away the complexity and variance in different source data and provides a unified coherent view of clinical concepts. CDIM is the pivot mechanism to view and exchange data throughout the TRANSFoRm system. The query workbench captures eligibility criteria in a computable representation, which is based on CDIM ontology so the criteria can be translated into executable query statements on the data source side using CDIM to data source model mappings.

The vocabulary service is another key component of TRANSFoRm, enabling data interoperability and it is an important associate service to the query formulation workbench. A vocabulary-controlled mechanism is necessary for an encompassing system to link many diverse data sources, where varied coding schemas, even local non-standard codes, are used. The TRANSFoRm integrated vocabulary service provides a unified platform to deliver terminology mappings and is shared by all entities in TRANSFoRm [30]. A centralized vocabulary service is easy to reuse, maintain and evolve.

The vocabulary content of the service is based on a Unified Medical Language System (UMLS) Metathesaurus subset of clinical concepts and associated terminologies relevant to primary care clinical practice and research. It also contains other terminologies and associated coding schemes, which they are not currently included in the UMLS Metathesaurus, such UK Read Codes version 2 and ICPC2 (version 2012). The UMLS Metathesaurus is a comprehensive multi-lingual biomedical terminology database, which covers many terminologies used for clinical care, basic and translational research, while it provides cross-mappings between these source terminologies [31]. The vocabulary service covers the most commonly used coding schemes in European primary care systems, including SNOMED CT, ICD-10, ICPC, ICPC2, Read Codes and others. In addition to English, the service also supports many European languages such as French, German, Dutch, Spanish. The vocabulary service provides, both a web interface and a web services API, for users to search and retrieve clinical concept definitions, codes and their associated mappings.

The first version of this service was implemented using LexEVS 5.1 and was based on UMLS Metathesaurus 2010AA release. The current version of the service has been upgraded to LexEVS 6 and is based on UMLS Metathesaurus 2012AA release. LexEVS is an open source general purpose terminology service solution and LexEVS 6 supports HL7 CTS 2 Draft Standard for Trial Use [32].

3.4. Distributed Infrastructure for Data Extraction and Linkage

The distributed infrastructure for data extraction and linkage enables communication between various TRANSFoRm applications, including but not limited to the query formulation workbench and EHR data sources. The infrastructure involves many distributed components, but as a whole provides federated secure data access and query of EHR data sources and

research databanks. Among others, the infrastructure includes the security framework which enables policy-driven authentication and authorisation. The query formulation workbench uses the infrastructure to invoke distributed query execution, monitor query progress and retrieve query results.

The distributed data extraction and linkage infrastructure provides secure messaging channels where all query requests and query results are encrypted. When clinical researchers design eligibility criteria and instruct the workbench to query selected EHR data sources, the workbench invokes the infrastructure API which encrypts the request using the security library, as described in deliverable D3.3 TRANSFoRm Security framework. The encryption of the query request happens before any external communication takes place.

The infrastructure implements an asynchronous messaging model, where each data source polls the infrastructure and uses the infrastructure API and security library to retrieve and decrypt query requests. Eligibility criteria are translated into local query statements suitable for individual data sources with the help of CDIM-DSM.

Figure 10 depicts the scenario of the interaction between the query formulation workbench and the TRANSFoRm distributed infrastructure for data extraction and linkage. When a query is run, the result is encrypted and pushed to an approved study data repository (Step 1). The data source uses the distributed infrastructure API & security library to pull the query from distributed infrastructure services. (Step 2). Following this step, the query is run and the result pushed to temporary encrypted data storage (Step 3). After a query is submitted, the workbench can request status updates using the infrastructure API (Step 4), informing the user if new results are available. If new query results are available, the workbench retrieves them from the data storage using the infrastructure API and aggregates the new result with existing results (Step 5). In addition to providing secure messaging, the distributed infrastructure interacts with the TRANSFoRm provenance service to capture relevant provenance information at each stage of the workflow.

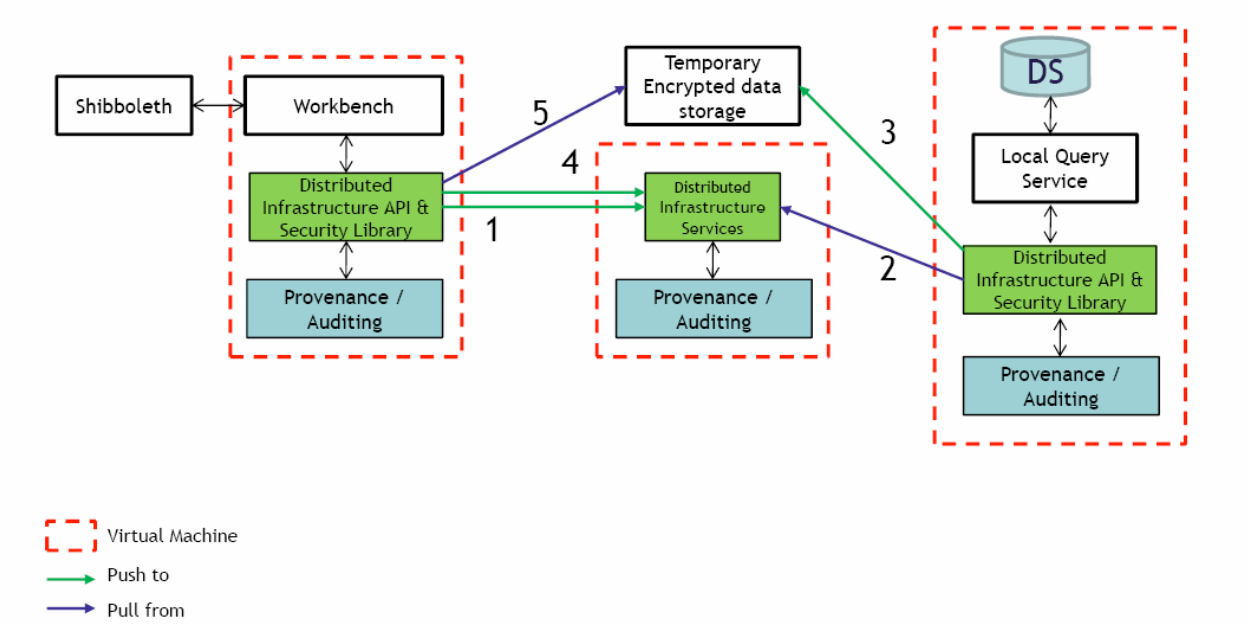

Figure 10: Scenario of interaction between the Query Formulation Workbench, the distributed infrastructure, security framework and provenance tool.

The following section discusses how CDIM handles the issue of the heterogeneity at the level of individual data sources.

3.5. Data Source and CDIM Mapping

Heterogeneous data models, at the data source end, are handled by CDIM ontology by a mechanism of data source model mappings to CDIM. The CDIM ontology hides specific details of each data source and provides a unified and consistent view of data elements. Each data source schema is captured in a representation which can be mapped to CDIM ontology. More specifically, database schemas and CDIM mappings are described in an ontology format and imported in a LexEVS instance. These mappings allow flexibility in defining how local objects are related to CDIM concepts. CDIM-DSM is then leveraged to translate generic queries based on CDIM into locally executable queries.

3.6. Provenance Framework

The TRANSFoRm provenance framework keeps track of the origin of various data and its evolution in different stages of the clinical studies [33]. Making TRANSFoRm systems provenance aware, one can enable the investigation of data sources and the services that produced a particular output, together with the individuals who instigated the requests and received the outputs. In such a way, user behaviour and data manipulation can be audited, to assess that correct decisions were made and appropriate procedures were followed.

Data privacy, legal and ethical regulations restrict provenance data from being stored in a central repository. The provenance framework mirrors the distributed EHR data access infrastructure, by implementing a decentralised platform for provenance capture, storage and querying. The query formulation workbench invokes the central provenance service to store provenance data for query execution and eligibility criteria design. This is achieved by using provenance templates for query execution process and eligibility criteria design process, respectively. More details about the provenance service can be found in [28].

3.7. Summary

This section provided brief description of the system architecture and its functional components, which form the context for the design and implementation of the TRANSFoRm query formulation workbench. The workbench integrates various TRANSFoRm services and presents a unified user interface for managing all aspects of eligibility criteria design and querying for eligible subjects. This service provides the essential support for clinical researchers to plan research studies participant recruitment and assess study feasibility at early stage. The following section discusses the query formulation workbench UI design and workflow.

4. Query Formulation Workbench User Interface Design

The Query Formulation Workbench UI design and workflow are described at this section, based on the functional and user requirements identified in section 2.3 and with reference to the other components of the TRANSFoRm digital infrastructure.

4.1. Interface Workflow

In this subsection, we discuss the UI workflow and describe it textually and with the use of UML activity diagrams.

All user interactions with the query formulation workbench are in the form of standard web application interactions, such as mouse clicks and simple text input. All interfaces are to be intuitive and standards compliant, while still retaining the functionality required for rich query design and analysis. Minimal training will be required to operate the web-based tool, which will operate across a wide range of devices ensuring compatibility across various types of user hardware.

Use of standard HTML visual elements, such as buttons, text boxes and dialog boxes ensures user familiarity. Sections of the UI are colour coded for consistency and navigation is simplified through clear menus and obvious button placements.

In the UI sections of “study creation” or “query design” users are presented with options to create elements, such as studies and criteria groups, only when requested by the user. At all times, the standard workflow of item creation, user data entry and final submission is adhered to. At no point is the user presented with a prompt for data entry unless a request has already been made to create an element (such as a study or criteria group).

Specifying a criteria group

Upon user request, users are guided through the process of creating a set of query criteria, including sub groups, each potentially containing multiple criteria or further nested sub groups. The UI is designed such that any number of sub groups can be nested, allowing users to design complex queries with rich logical structures. Creation of a single criteria group and criterion is described in Figure 11.

User specification of an eligibility criteria group and criterion requires the user to first add a criteria group. A UI element, such as a button is clicked by the user, triggering a prompt to appear requesting the user to enter more information about the criteria group, such as name or description.

After the creation of a group, the user can now add individual criteria, again using a UI element such as a button, which triggers another prompt for user entered data to appear. Alternative approaches to user data entry include simple searching for terms using the TRANSFoRm integrated vocabulary service or more advanced prompting for criteria attributes, using a set of pre-defined CDIM concepts. In Figure 11, a CDIM concept is first selected by the user, modifying the UI to display a prompt for searching for a specific term. Searching for a term communicates with the vocabulary service to match user input against

the list of available concepts. If search results are found, a user selects a specific term, which updates the UI.

Approaches to updating the UI appearance, include adapting the UI upon selection of a search term to show user data entry fields for a list of criteria attributes as determined by CDIM, or alternatively, the user is simply prompted to fill in a set list of attributes, regardless of search term result. In Figure 11, the term searched for and selected by the user is used to determine the list of attributes. The user then fills in all relevant criteria attribute fields, including any that may be data source specific, and completes the process of adding a single criterion. The query formulation workbench UI is then updated to show the newly created group and criteria contained within.

Running a query

Once the structure of a query has been designed, using the query formulation workbench, it can be stored for future use or submitted directly. Figure 12 describes the overall process of running a previously designed query.

Once a user is authenticated and authorised to run a query they can follow the process of submitting a query and displaying the results on the screen. Although many TRANSFoRm components, including the distributed infrastructure, are used in the query submission process, the user experience is constrained entirely within the query formulation workbench interface and none of the background processes are visible to the user.

In the workflow described in Figure 12, a user selects a study and is directed to a list of queries already designed for that study. When a predefined query is selected, the user is prompted to select from the available data sources. On confirmation of data source choice the distributed infrastructure is used to translate and process the query to a suitable format for each data source.

Submission of a query by the user will update the UI to display the progress of the query. The amount of information given to the user on query progress depends on the ability of the distributed infrastructure to support partial results. An alternative to displaying partial results is to only show the results on query completion. In either case, the UI can either be updated automatically, or on user prompted update, to show the status of the query.

On completion of the query the user is redirected to a full results view, where all patient counts are shown for each individual criterion and group, as designed by the user when the query was created. If information on country, region and organisation is available, the user can further filter the query results.

Specifying eligibility criteria group

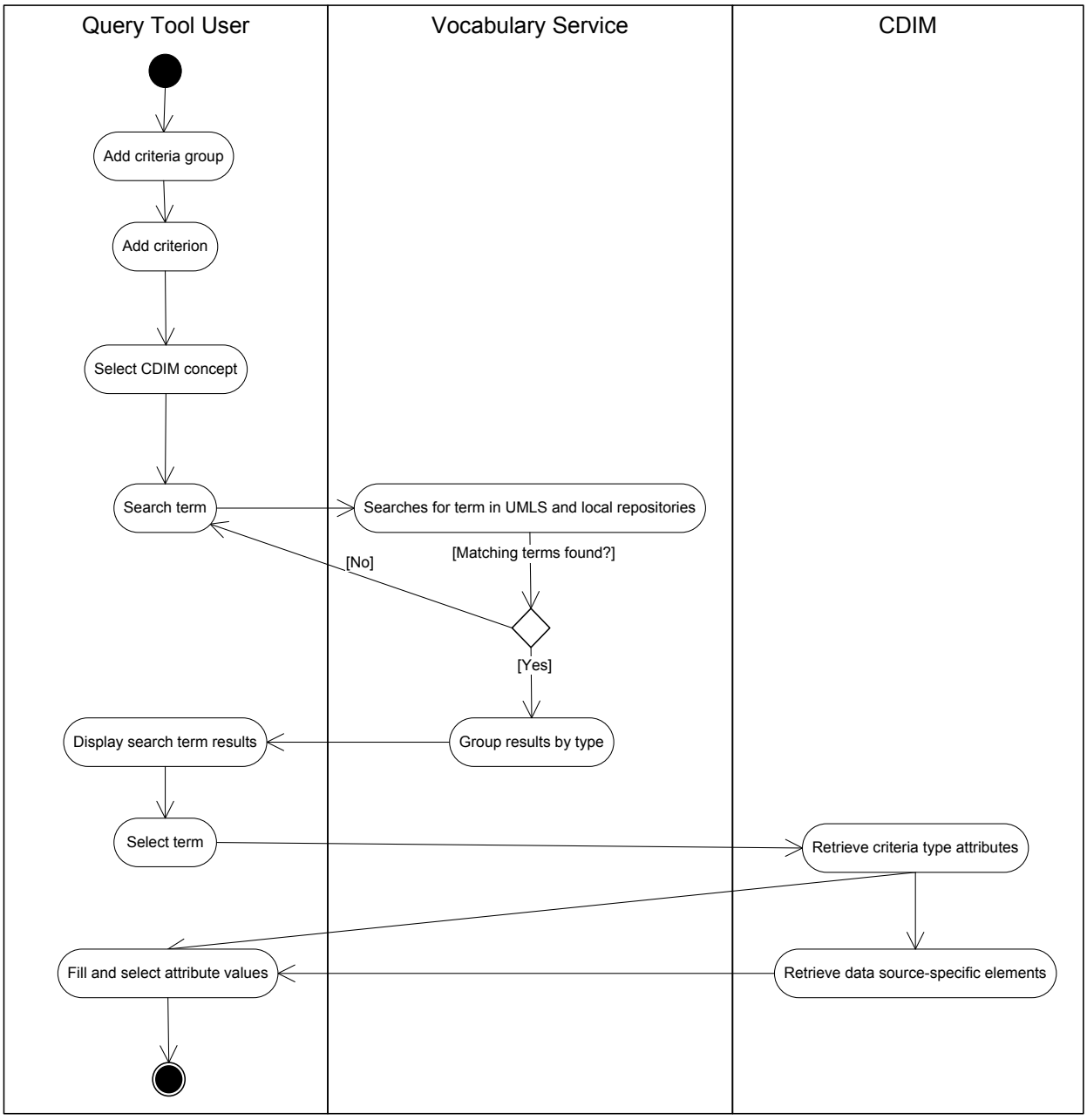

Figure 11: Typical workflow for specifying eligibility criteria group

Running a query

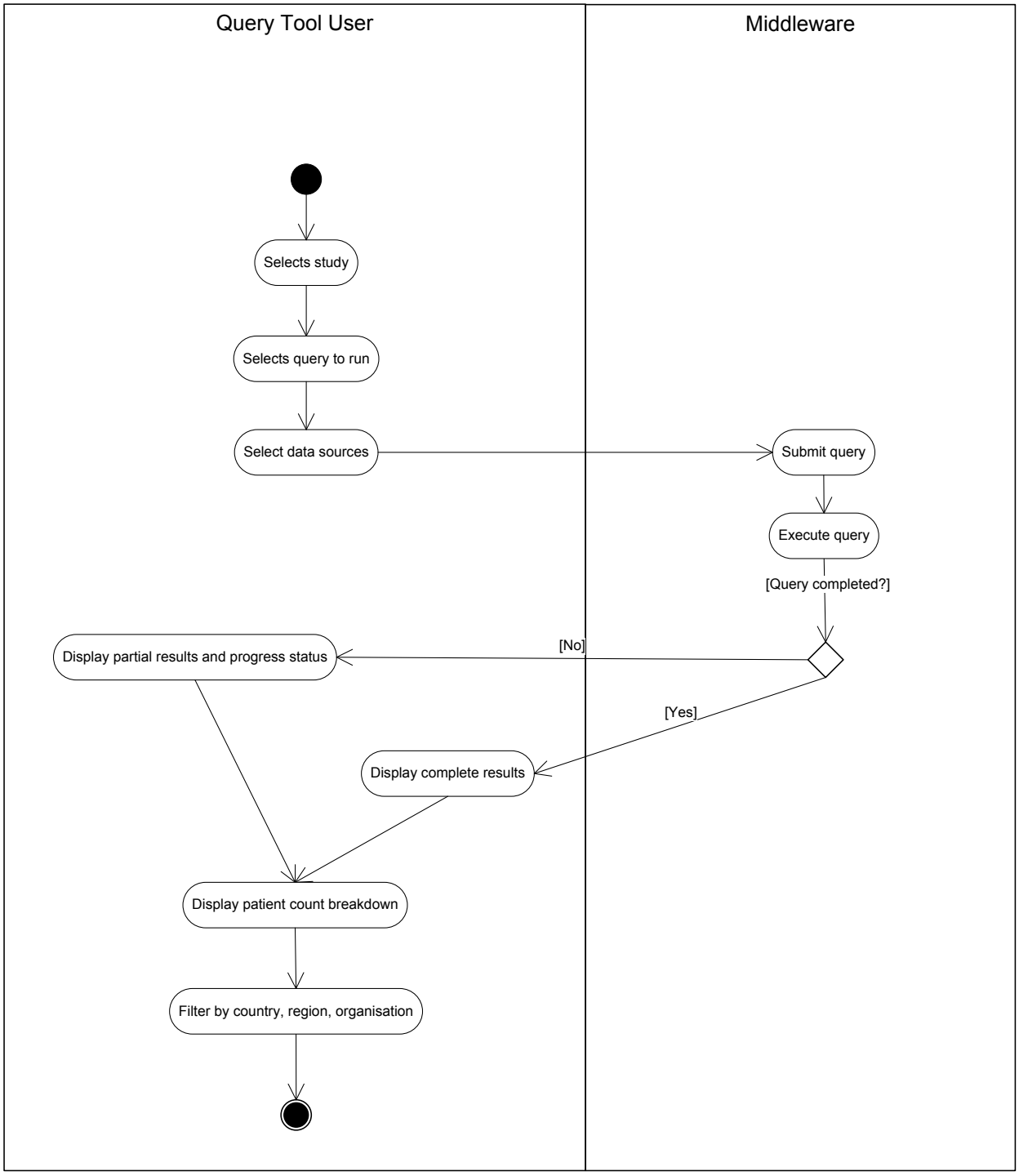

Figure 12: Typical workflow for running a query

4.2. Specific UI Considerations

Based on the desiderata and previously identified requirements, the UI design and workflow, includes the following specific considerations for the functionality of the workbench's interface.

Collaborative study design

Study and query design can be collaboratively worked on by multiple groups of users. While each study can have a number of queries, only one user at a time can modify and run a query. Other users can only view the query and are unable to edit.

Role based access control, as described by the TRANSFoRm security framework, will allow for restricted user access based on individual and group permissions. Complex authorisation rules are captured as part of the participatory design process and are seamlessly integrated with the UI to tailor the user's experience. Restricted UI components are either not shown to unauthorised users or replaced with warnings.

Automatic grouping of term search results

Ideally, a team of clinical subject matter experts would group concepts (UMLS semantic type or coding system) into the main eligibility criteria types to be used by the query formulation workbench. This information would be represented in CDIM and made available to the UI through the TRANSFoRm distributed infrastructure.

As an alternative, users can select a specific criterion type when adding new criteria, modifying UI elements to allow suitable data entry. For instance, a short list of criterion type can include: demographics, diagnoses, procedures, medications, measurements (lab tests and vital signs). When a demographics criterion is selected by the user, options such as age, gender, and ethnicity will be displayed with the appropriate attributes (constraints on time, value, etc). For other criteria types, including where a coding system is to be used, the vocabulary service can be invoked. Development of a generic model of criteria will allow multiple attributes per criterion type, with an expected mapping between each criterion type and its child attributes.

As the CDIM concept needs to be selected by the user, this may be comparable to the criterion type described above.

Units and concept-specific attributes

The TRANFoRm project will address the issue of varying unit types between data sources and the associated unit conversion issue.

For example, when the user selects HbA1c as the term, the list of units the user should see are %, mmol/mol, mmol/L and iu/L, as these are used by the data sources NIVEL and

GPRD. If the user selects % as a unit type, data in other units should be converted to % (if possible). The conversion of units will be handled by the CDIM translation web services, again accessed through the TRANSFoRm distributed infrastructure.

Based on feedback obtained from the wider TRANSFoRm team and domain experts, one solution for the display of units and unit conversion is as follows: the units available per concept are provided by the CDIM-data source translation services. Information on which units can be converted is also provided. This information can be displayed to the user and when a unit is selected, any possible conversions are done via the CDIM-data source mapping, to include as many relevant records as possible.When the user selects a concept using the vocabulary service, this concept (UMLS concept ID or local code) needs to be mapped to the CDIM concept for the subsequent correct display of units of measure per concept.

Results breakdown

The patient count returned for the query is not only as a single total, but is returned as a breakdown to the level of individual criteria and criteria groups as designed by the user in query creation, taking into account the issue of vicarious identification. The results can be filtered by country; region, etc. provided this information is available from the data source.

4.3. Query Formulation Workbench UI Storyboards

Storyboards covering the main functionality provided by the query formulation workbench UI are shown below. Through the storyboarding, we demonstrate the full workflow, interface considerations and requirement implementation alternatives. Full size mock up screens are available in Appendix A.

Medium fidelity mock ups, showing creation of a query in an example web application using HTML and the jQuery JavaScript framework are shown in Appendix B. These mock ups demonstrate the creation of the diabetes eligibility criteria example study, described in the following subsection.

4.3.1. Example Study: Type 2 Diabetes Use Case

Part of the Type 2 Diabetes use case, described in the TRANSFoRm Deliverable D1.1 [2], is used for the UI workflow description (storyboards), describing functionality that relates to a set of typical eligibility criteria creation and running with the TRANSFoRm query formulation workbench.

Based on the specification of the workbench's UI, the diabetes eligibility criteria can be specified as the following groups:

Inclusion Eligibility Criteria Group: Diabetes **AND**
Inclusion Eligibility Criteria Group: Drug Therapy **AND**
Inclusion Eligibility Criteria Group: Laboratory Data **AND**
Inclusion Eligibility Criteria Group: Demographics

A. Inclusion Eligibility Criteria Group: Diabetes

Subgroup: Definite T2D **OR**
Subgroup: Probable T2D **OR**
Subgroup: Possible T2D

A.1. Subgroup: Definite T2D

Type 2 Diabetes

A.2. Subgroup: Probable T2D

Maturity-onset diabetes **OR**
Non-insulin dependent diabetes mellitus **OR**
NIDDM **OR**
Subgroup (Type 1 diabetes **AND** Type 2 Diabetes)

A.3. Subgroup: Possible T2D

Steroid Induced Diabetes OR
Gestational Diabetes OR
Diabetes Mellitus

B. Inclusion Eligibility Criteria Group: Drug Therapy

Subgroup: Insulin Only **OR**
Subgroup: Insulin, OGLA, Metformin **OR**
Subgroup: OGLA, Metformin **OR**
Subgroup: Metformin **OR**
Subgroup: No Drug Therapy

B.1. Subgroup: Insulin Only

Insulin **AND**
Daily frequency < 3

B.2. Subgroup: Insulin, OGLA, Metformin

Subgroup (Insulin **AND** Oral glucose lowering agents) **OR**
Subgroup (Insulin **AND** Oral glucose lowering agents **AND** Metformin)

B.3. Subgroup: OGLA, Metformin

Oral glucose lowering agents **OR**
Subgroup (Oral glucose lowering agents **AND** Metformin)

B.4. Subgroup: Metformin

Metformin

B.5. Subgroup: No Drug Therapy

Assumption: No drug therapy means that patients with diabetes are not on any diabetes drugs (insulin, OGLAs, metformin).

Exclusion group (Insulin **OR** OGLA **OR** Metformin)

C. Inclusion Eligibility Criteria Group: Laboratory Data

Fasting blood glucose >= 7.0 mmol/l **OR**
Random blood glucose >= 11.1 mmol/l **OR**
HbA1c >= 6.5%

C.1. Inclusion Eligibility Criteria Group: Demographics

Age >= 35

4.3.2. Study Creation and Collaborative Work Storyboard

Figure 13 describes user authentication, study selection and review of study collaborators. UI 1 shows a typical prompt for user entered data, in this case a user's authentication details. On entering these details they are confirmed using the distributed infrastructure and security framework before the user is then redirected to a list of available studies as in UI 2. The list of available studies is determined by the role of the user such that only studies where the user is an authorised collaborator are shown.

Selection of a study, using a UI element such as a link or button, redirects the user to more detailed information on the study as in UI 3. As well as general information on the study, a list of protocols previously designed using the query formulation workbench are shown.

The menu options available to the user are also updated, providing the user with the option of showing the study collaborators. On selection of this menu link the user is redirected to a list of study collaborators as in UI 10. Changes can then be made to the list of study collaborators.

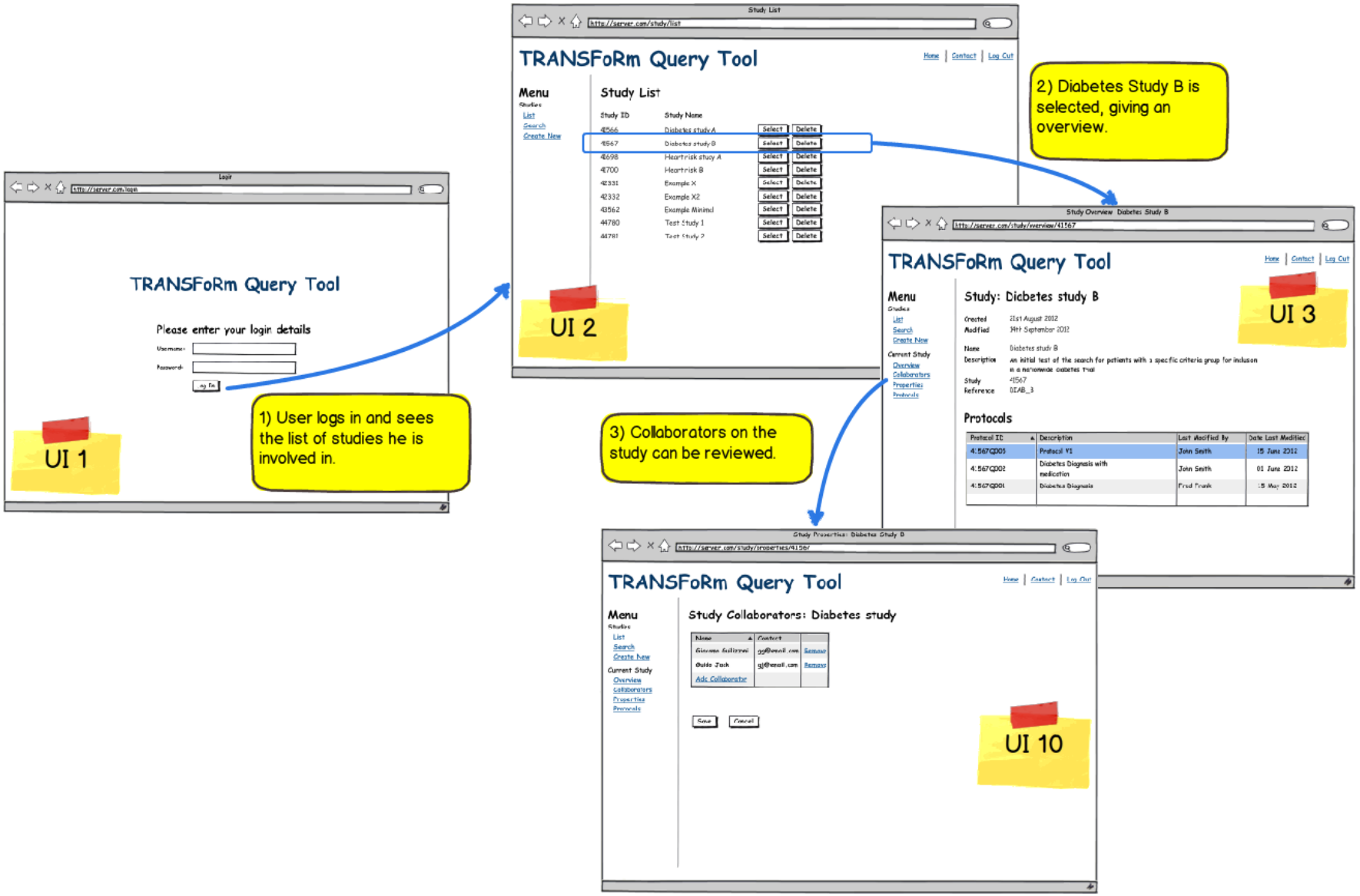

Figure 13: Storyboard of Studies and Collaborative Work

4.3.3. Protocol Building Storyboard

A study can have a number of query designs, or protocols, defined to reflect protocol versioning. Figure 14 shows the workflow for the building of a single protocol for a diabetes study. The TRANSFoRm query formulation workbench enables users to define eligibility criteria by selecting the data sources to be queried (UI 12) and by creating criteria groups, that allow flexible and logical criteria definition with inclusion, exclusion criteria, as well as logical operators. A criteria group represents a logical combination of concepts and their attributes. UI 13 describes a typical dialog prompting the user to enter details on a criteria group. When a user has entered all required data to describe a criteria group they click on a submission element, such as a button, and are returned to the query designer which has now been updated to include the new criteria group.

Figure 15 describes concept selection and concept attribute definition using the example of searching for the CDIM concept “Disease” with the “Diabetes Mellitus” concept. First the user must request addition of a criterion as in UI 14, which will update the UI to show the dialog to enter all required information as in UI 17. UI 17 also shows the integration between the query formulation workbench and the vocabulary service on searching for the concept “Diabetes Mellitus”. The user is offered a list of possible concepts with associated detailed information, including the concept’s representation across multiple source vocabularies.

On selection of a concept from the results list the UI is updated to appear as in UI 18. The user is then able to enter attributes using simple on screen elements such as text boxes and date pickers. These attributes are used to further constrain the eligibility criteria.

4.3.4. Query Execution and Results

Figure 16 describes successful submission of a designed protocol and display of the results on screen. In this example the natural delay due to query translation and submission to individual data sources using the distributed infrastructure is not shown. UI 22 shows a typical dialog providing the user with information on the current stage of the workflow, successful query submission. UI 9 shows the results page, which in this example is automatically redirected to on successful submission. The overall patient count is shown as are individual counts for all elements of the protocol, including individual criteria and groups.

The results can be further filtered by the user as in UI 24, which on confirmation of filter properties, will update the results page to show the new counts as in UI 25. Queries can be executed multiple times, producing multiple results, each of which can be viewed by the user at any point. Figure 17 describes the listing of query results for a single query in UI 8. A user can choose to edit the protocol, as in UI 7 as described above, or simply view the results as in UI 9 and above.

Partial result sets may be provided by the distributed infrastructure, which can be displayed to the user in a similar manner to full result sets. Figure 18 shows an alternative, where only a running count of results is shown to the user for partially completed queries.

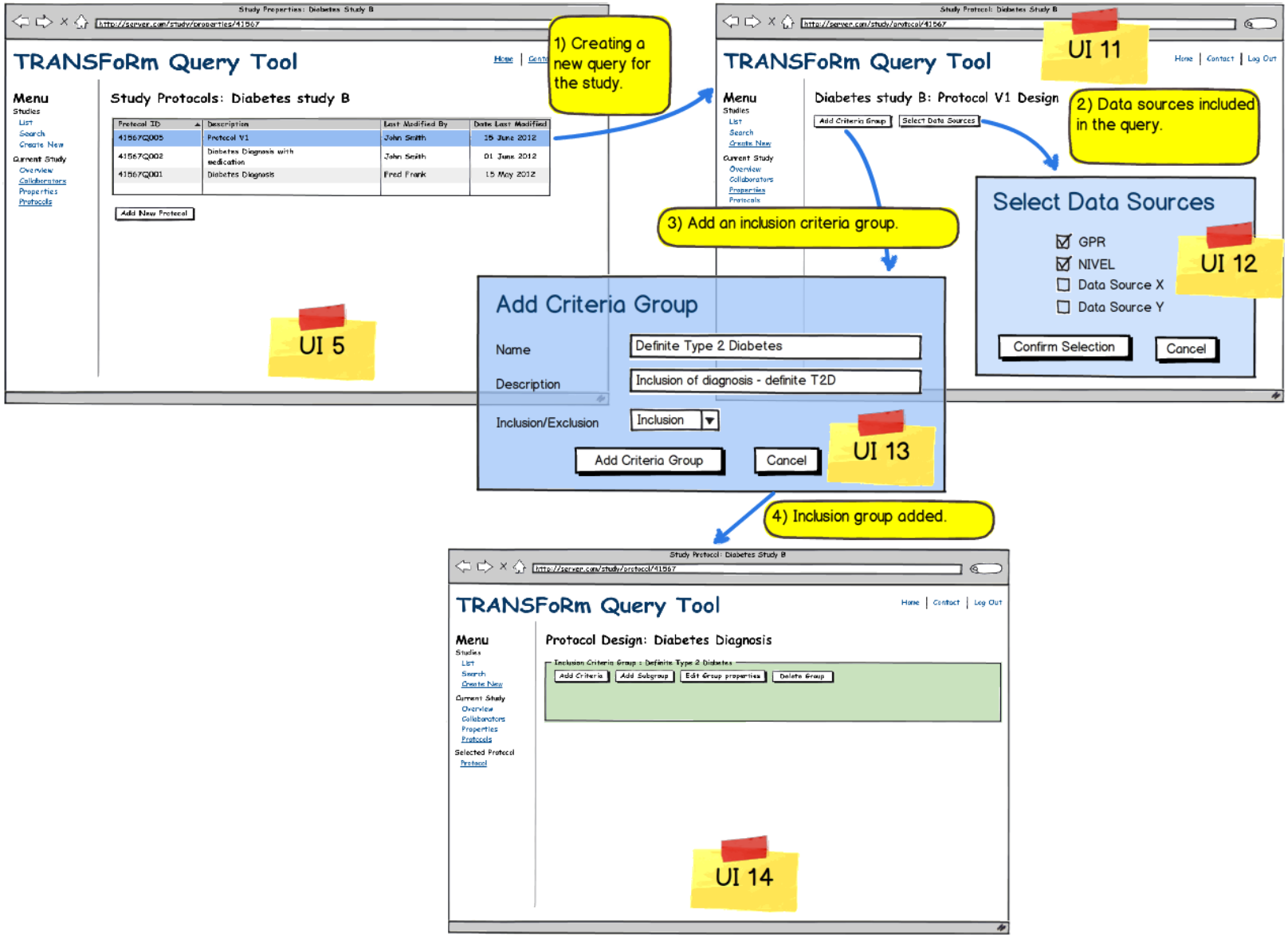

Figure 14: Storyboard for Protocol Creation and Adding Criteria Group

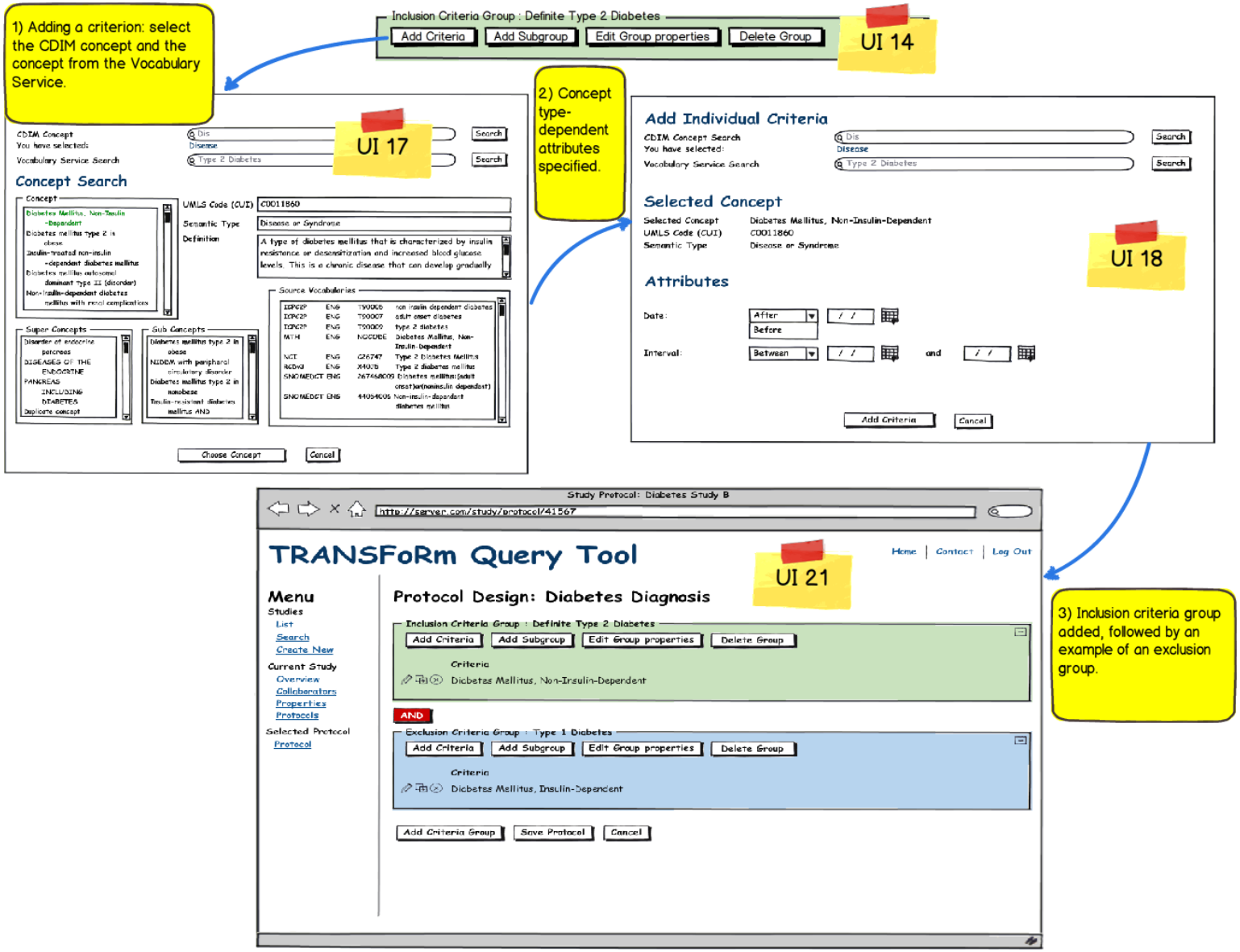

Figure 15: Storyboard for adding concepts defining eligibility criteria.

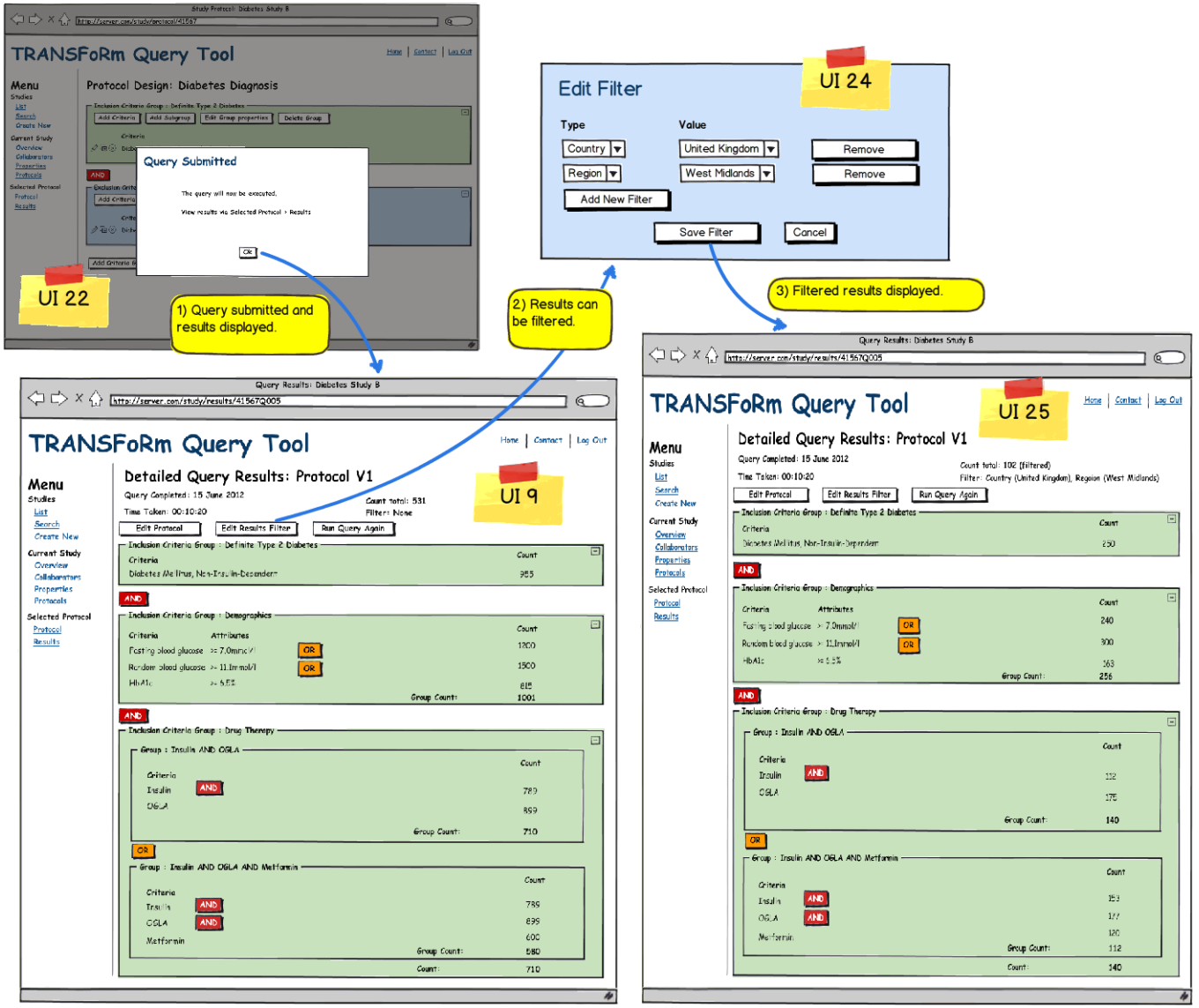

Figure 16: Storyboard for query submission and results display.

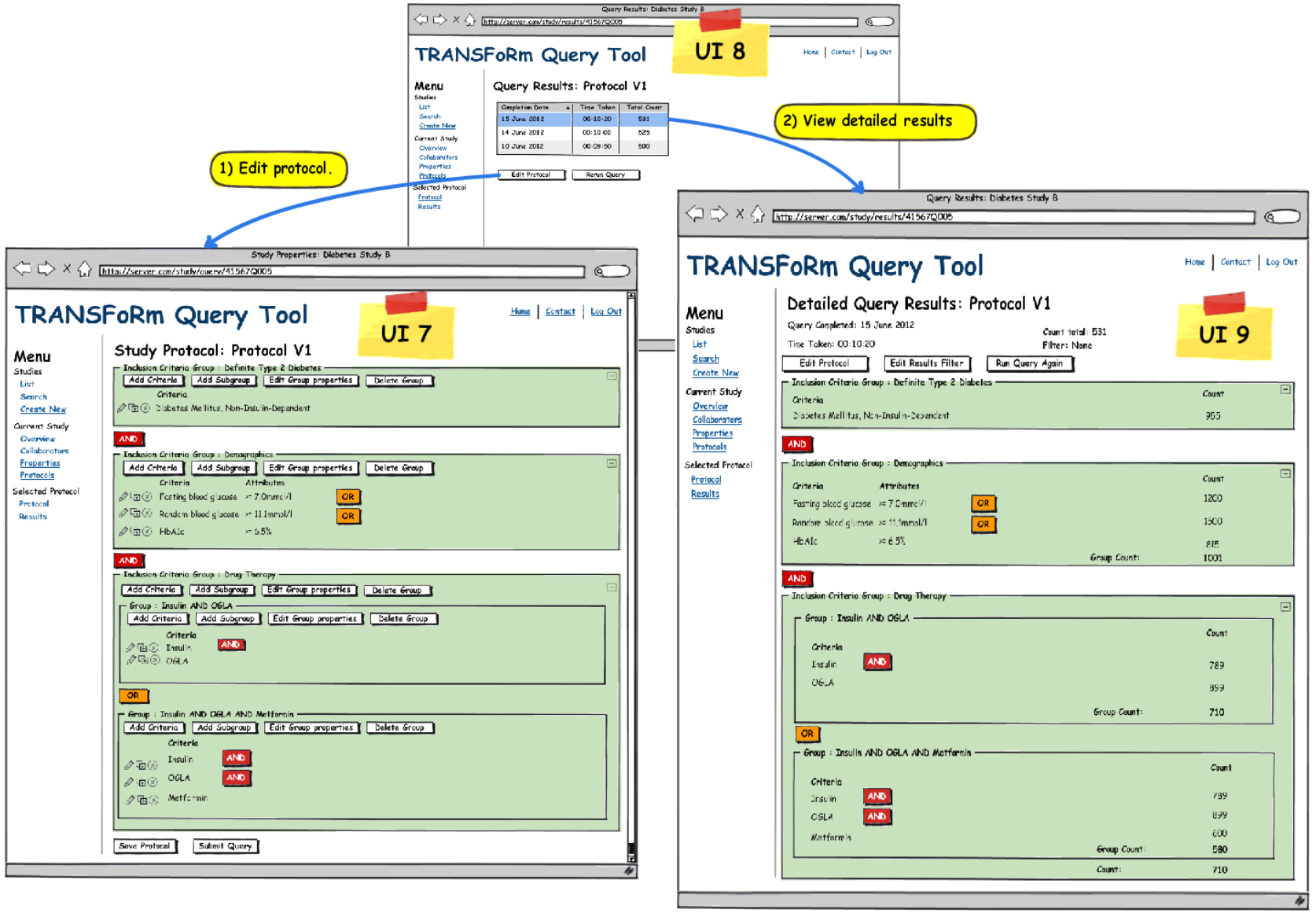

Figure 17: Storyboard of query results and protocol update.

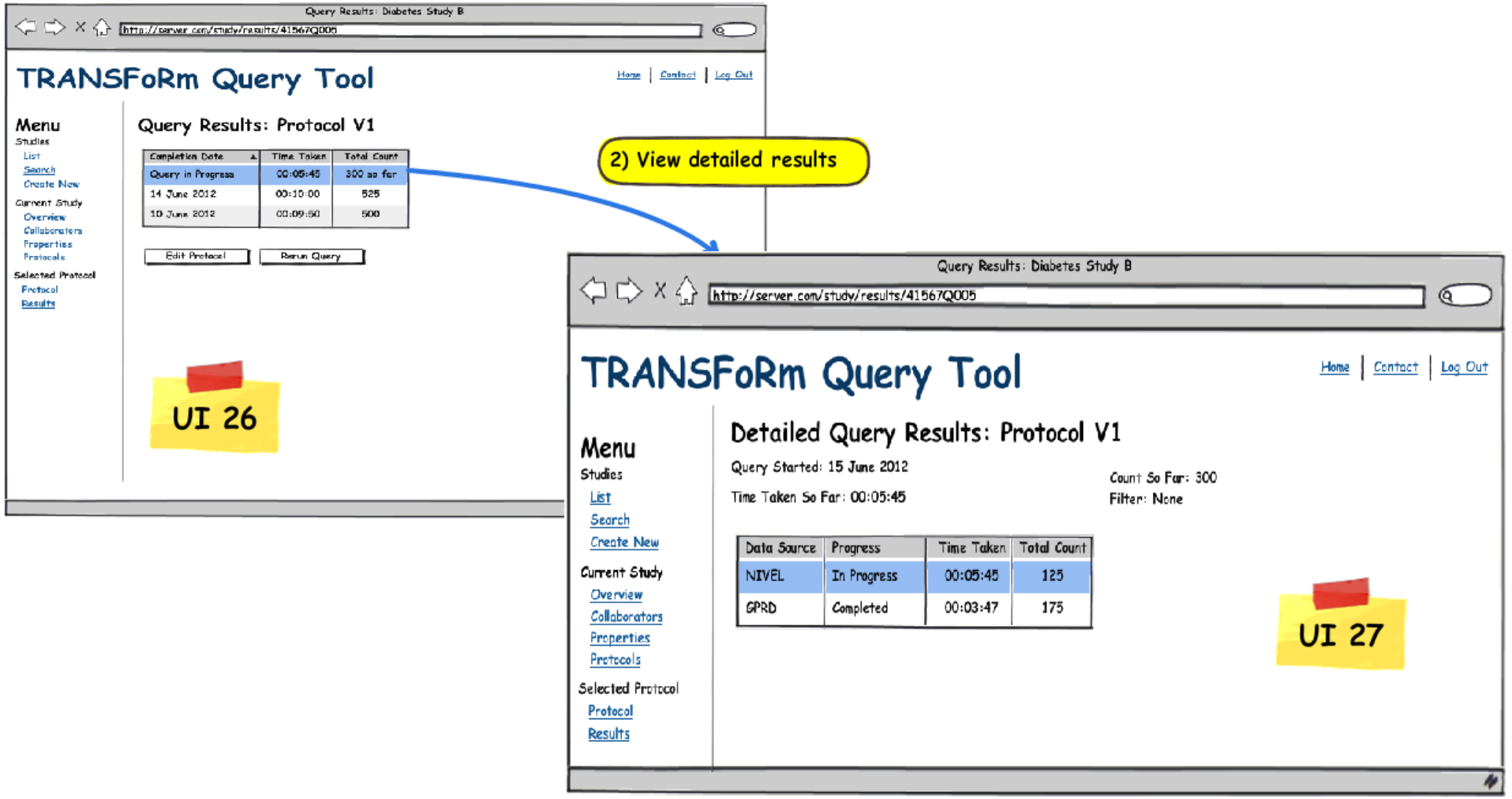

Figure18: Storyboard of viewing incomplete query results

4.4. Summary

This section has discussed the query formulation workbench UI design and workflow, demonstrating the overall functionality the software tool. The storyboards used have demonstrated the realisation of the workbench's functional requirements and associated UIs, as derived by a participatory design process, discussed in section 2. Translation of the diabetes eligibility use case to groupings of eligibility criteria has allowed for development of storyboards of interfaces based on a simple visual representation. Groupings of eligibility criteria can be easily managed by users and complex procedures such as adding concepts can be presented as a series of steps, guiding the user through the process. The following section discusses the technical implementation of the Query Formulation Workbench software tool and its associated user interface.

5. Implementation of the Web-based Query Formulation Workbench Software Tool.

This section presents technical implementation details of the web-based query formulation workbench software tool, together with associated and supporting technology components from other parts of the TRANSFoRm digital infrastructure.

5.1. Functionality Overview

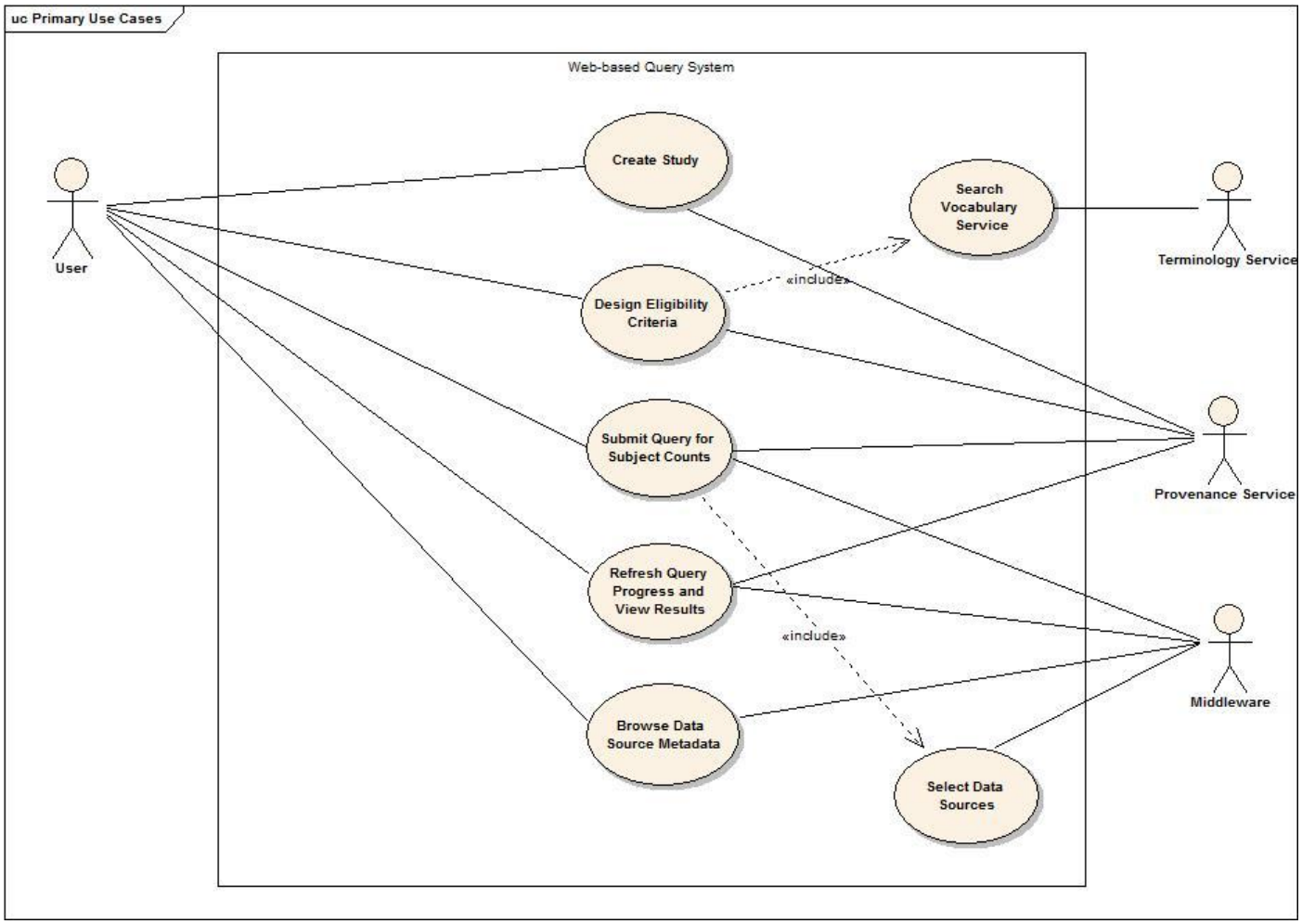

Figure19: Functional description

Based on the functional requirements and UI design specifications, a UML use case diagram is provided in this section to illustrate the functionality that the current version of the TRANSFoRm query formulation workbench deliver (Figure 19): users can browse data source information, design eligibility criteria, select data source and count eligible subjects. Users can browse EHR data sources registered with TRANSFoRm and find basic information about each data source, such as its country, region, organisation, size and used clinical coding schemes. A basic knowledge of available data sources helps users to decide which data sources are relevant for each study and should be selected during query creation.

Design of specific eligibility criteria and analysis of the returned results allow users to find counts of potential subjects who are eligible to participate in their studies. More specifically, users can create studies,

search for concepts using the vocabulary service, design eligibility criteria, select data sources and run queries. Users can also actively monitor query progress for long-running queries and view up-to-date results on demand if this functionality is made available through the TRANSFoRm distributed infrastructure.

The query formulation workbench needs to interact with various other TRANSFoRm components and services to fulfil most of the required functionality. The workbench's UI includes a vocabulary search interface that allows the user to seamlessly connect to the TRANSFoRm integrated vocabulary service, in order to search for clinical concepts and associated codes. The provenance service is integrated at each stage in the workflow of finding eligible subjects: create study, design eligibility criteria, submit query and retrieve query results. The query formulation workbench also uses various provenance templates provided by provenance service to capture each activity's provenance information and stores the data in a central provenance database. Users can instruct the query formulation workbench to search across a list of selected EHR data sources, represented in the UI as a list with appropriate additional information. In the background, the query formulation workbench invokes the TRANSFoRm distributed infrastructure to serve requests. Finally updating the UI information is provided where appropriate.

Designing and running clinical studies is intrinsically a collaborative process, which involves many people to work together. Furthermore, developing research protocol and eligibility criteria is by nature an iterative process with elements being updated constantly until a final version is reached and approved. The TRANSFoRm query formulation workbench implements proper versioning and concurrent access control mechanisms and thus provides a team-based collaboration environment.

5.2. Application Architecture and Implementation

The TRANSFoRm query formulation workbench UI is implemented as an Ajax-enabled web application, using JQuery technology, which delivers a responsive and rich UI and is easily accessible across diverse client platforms. This section describes the application architecture and how the application interacts with other TRANSFoRm components, to fulfil the previously defined functional and user requirements (section 2); particularly designing eligibility criteria, running queries and retrieving results.

5.2.1. Application Architecture

Following the typical layered web application design, the TRANSFoRm web-based query formulation workbench comprises of three layers: a presentation layer, service layer and persistence layer (Figure 20).

The presentation layer provides an Ajax web client implemented using JQuery. The Ajax application is delivered to clients'web browsers and supported by the Spring MVC framework. The Ajax client includes a vocabulary search interface for users to quickly search the TRANSFoRm integrated vocabulary service. In order to reduce network latency and improve throughput, the Ajax client is designed to connect directly to the vocabulary service through its RESTful interface.

Processing logic and interaction with other TRANSFoRm components/services are implemented in the service layer. In response to requests from the Ajax client, the Spring MVC framework invokes the service layer components and returns their responses to the user.

The persistence layer is implemented based on Object-Relational Mapping (ORM) technology to build domain objects. These objects are mapped to relational database tables through Hibernate, a popular open source ORM solution which implements Java Persistence API (JPA).

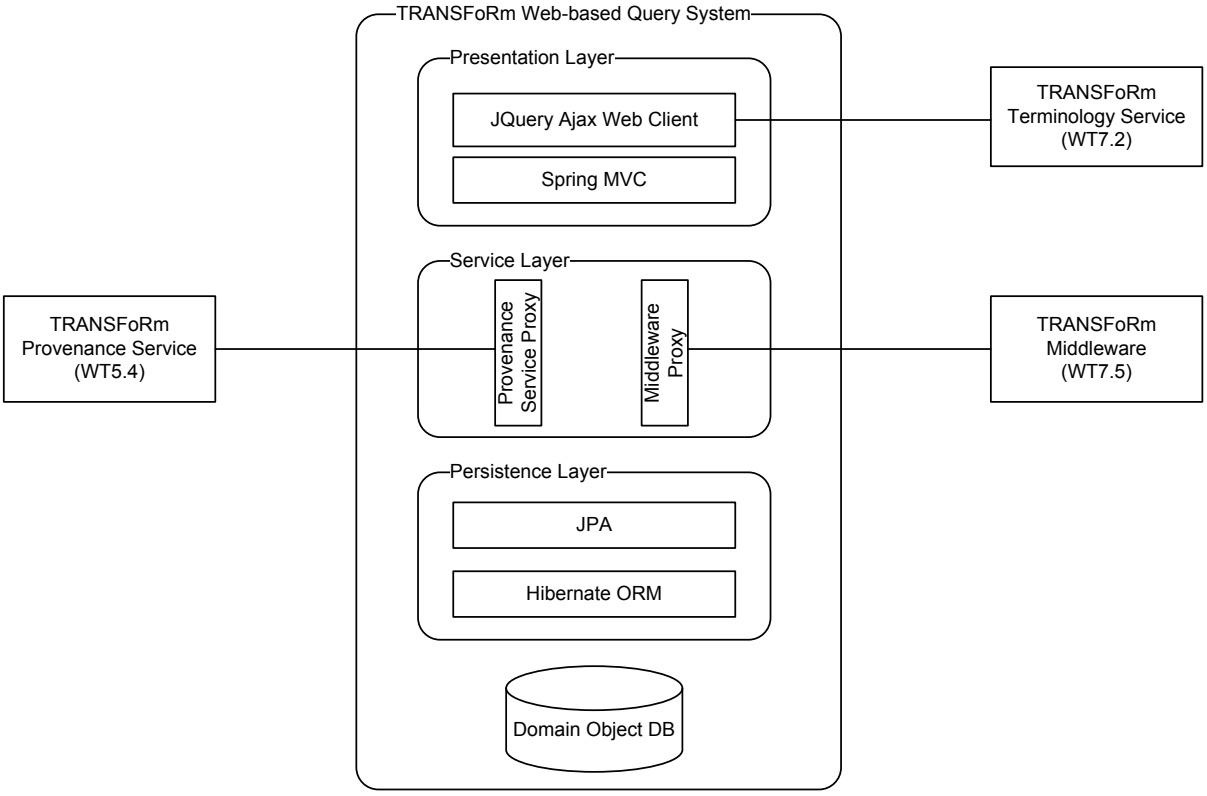

Figure 20: Layered Application Architecture

5.2.2. Technology Stack

The current query formulation workbench is built using the following set of technologies:

- CentOS 6.2
- Oracle JDK 6u32
- Apache Tomcat 6.0.35
- JQuery Framework 1.7.2
- Spring Framework 3.1.1
- Hibernate 4.1.4
- MySQL 5.5.25

The Ajax component of the TRANSFoRm query formulation workbench is implemented with JQuery framework 1.7. JQuery is a fast and concise JavaScript library that simplifies JavaScript programming and Ajax interactions for rapid web development.

The server side component of the application is based on Spring framework 3 and Hibernate 4. Spring is an open source application framework and is amongst the most popular for enterprise level Java software development. MySQL 5.5, the most popular open source relational database, is used as the database server. The whole application is bundled as a Java servlet application and hosted by Apache Tomcat 6. The application is developed with the latest release of JDK 6 from Oracle.

Although the application is based on Java and is able to run on many platforms, the target operating platform is CentOS 6.2. CentOS is a free enterprise-class Linux distribution derived from the Red Hat Enterprise Linux (REHL) distribution.

5.2.3. Implementation Workflow – Designing Eligibility Criteria

Figure 21 describes the implementation level workflow of designing eligibility criteria and shows the interaction between the query formulation workbench and other TRANSFoRm services, such as the vocabulary and provenance services.

Clinical research users start the process of designing eligibility criteria by opening a study and creating an empty query. The user adds criteria groups and subgroups based on their desired Boolean logic structure. The user can decide between AND or OR as the Boolean operator between groups as well as between criterions within each group.

When adding specific criteria to a group, the user is prompted with a list of semantic type choices driven by the CDIM information service, such as diagnosis, laboratory measurement, prescription, age, gender etc. This list of concepts can be either encoded in the query formulation workbench internal database or retrieved from the CDIM information service using the TRANSFoRm distributed infrastructure. In either case, the UI is developed such that any change to a user's selection updates the form used for data capture.

Where relevant, the user can use the vocabulary service to search for a suitable clinical concept. The workbench invokes the vocabulary web service to do the search and updates the UI as soon as results are returned. The vocabulary service provides a web service API for either term or code based search. A term based search matches the concept description with user input phrases. The TRANSFoRm vocabulary service has multi-lingual support, so search terms can be in any language that the vocabulary contents support. Alternatively, if familiar with a specific coding system, the user can enter a known code and find the concept through a code based search.

Apart from clinical codes, for some criteria (such as a laboratory measurement or prescription) the user may input value or unit constraints in order to build a complete rule. Users may also add temporal restrictions to the criterion, for example, recordings before or after a certain date.

During the eligibility criteria design process, users can periodically save their progress. On a user's request to save the current eligibility criteria design, the query formulation workbench persists to any changes to the local database and invokes the TRANSFoRm provenance service to persist the saved criteria as provenance data through the provenance templates for eligibility criteria design. User's behaviour and design activities are captured so design patterns can be analysed later.

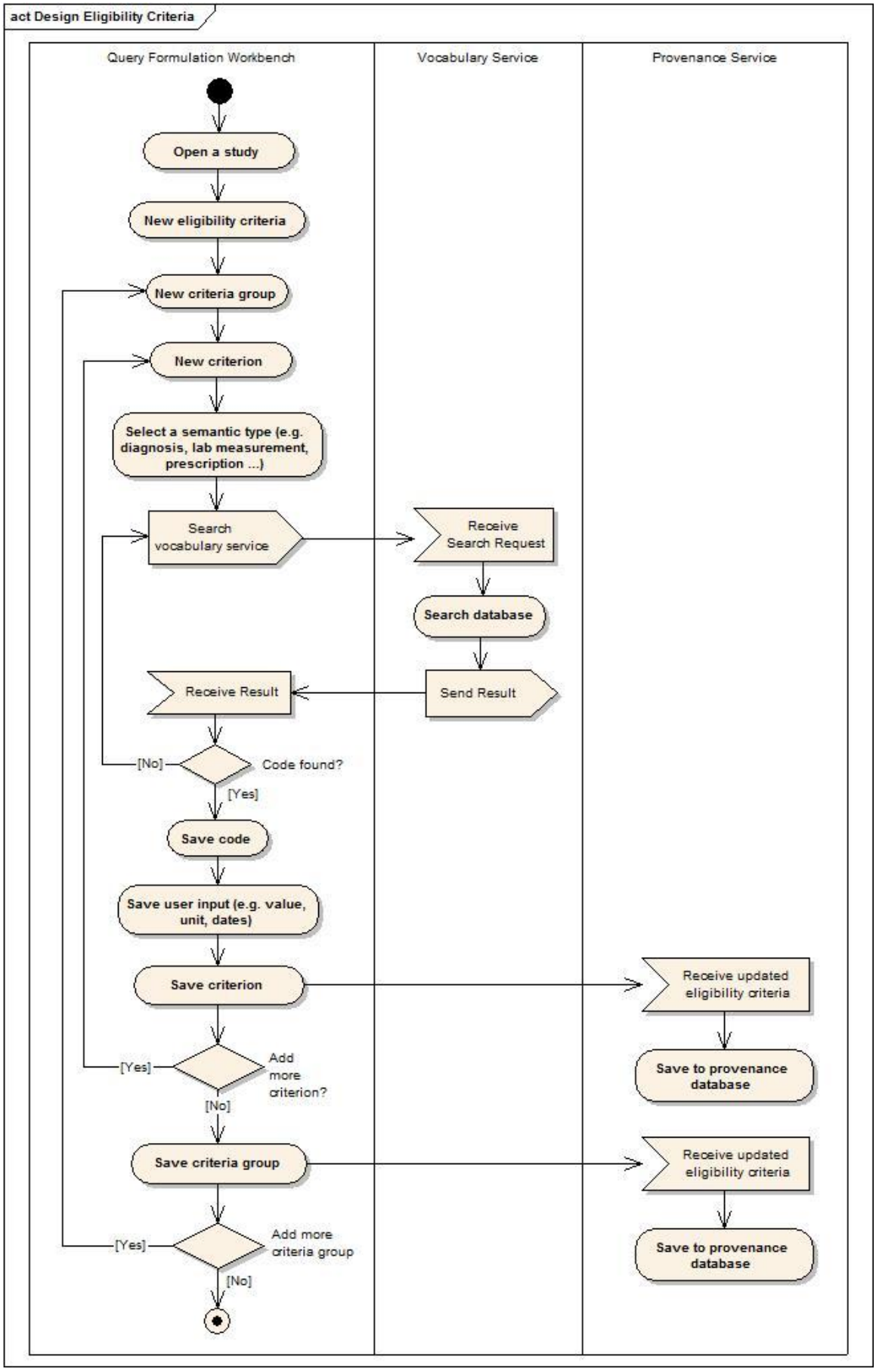

Figure 21: Design Eligibility Criteria Workflow

5.2.4. Workflow – Running a Query and Reporting Results

Figure 22 describes the typical workflow involved when running a query and shows use of the query formulation workbench, the distributed infrastructure for data extraction and linkage, and the provenance service. The workflow comprises of query submission, checking query progress and retrieving results.

In order to run a query and return results, firstly a user created eligibility criteria design is selected for submission. The workbench invokes the data source registry service via the distributed infrastructure to retrieve the list of available data sources. The TRANSFoRm data extraction and linkage infrastructure provides a semantically rich registry service, where all data sources are registered. The service provides various metadata for each data source, such as coding schemes used by the data source, the country and geographical regions that the data source covers, the covered population size and how up-to-date the data source is. The provided extra information on each data source helps users to make an informed selection.

Once the user has selected a set of data sources to search, the workbench sends the eligibility criteria and the list of selected data sources to the distributed infrastructure by invoking the distributed infrastructure API which encrypts the query request using the security library. In order to retrieve the query, data sources regularly poll the infrastructure and use the infrastructure API and security library to retrieve and decrypt the request. With the help of CDIM-DSM, the eligibility criteria are translated into query statements suitable for execution on local data sources. The query is run and any results are encrypted and pushed back to an approved study data repository via the distributed infrastructure.

When the workbench passes the query request to the infrastructure, it informs the provenance service that a query request has been made and a query execution process has been initiated so the relevant provenance information is saved. The distributed infrastructure also interacts with the provenance service to record provenance of each activity taking place inside the infrastructure.

Users can request query progress reports once the query has been submitted. The workbench invokes the distributed infrastructure API to retrieve status updates based on the response of individual data sources. The infrastructure retrieves the query status on each data source and reports whether the query task is being processed, completed, or failed. If the query task has failed, the infrastructure returns an error code and suitable message to the query formulation workbench to indicate to the user the exact reason for the failure.

The distributed infrastructure also reports other information, such as the start time and end time of each query task. If any new result is available, the workbench retrieves the result from the data storage using the infrastructure API and decrypts the result. The workbench then aggregates the new result with existing results and presents this to the user. Once all results have been received, the workbench reports the aggregated result to provenance service and indicates the query execution process has completed. All the communication between the workbench and provenance service are based on the provenance templates as defined by the provenance service for the query execution process.

At any point of time after the query is submitted, the user may decide to stop the query, typically because some data sources have taken too long to process the query or the results are deemed unreliable. Whatever the reason, when the user instructs the workbench to stop the query, the workbench instructs the distributed infrastructure using the API to terminate any pending queries. The workbench marks unfinished query tasks as being terminated by the user and again informs the provenance service in order to capture this information.

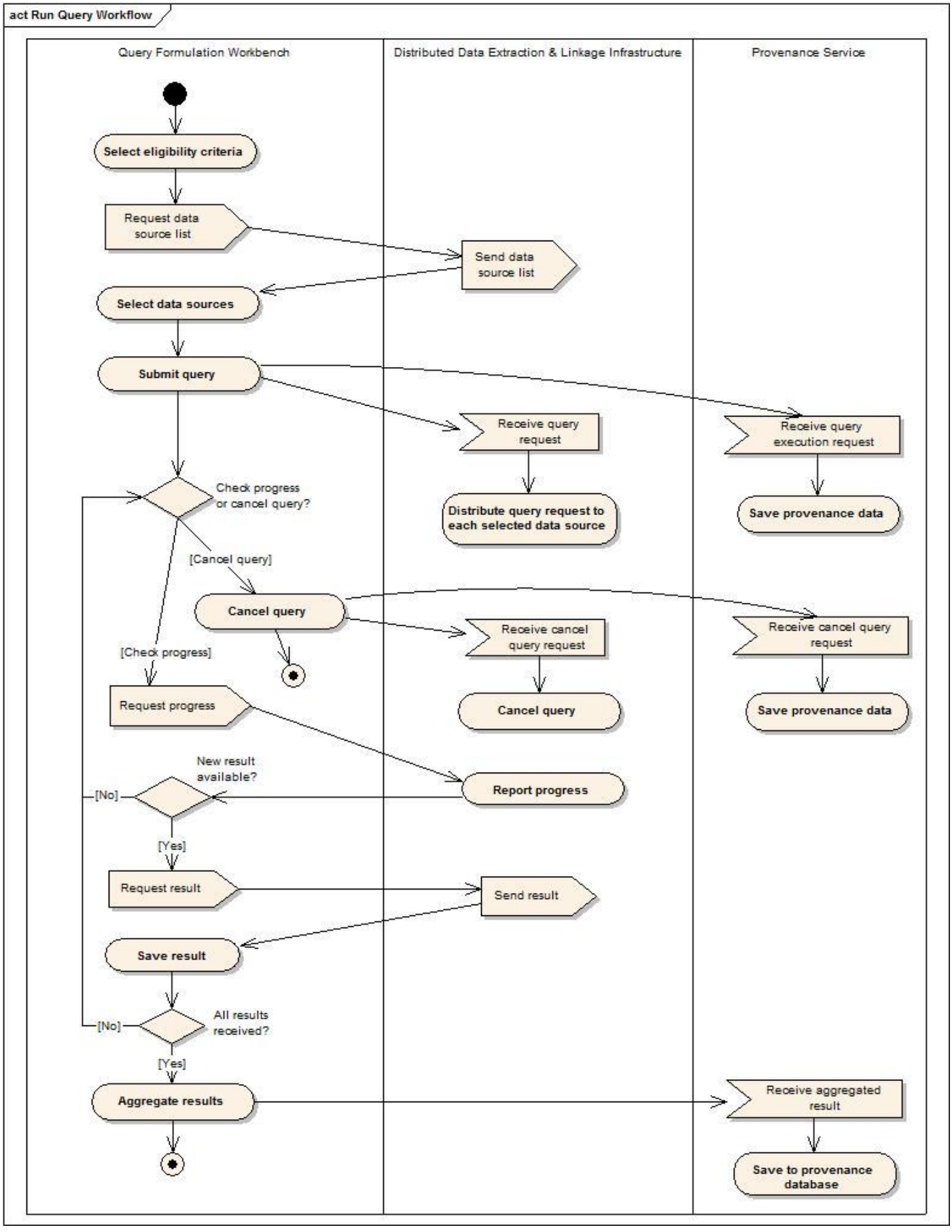

Figure 22: Workflow for Running a Query and Reporting Results

5.3. Concurrency Considerations

The TRANSFoRm query formulation workbench is designed as a team-based collaboration environment, allowing multiple users to work concurrently on the same study. In such a multi-user environment, proper concurrency control mechanisms need to be in place to prevent potential conflicts, when more than one user attempt to update the same object or invoke the same function. The core business objects within a study, which have shared access, currently include protocol, eligibility criteria and query. Simple policies are introduced to protect concurrent access to these objects:

- Only one user can update the protocol and eligibility criteria at any point in time;
- Only one user can initiate a new query for a specific eligibility criteria at any point in time, but new queries can be submitted for another eligibility criteria within the same study as long as that eligibility criteria has no running query;
- Read access has no restrictions.

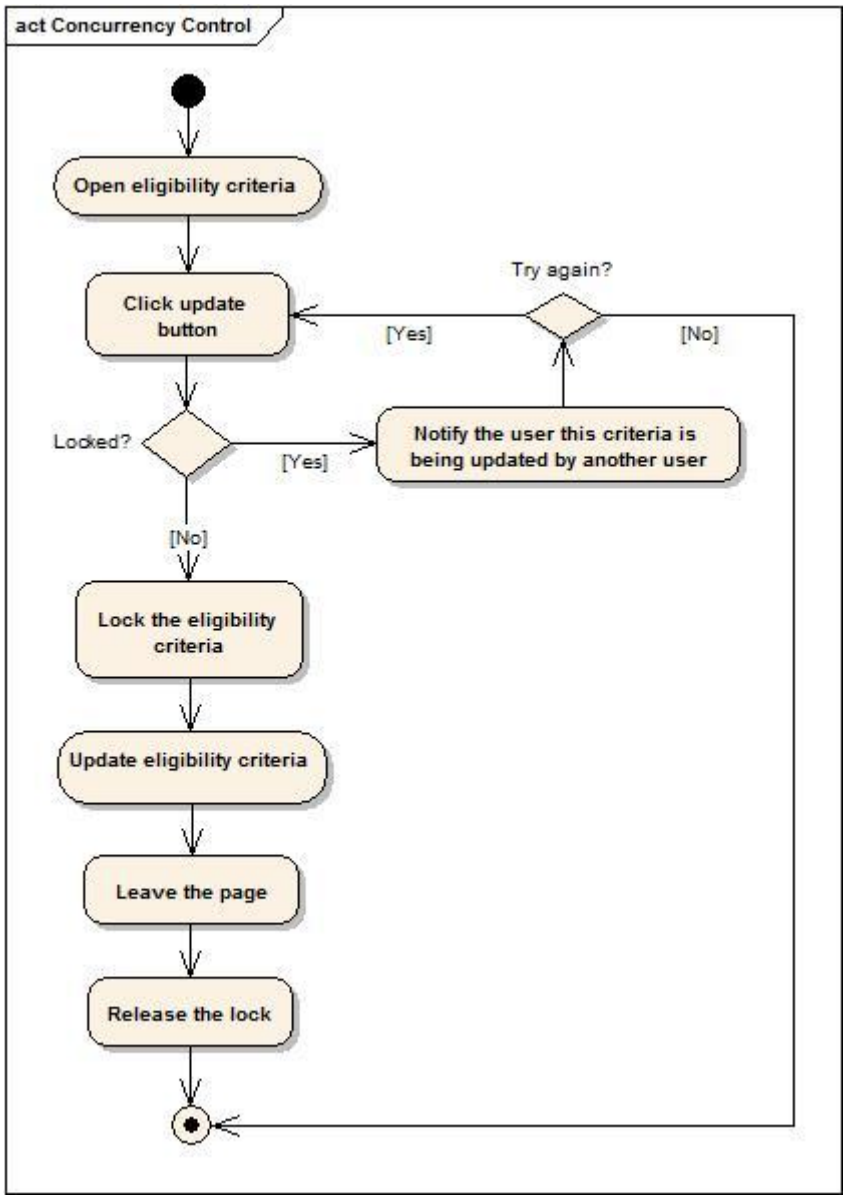

Figure 23: Concurrency Control for Eligibility Criteria Update

As an example, Figure 23 shows the scenario of updating eligibility criteria and implements concurrency control using a locking mechanism. Note that because concurrent reads have no restrictions, users may view stale data. Since the number of users attempting concurrent access is expected to be small, a simple write lock is sufficient in this situation.

5.4. Object Model of Query Formulation Workbench

5.4.1. CRIM Model

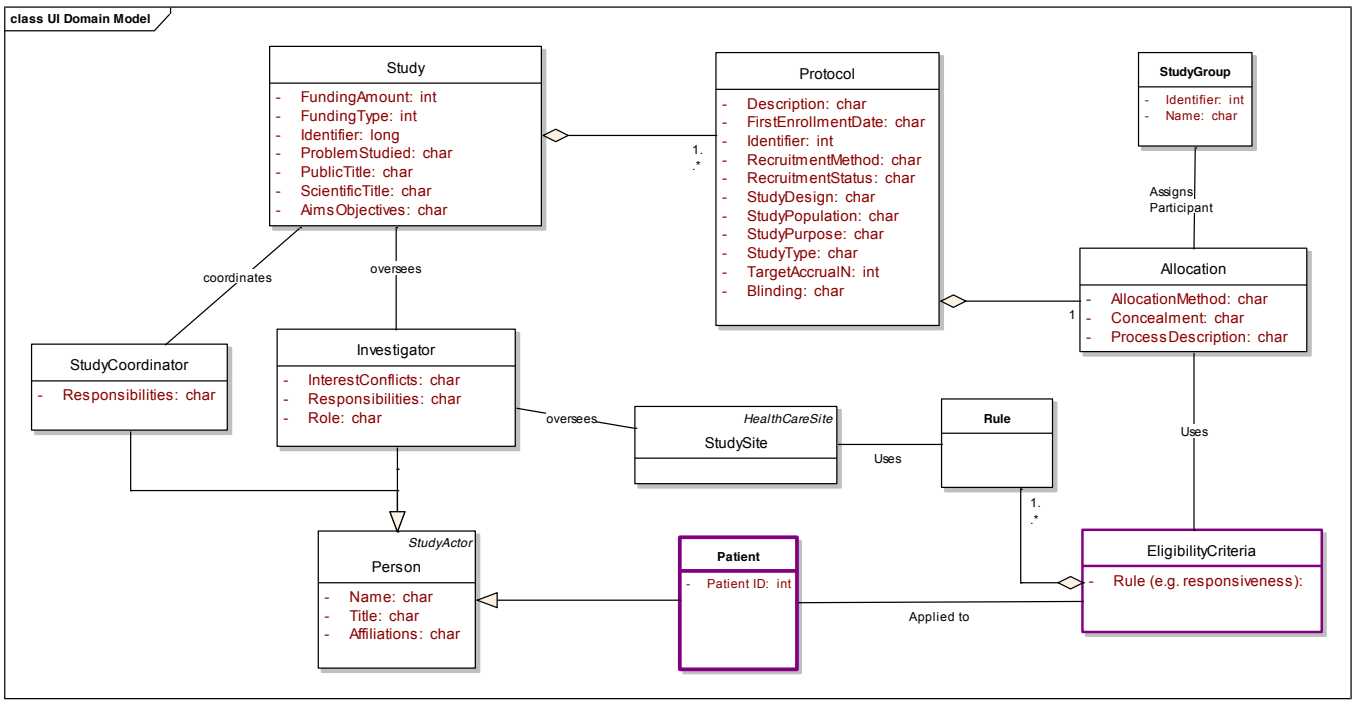

Figure 24: Relevant Part of CRIM Model ([29], Figure 27)

The object-oriented (OO) implementation of the query formulation workbench is based on CRIM, which defines business workflows for TRANSFoRm use cases and develops high-level domain information models. CRIM identifies conceptual entities and relationships between the entities in the clinical research domain. Figure 24 shows a subset of the CRIM domain model, which is relevant for the current system implementation. This subgraph is extracted from the complete UML class diagram [29]. The model identifies conceptual domain entities such as Study, Protocol, EligibilityCriteria, StudySite, Investigator, StudyCoordinator, etc. These entities represent important domain objects and need to be used as the basis to drive software implementation.

5.4.2. Conceptual Object Model of Query Formulation Workbench

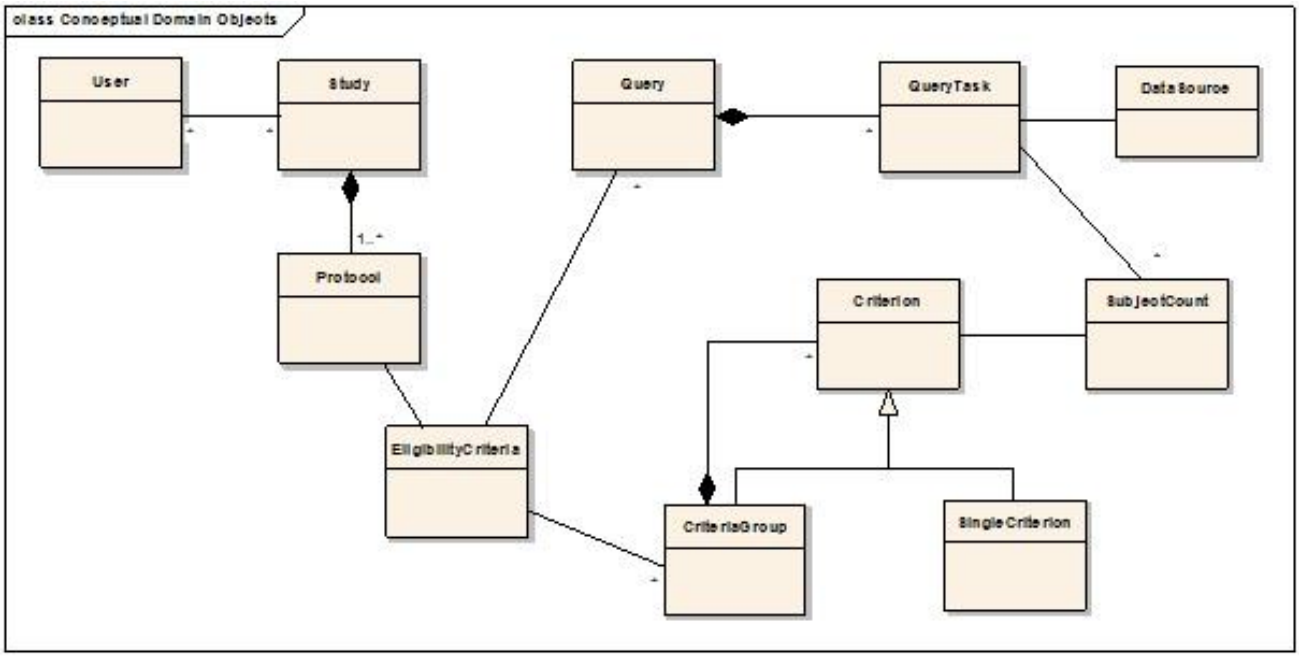

Figure 25: Conceptual Object Model

While CRIM has captured the most important domain entities, a few other entities are missing from the model which are also essential to be represented from a practical implementation perspective, particularly the federated query, counts, etc. A high-level model design for UI system implementation bridges the gap by capturing the missing elements (Figure 25). Each of the elements is elaborated below:

- `User`
  A user can potentially be involved in several studies and is not restricted to one study at a time. On the other hand, since the query formulation workbench is designed as a team collaboration environment, it allows multiple users to work together on the same study. So `User` has many-to-many association with `Study`. `User` may have different roles and so different privileges and responsibilities in different studies. Since `User` is closely linked with authentication and authorisation, it is generally managed by the TRANSFoRm security framework which provides interfaces for other components to retrieve user information.

- `Study`
  `Study` represents a clinical study. It is the central object that encompasses all the components about a study, such as protocol, staff, sites, eCRFs, data results, etc. `Study` is uniquely identified across the whole TRANSFoRm system.

- `Protocol`
  `Protocol` is a realisation of a study's design elements, processes, activities and details. As `Protocol` is modelled in CRIM as many-to-one association with `Study`, one study may have more than one protocol. Each protocol may either represent a different version, or represent different design of the study, which gives users flexibility in managing their study protocol development.

- `EligibilityCriteria`
  `EligibilityCriteria` is a defined list of rules that specifies the eligibility of study participants, specifying patients that are eligible or not to participate in a study. It can include separately inclusion and exclusion rules. `Protocol` has one-to-one association with `EligibilityCriteria`. As

described in section 4, the query formulation UI is designed in such a way that the whole eligibility criteria comprise a number of criteria groups. Each criteria group is either an inclusion criteria group or an exclusion criteria group, so the user has the flexibility to organise inclusion and exclusion rules. Each criteria group can have subgroups so as to form a hierarchical structure. At the lowest level, each group/subgroup comprises a number of single criteria. Each single criterion can either specify a clinical concept such as diagnosis, medication, laboratory measurement etc, or define a demographic rule such as age, gender etc. Criteria groups/subgroups and single criteria are connected using the Boolean operators `AND` and `OR`. The hierarchical structure of criteria groups gives the user the flexibility to organise complex Boolean logic statements, in a nested form. Criteria group and single criterion are modelled as `CriteriaGroup` and `SingleCriterion` respectively in a composite design pattern.

- `Query`
  `Query` represents the federated query and comprises a number of single query tasks, represented by `QueryTask`, one per data source. `Query` records overall query status and aggregates individual query results (i.e. counts in the current deliverable). The user may search different data sources using the same eligibility criteria or do the same search again from time to time. In order to support this flexibility, `EligibilityCriteria` has one-to-many associations with `Query` so multiple query instances can be based on the same eligibility criteria. `Query` has one-to-many relationship with `QueryTask`.

- `QueryTask`
  `QueryTask` represents the actual query task on a single data source, so it has a one-to-one association with `DataSource`. `QueryTask` reports query status for the single query and records the query result such as count returned by the data source. `QueryTask` has one-to-many association with `SubjectCount`, which represents a breakdown of the count result by criterion.

- `SubjectCount`
  A breakdown of the total count by criteria group and individual criterion is a useful feature to help users evaluate the contribution of each criterion or criteria group and fine tune the criteria for a better match. `SubjectCount` is introduced to support this requirement, where each `Criterion` (the superclass of `CriteriaGroup` and `SingleCriterion`) has one-to-one association with `SubjectCount`.

- `DataSource`
  `DataSource` is the abstraction of a real data controller, who hosts and provides access to EHR data. Various metadata about the data source, such as organisation information, data quality, technical configuration, etc., are recorded with a registry service. The service is managed by the TRANSFoRm distributed infrastructure, which provides an API for other components to retrieve and populate data source information.

5.4.3. CDIM Artefact-based Eligibility Criteria Implementation Model

The eligibility criteria model is the core component of the domain object model. As described before, `EligibilityCriteria` comprises `CriteriaGroup` and `SingleCriterion`, which are designed using a composite pattern. Each `CriteriaGroup` can be either an inclusion criteria group or an exclusion criteria group, which is indicated by the `InclusionOperator`. `CriteriaGroup` can have subgroups which are connected using `BooleanOperator` AND/OR. The recursive nature of the subgroup relationship makes it possible to build a nested hierarchy, the leaves of which are `SingleCriterion`. The

specification of individual criteria is based on the concept of CDIM artefacts (Figure 26). A CDIM artefact is an archetype that groups together closely related CDIM ontology concepts in order to construct application friendly reusable units.

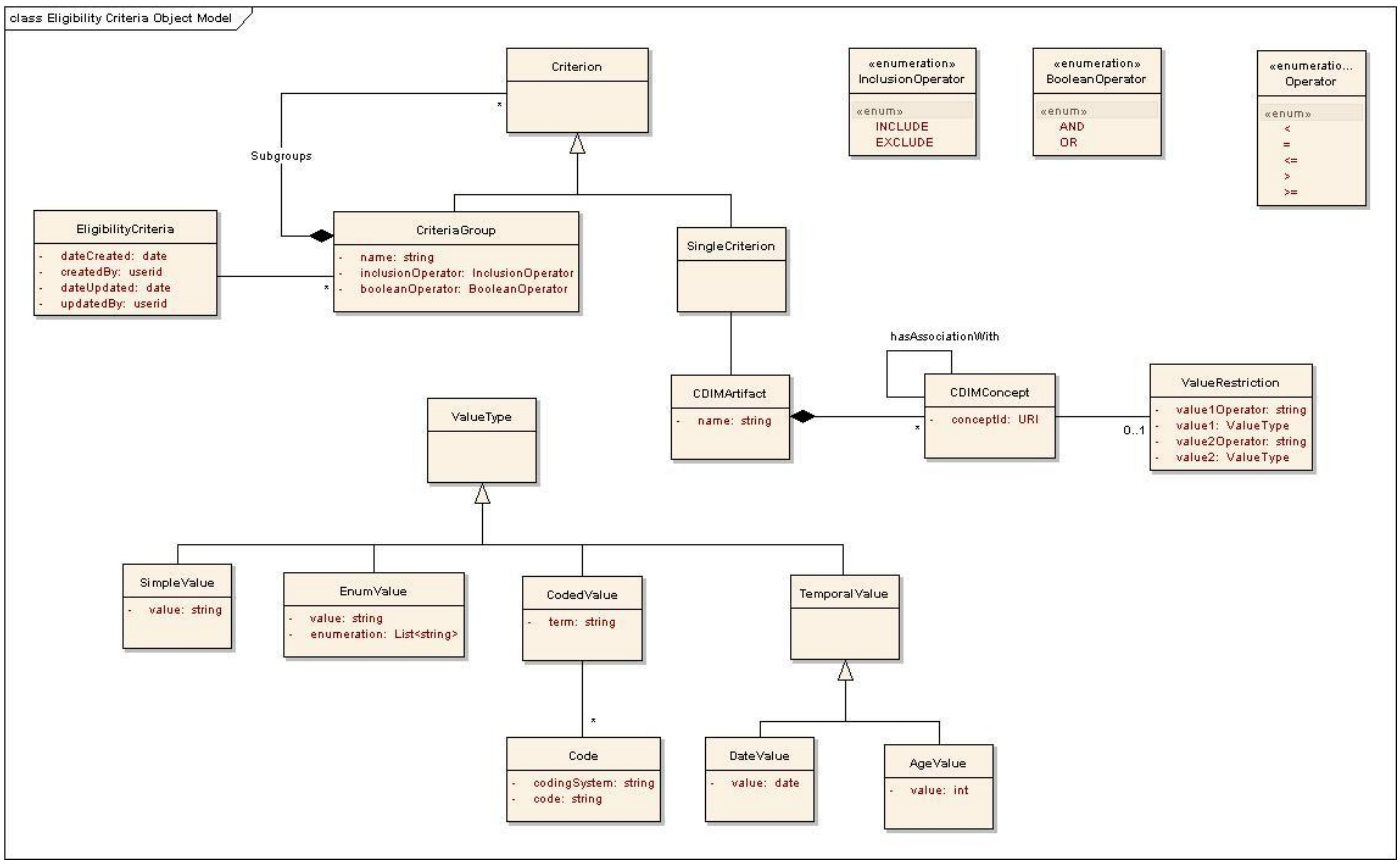

Figure 26: CDIM Artefact based Eligibility Criteria Object Model

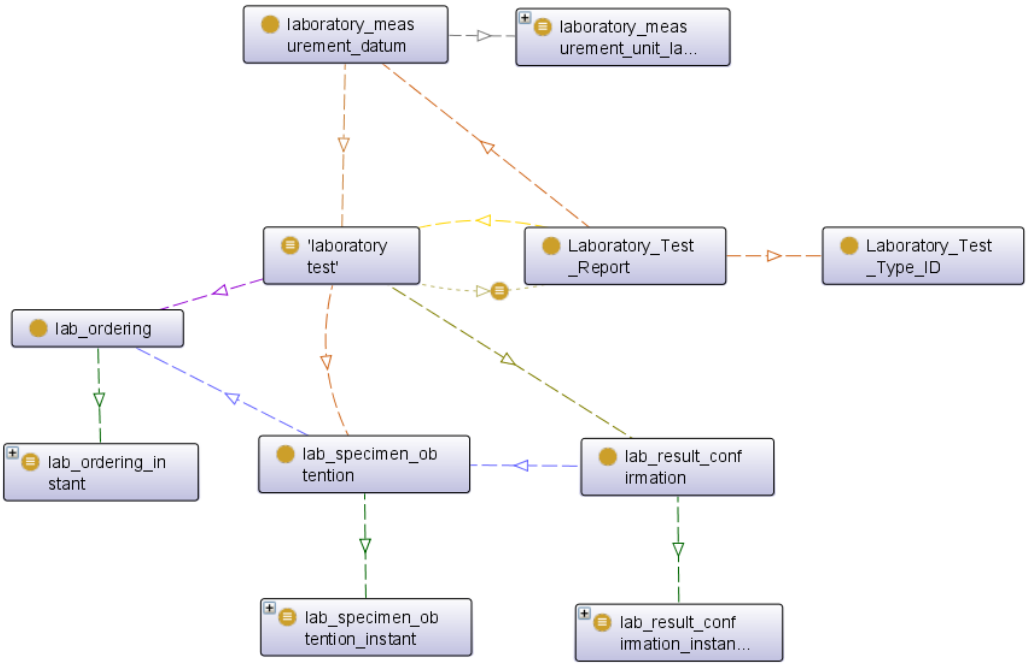

Figure 27: CDIM Artefact "Laboratory Test"

Figure 27 shows the Laboratory Test artefact, which aggregates relevant concepts closely related to the laboratory test concept extracted from CDIM. Note that some reciprocal relationships have been omitted to enhance figure readability. Table 4 lists the CDIM ontology concepts captured in the Laboratory Test artefact. At the moment, 7 artefacts have been defined for use with eligibility criteria model, including Diagnosis, Procedure, Laboratory Test, Vital Sign, Prescription, Age, and Gender. New artefacts can be easily added as per user requirements.

Core Concept: Laboratory test

Ontology Label	URI
Laboratory_Test_Type_ID	http://www.transformproject.eu/cdim_1_7.owl#CDIM_6880734
laboratory_measurement_datum	http://www.transformproject.eu/cdim_1_7.owl#CDIM_2520788
laboratory_measurement_unit_label	http://www.transformproject.eu/cdim_1_7.owl#CDIM_1484306
lab_ordering_instant	http://www.transformproject.eu/cdim_1_7.owl#CDIM_4336497
lab_specimen_obtention_instant	http://www.transformproject.eu/cdim_1_7.owl#CDIM_9589456
lab_result_confirmation_instant	http://www.transformproject.eu/cdim_1_7.owl#CDIM_0196231

Table 4: CDIM Concepts in Laboratory Test Artifact

In order to build logical rules with CDIM artefacts, a class `ValueRestriction` is introduced so each CDIM concept in the artefact can be attached to an operator and a value. `ValueRestriction` actually has two pairs of operators and values. The second pair of operator and value is optional and only used when defining a range is required. Value representation requirements would be different as per specific concepts. A simple type system is introduced to facilitate value specification. `SimpleValue` represents primitive value types such as integer, decimal, literals etc. The value itself is encoded in a string. Since the value is generally passed between JavaScript/HTML and remote database systems, a string encoding is often good enough. `EnumValue` is used for enumerated values. For example, gender will have enum values: `Male`, `Female`, `Both`. `CodedValue` is used when a list of clinical codes need to be attached to the concept. A typical usage of `CodedValue` is searching vocabulary service and binding found codes with the criterion. Finally, `TemporalValue` is needed to specify temporal restrictions. In practice, two forms of temporal restrictions are often used: date-based and age-based. Date-based restriction uses an absolute date value e.g. > 2010/01/01, while age-based restriction is relative to subject's date of birth. An example of age-based temporal restriction is presented below (Table 5). Date-based and age-based are represented by `DateValue` and `AgeValue` respectively.

A criterion of HbA1c > 6.0 % for patient aged more than 50 at the time of laboratory result confirmation could be expressed as:

Ontology Label	Operator	Value
Laboratory_Test_Type_ID	=	([LOINC;4548-4])
laboratory_measurement_datum	>	6.0
laboratory_measurement_unit_label	=	([UO;0000187])
lab_ordering_instant		
lab_specimen_obtention_instant		
lab_result_confirmation_instant	>=	(50 yrs + {http://www.transformproject.eu/cdim_1_2.owl#CDIM_8130944}

Table 5: An Example of Laboratory Test Measurement

Notes:

- LOINC: Logical Observation Identifiers Names and Codes
- 4548-4: LOINC code for "HbA1c"
- UO: The Ontology of Units of Measurement, http://code.google.com/p/unit-ontology/
- 0000187: % (ratio)
- Units are represented as a special form of CodedValue where the coding system is UO. The code value is automatically populated by the system. No knowledge of UO is assumed for end users.
- CDIM_8130944: birth_instant
- The age-based temporal restriction translates into CDIM concept birth_instant + 50 years when this eligibility criterion is submitted as a query.

5.4.4. Query Formulation Workbench: Complete Domain Object Model

The complete domain object model for the software implementation of the query formulation workbench is presented in Figure 28. It is based on the conceptual model developed in section 5.4.2. This domain model includes the eligibility criteria model as elaborated in the previous section and other domain objects such as `Study`, `Protocol`, `Query`, as well as query results. The `Study` class records various pieces of information about a study, such as name, reference number, when the study was created, who created the study, when the study was updated, etc. A study has multiple protocols, with each protocol being a different version or different study design. Each `Protocol` has one `EligibilityCriteria`, which comprises one or many `CriteriaGroup`. Each `CriteriaGroup` contains name, inclusionOperator and booleanOperator. `InclusionOperator` indicates whether this group is an inclusion group or an exclusion group. `BooleanOperator` decides whether the subgroups and single criterions within the group are ANDed or ORed. `SingleCriterion` is based on `CDIMArtifact`, the name of which indicates its semantic type such as diagnosis, laboratory measurement, prescription, demographic information such as age, gender, etc. Each `CDIMArtifact` is mapped to a list of `CDIMConcept` following CDIM ontology definition. `CDIMConcept` may be attached with value restriction to establish query rules. `ValueRestriction` can have two operator/value pairs. The second pair is optional and only used to define a range. The value may be a simple numeric value, enumerated value, coded value, or a temporal value, so defined as 4 value types. One `EligibilityCriteria` can be submitted for query multiple times so is associated with multiple `Query`. The `Query` class tracks the overall status of the federated query, reports the aggregated count, and records the start time and end time. The overall status can be all

query tasks running and no data source returns, partially completed with some results available, all tasks completed, partially completed and terminated by user. Effectively each data source has one query task. `QueryTask` tracks the status of individual query task and records its returned count and its end time. The status of a single query task can be processed, completed, failed or terminated by user. If that query task has failed, an error code is returned to indicate the exact failure reason, and a human readable text message will be displayed to the end user about the error. Criterion sensitivity analysis is supported with a breakdown of counts by criterion (`CountPerCriterion` links to both `CriteriaGroup` and `SingleCriterion`).

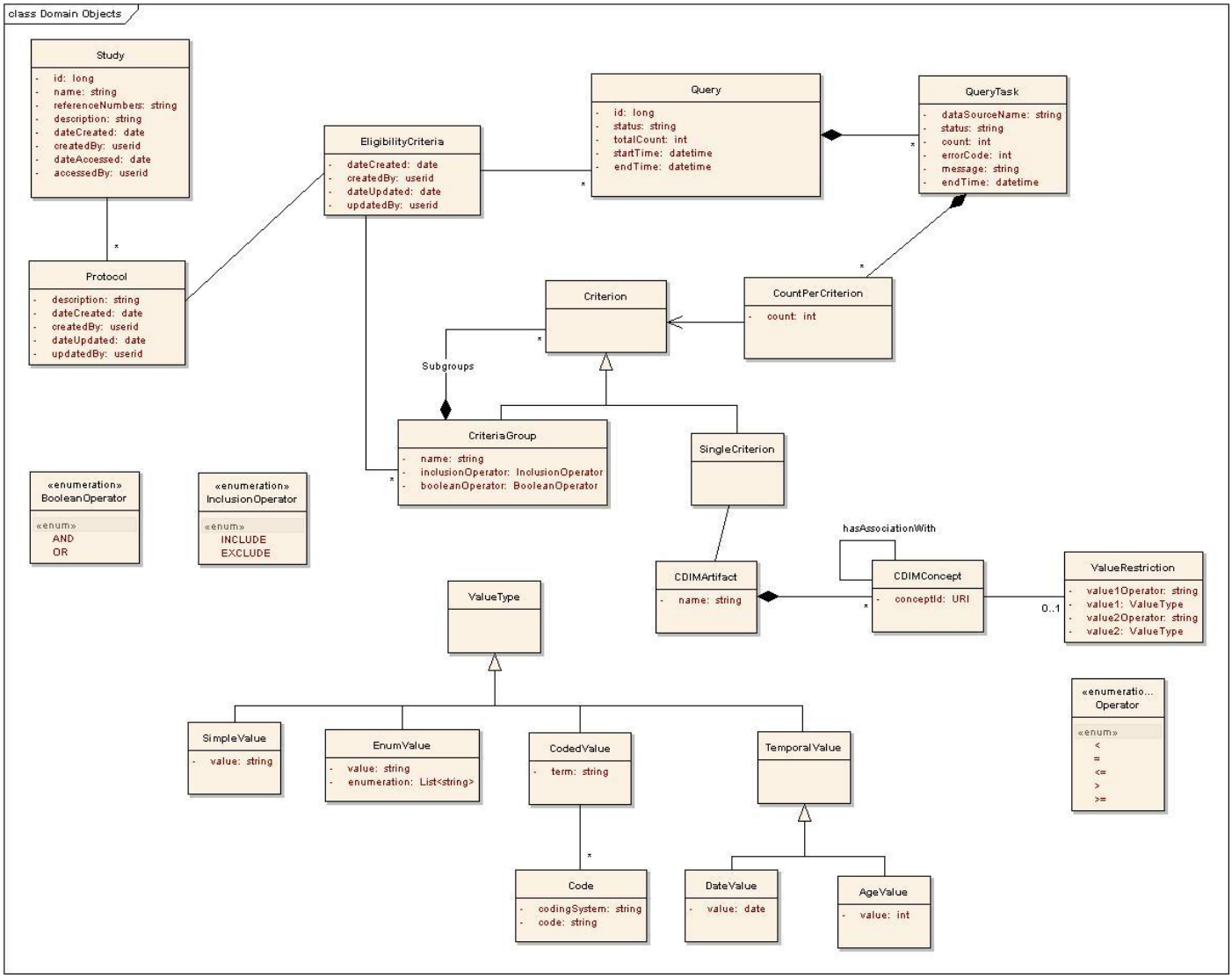

Figure 28: Domain Object Model

5.5. Summary

In the previous section we presented the technical and implementation details of the TRANSFoRm web-based query formulation workbench. The system is designed as a typical layered web application based on Spring framework and Hibernate ORM. An Ajax-based rich UI is developed using JQuery to allow for dynamic interfaces and easy integration with the underlying web application.

Design and running of queries is assumed to be team based, where concurrent access is typical. With the introduction of a simple write locking mechanism, the system protects shared objects from conflicting updates.

The main aim of the query formulation workbench at this stage is to support eligibility criteria design and retrieval of an eligible subject count, the core of which is built around a sound eligibility criteria model. The whole TRANSFoRm system is adopting an ontology-driven approach where a unified and consistent view of diverse data source representations is provided through an ontology model, i.e. CDIM.

The query formulation workbench eligibility criteria implementation model is based on CDIM artefacts, where closely related CDIM concepts are grouped together to form a higher-level abstraction which closely matches the application domain. The eligibility criteria model is translated into CDIM concepts with associated value restrictions when the model is submitted for query.

The query formulation workbench and its associated UI is made provenance-aware, and interoperates with a variety of other TRANSFoRm components, including the vocabulary service, distributed infrastructure, CDIM information services and provenance service.

6. Concluding Remarks

The current deliverable discusses the design and implementation of the TRANSFoRm Query Formulation Workbench. This is a semantically aware software tool that supports the easy authoring of distributed searches to EHR and other clinical data sources, using a controlled vocabulary service and appropriate standards-based technological solutions. The main aim of the workbench is to automatically identify 'prevalent cases' for research, where the searches report back counts of eligible subjects in the EHRs, flagging the subjects for recruitment and consent by the local clinical care team, in full compliance with data protection legislation and best practice.

The design and implementation of the query formulation workbench was based on a requirements-driven approach. To gather the appropriate functional and user requirements for the tool, we followed the combined analysis of the use cases provided by deliverable D1.1 and the information elicited by conducting participatory task modelling, involving a group of expert users. We enhanced this knowledge through task analysis work, based on users and existing literature. As part of latter, we conducted a literature review of existing eligibility criteria query interfaces describing the desirable features of a query tool. As part of this review, we also highlighted, by means of comparison, some of the limitations of existing solutions and elaborated on how TRANSFoRm addresses and improves these on its query formulation workbench solution.

We demonstrated the overall functionality of the query formulation workbench requirements and associated UI on an example study. We showed how the tool can translated the eligibility needs of the diabetes use case. The UI provides mechanisms and simple visual representations of groupings of eligibility criteria, which can be easily managed by users as a series of simple interaction steps, guiding the user throughout the process.

Integration with other TRANSFoRm components, such as the security framework and CDIM information service, has provided a successful working functionality of the query formulation within the TRANSFoRm distributed infrastructure. Users can work together on study and protocol design, with rules for collaborative access, driven by the predefined user requirements. Study protocols are designed using an intuitive interface where authorisation rules are used to restrict access to a user's individual or group permissions. The use of the TRANSFoRm terminology services, in conjunction with the Clinical Data Integration Model CDIM allows the capturing eligibility criteria in a computable representation, based on CDIM ontology, so the criteria can be translated into executable query statements at the individual EHR data sources. Integration with the TRANSFoRm provenance service also allows for further auditing of user actions.

References

1. TRANSFoRm, Translational Research And Patient Safety In Europe, ICT-2009.5.2-247787, Annex I-Description of Work, 09/12/2011.

2. Leysen P, Bastiaens H, van Royen P, Agreus L, Andreasson AN. Development of Use Cases [Internet]. 2011 Feb. Report No.: TRANSFoRm Deliverable D1.1, V2.1. Available from: https://transform.kcl.ac.uk/groups/publicdeliverables/wiki/welcome/attachments/ff213/TRANSFoRm%20WP1%20Detailed%20Use%20Cases_V2.1.pdf

3. O'Neill E, Johnson P. Participatory task modelling: users and developers modelling users' tasks and domains. Proceedings of the 3rd annual conference on Task models and diagrams [Internet]. New York, NY, USA: ACM; 2004 [cited 2012 Jun 25]. p. 67–74. Available from: http://doi.acm.org/10.1145/1045446.1045460

4. Muller MJ, Kuhn S. Participatory Design. Communications of the ACM. 1993 Jun;36(6):24–8.

5. Bastiaens H. Data elements for the use case diabetes (DM_usecase_for_GIM_06jan2012.doc). 2012.

6. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: A literature review. Journal of Biomedical Informatics. 2010 Jun;43(3):451–67.

7. ePCRN. The electronic Patient Care Research Network [Internet]. Available from: http://www.epcrn.bham.ac.uk/

8. Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FDR. Envisioning a Learning Health Care System: The Electronic Primary Care Research Network, A Case Study. Annals of Family Medicine. 10(1):54–9.

9. US National Library of Medicine. Chapter 2: Metathesaurus. UMLS Reference Manual [Internet]. 2009 [cited 2011 Aug 15]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9684/pdf/ch02.pdf

10. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. Journal of the American Medical Informatics Association : JAMIA. 2009; 16(5):624–30.

11. Partners Healthcare. i2b2 Web Client version 1.6.04 [Internet]. [cited 2012 May 29]. Available from: https://www.i2b2.org/software/#downloadables

12. McMurry A. Shrine Core Ontology [Internet]. 2011 [cited 2012 May 30]. Available from: https://open.med.harvard.edu/display/SHRINE/Core+Ontology

13. SHRINE Data Reference [Internet]. Available from: https://shrine.catalyst.harvard.edu/shrine-webclient/Help/DR_Help_Files/Toc252797232.htm

14. SHRIMP User Guide [Internet]. [cited 2012 May 30]. Available from: https://open.med.harvard.edu/display/SHRIMP/User+Guide

15. North West e-Health [Internet]. [cited 2012 Apr 27]. Available from: http://www.nweh.org.uk/

16. Ainsworth J, Buchan I. Preserving consent-for-consent with feasibility-assessment and recruitment in clinical studies: FARSITE architecture. HealthGrid 2009 [Internet]. Berlin; 2009. Available from: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:7155

17. Medical Research Council. Data and Tissues Toolkit: Draft Guidance on Consent for Consent [Internet]. 2007. Available from: http://www.dt-toolkit.ac.uk/_db/_documents/Consent_for_Consent_DRAFT_200903034252.pdf13.

18. Medical Research Council. Consent arrangements: Should consent be sought? [Internet]. Data and Tissues Tool Kit. [cited 2012 Apr 30]. Available from: http://www.dt-toolkit.ac.uk/routemaps/station.cfm?current_station_id=427

19. Medical Research Council. Personal Information in Medical Research (MRC Ethics Series) [Internet]. 2000 [cited 2012 Apr 30]. Available from: http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002452

20. Elliot M, Purdam K, Smith D. Statistical disclosure control architectures for patient records in biomedical information systems. Journal of Biomedical Informatics. 2008 Feb;41(1):58–64.

21. Rector AL, Rogers JE, Zanstra PE, van der Haring E. OpenGALEN: Open Source Medical Terminology and Tools. AMIA Annual Symposium Proceedings 2003 [Internet]. American Medical Informatics Association; 2003 [cited 2012 May 3]. p. 982. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480228/

22. NorthWest e-Health. FARSITE Feasibility Demonstration Video [Internet]. 2012. Available from: http://youtu.be/pOTfGXbW0Zg

23. Thew S, Leeming G, Ainsworth J, Gibson M, Buchan I. FARSITE: evaluation of an automated trial feasibility assessment and recruitment tool. Trials 2011. 2011;12 (Suppl 1):A113.

24. Zhang G-Q, Siegler T, Saxman P, Hunscher D, Arabandi S. VISAGE: A Query Interface for Clinical Research. AMIA Summits on Translational Science Proceedings. 2010;2010:76–80.

25. Case Western Reserve University. Physio-MIMI - Multi-Modality Multi-Resource Physiological and Clinical Informatics Infrastructure [Internet]. [cited 2012 May 31]. Available from: http://physiomimi.case.edu/physiomimi/index.php/Physiomimi:About

26. Physio-MIMI Development Team. Physio-MIMI VISAGE End User Guide [Internet]. 2010 [cited 2012 May 31]. Available from: http://physiomimi.case.edu/physiomimi/images/8/84/VISAGE_UserManual.pdf

27. ASTM International. ASTM E2369 - 05e1 Standard Specification for Continuity of Care Record (CCR) [Internet]. [cited 2011 Aug 16]. Available from: http://www.astm.org/Standards/E2369.htm

28. Danger R, Curcin V. Application of Provenance Model [Internet]. 2012 May. Report No.: TRANSFoRm Deliverable D5.2, V2.0. Available from: https://transform.kcl.ac.uk/groups/publicdeliverables/wiki/welcome/attachments/d907a/D5.2%20Application%20of%20Provenance%20Model_final.pdf

29. Kuchinke W, Karakoyun T, Ohmann C. Clinical Research Information Model [Internet]. 2012 May. Report No.: TRANSFoRm Deliverable D6.2, V1.0. Available from: https://transform.kcl.ac.uk/groups/publicdeliverables/wiki/welcome/attachments/03b0e/TRANSFoRm-Del_6%202_CRIM-v1_fin_31.5.2012-finf.pdf

30. Lim Choi Keung SN, Zhao L, Tyler E, Arvanitis TN. Integrated Vocabulary Service for Health Data Interoperability. Fourth International Conference on eHealth, Telemedicine and Social Medicine (eTELEMED 2012). Valencia, Spain: IARIA; 2012. p. 124–7.

31. US National Library of Medicine. Unified Medical Language System (UMLS) [Internet]. 2012 [cited 2012 Jun 16]. Available from: http://www.nlm.nih.gov/research/umls/

32. National Cancer Institute. LexEVS Server and API [Internet]. caBIG. [cited 2012 Jun 15]. Available from: https://cabig.nci.nih.gov/community/tools/LexEVS_Server

33. Anjum A, Curcin V. TRANSFoRm Provenance Framework [Internet]. 2011 Mar. Report No.: TRANSFoRm Deliverable D3.1, V1.0. Available from: https://transform.kcl.ac.uk/groups/publicdeliverables/wiki/welcome/attachments/d8849/ProvenanceFrameworkDesign_v1.0.1.pdf

Appendix A

Balsamiq is a rapid prototyping package for user interface design which is typically used in collaborative environments. Prototyping software such as Balsamiq allows users to quickly see visual mock ups, without developing the underlying software necessary to support them as full applications. The following mock ups describe a series of typical workflows in the prototype TRANSFoRm query tool.

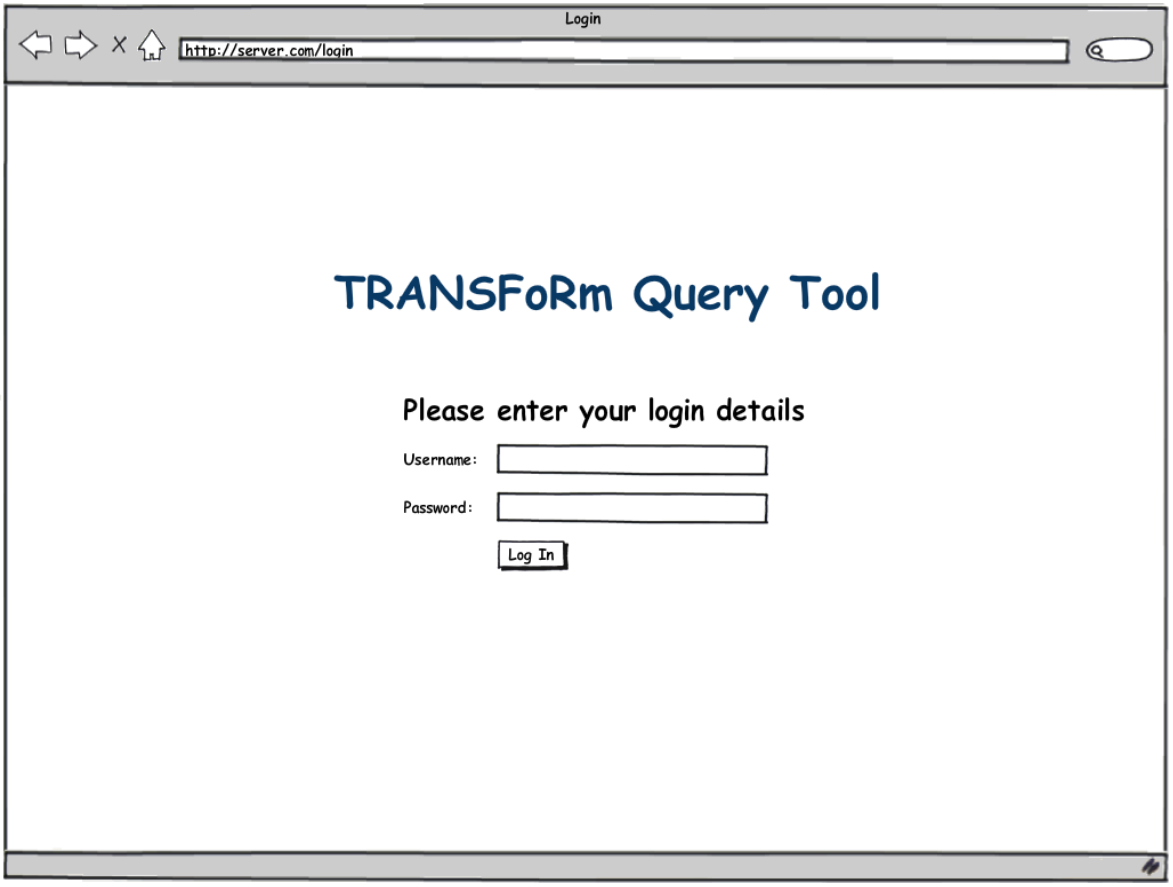

Figure 29: UI 1 Login

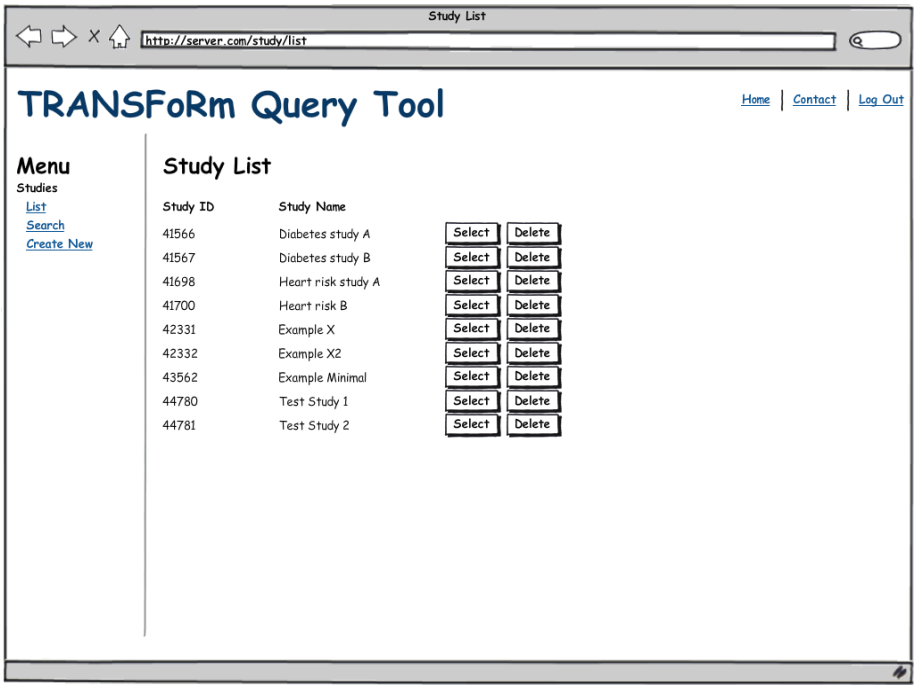

Figure 30: UI 2 Study List

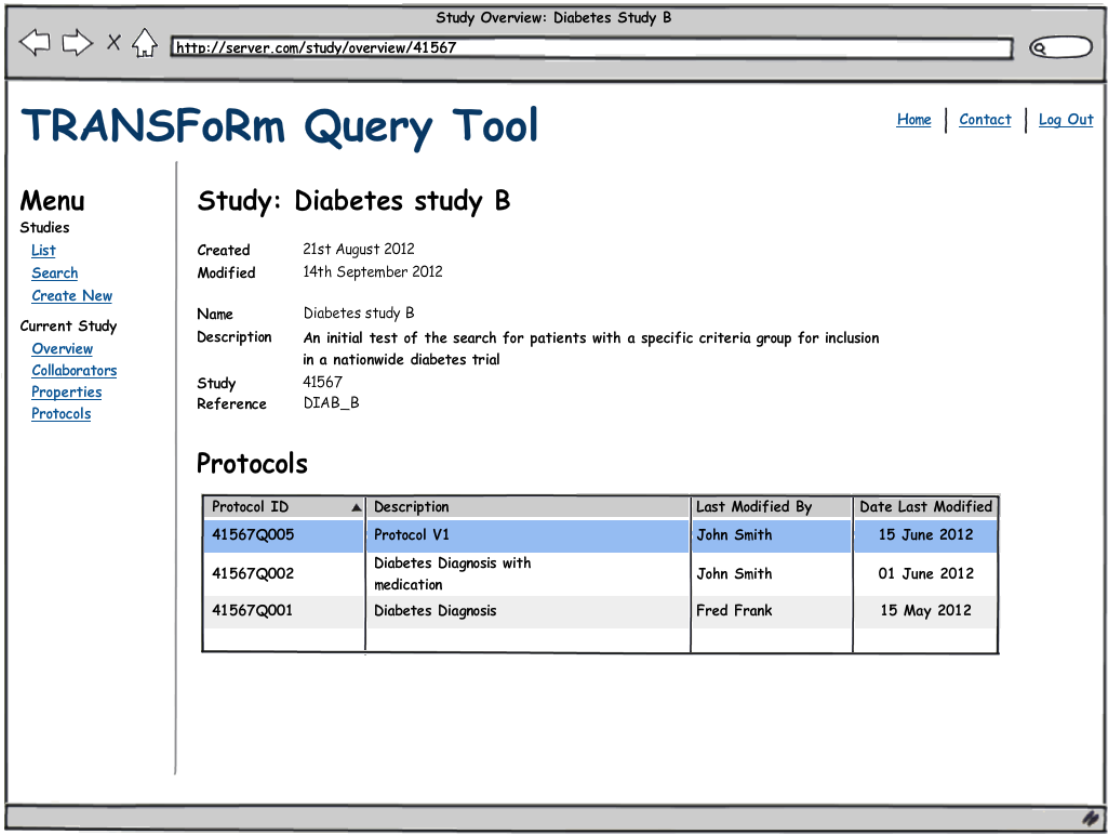

Figure 31: UI 3 Study Overview

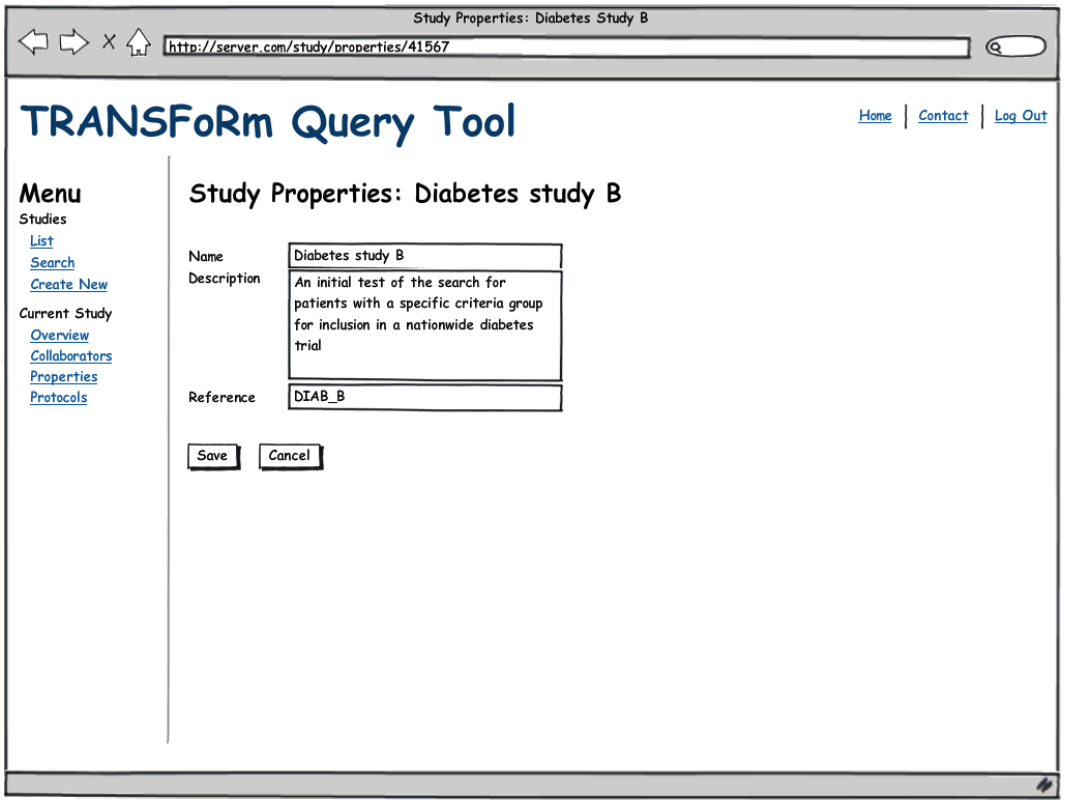

Figure 32: UI 4 Study Properties

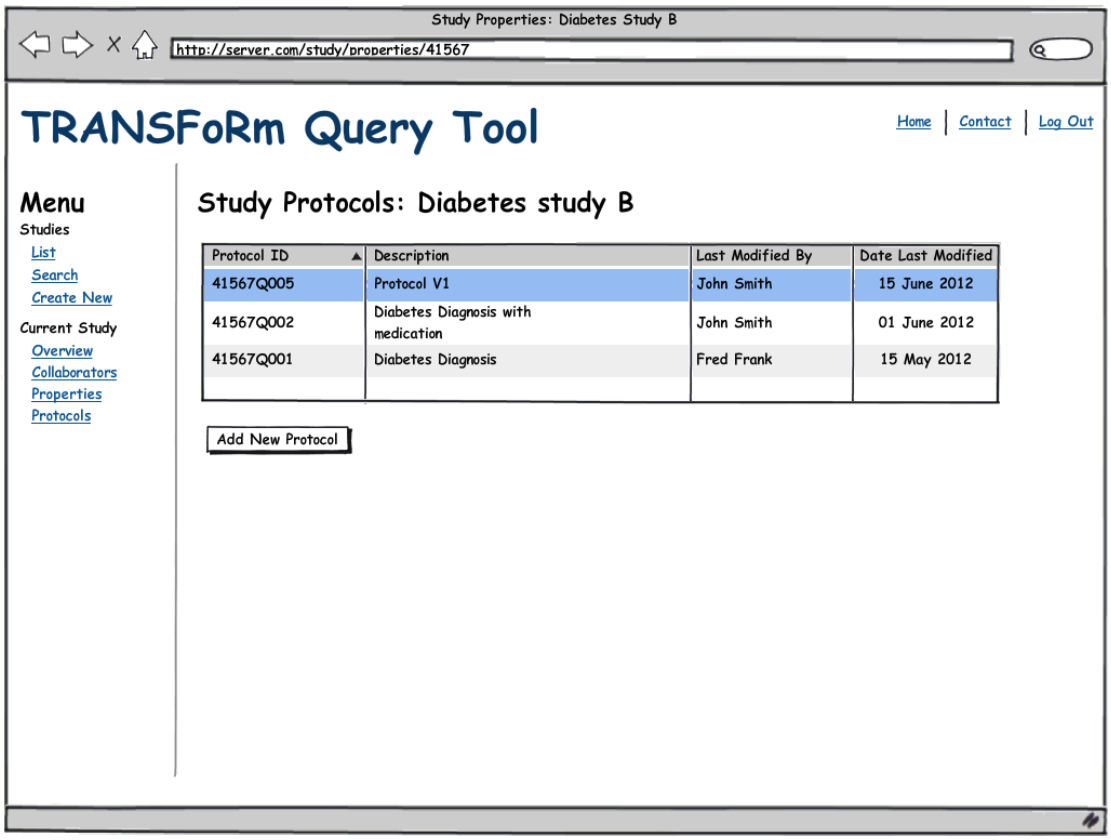

Figure 33: UI 5 Study Protocols

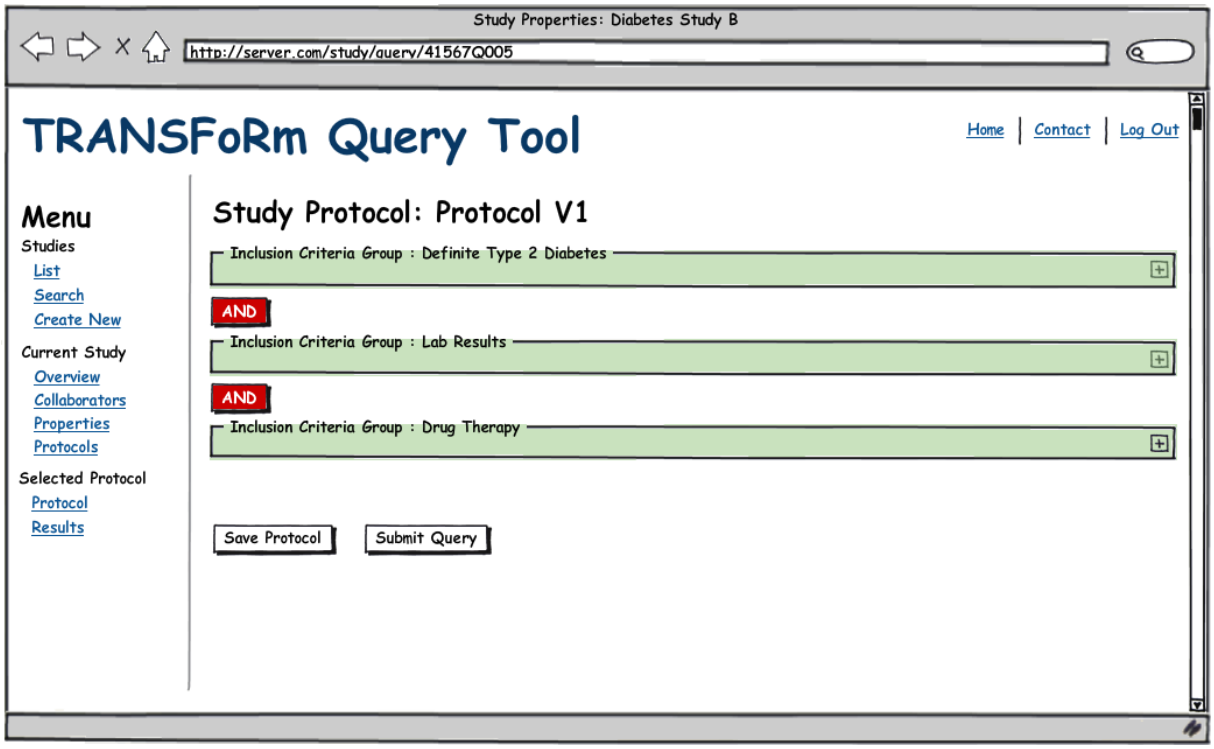

Figure 34: UI 6 Protocol – Collapsed

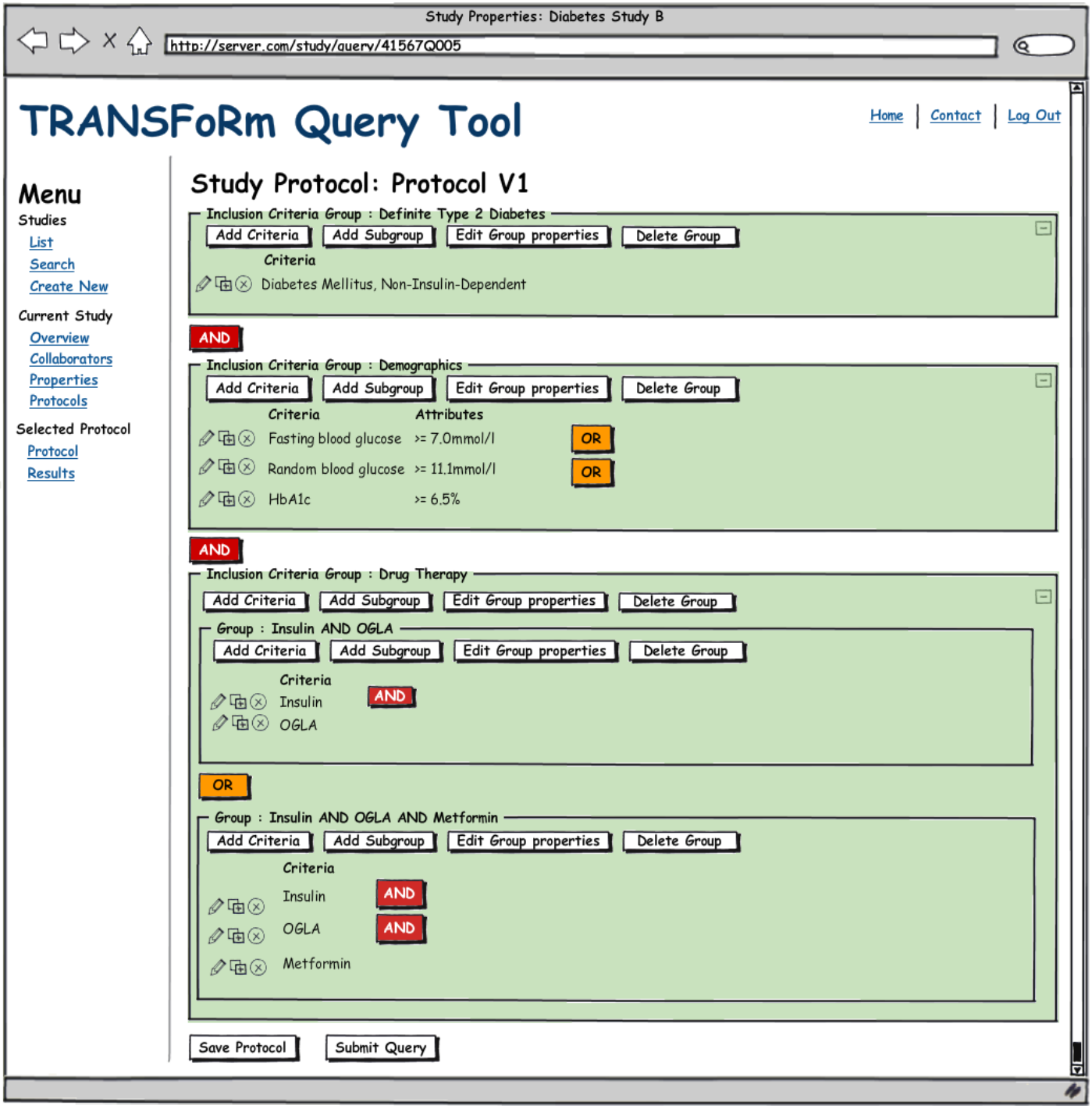

Figure 35: UI 7 Protocol - Expanded

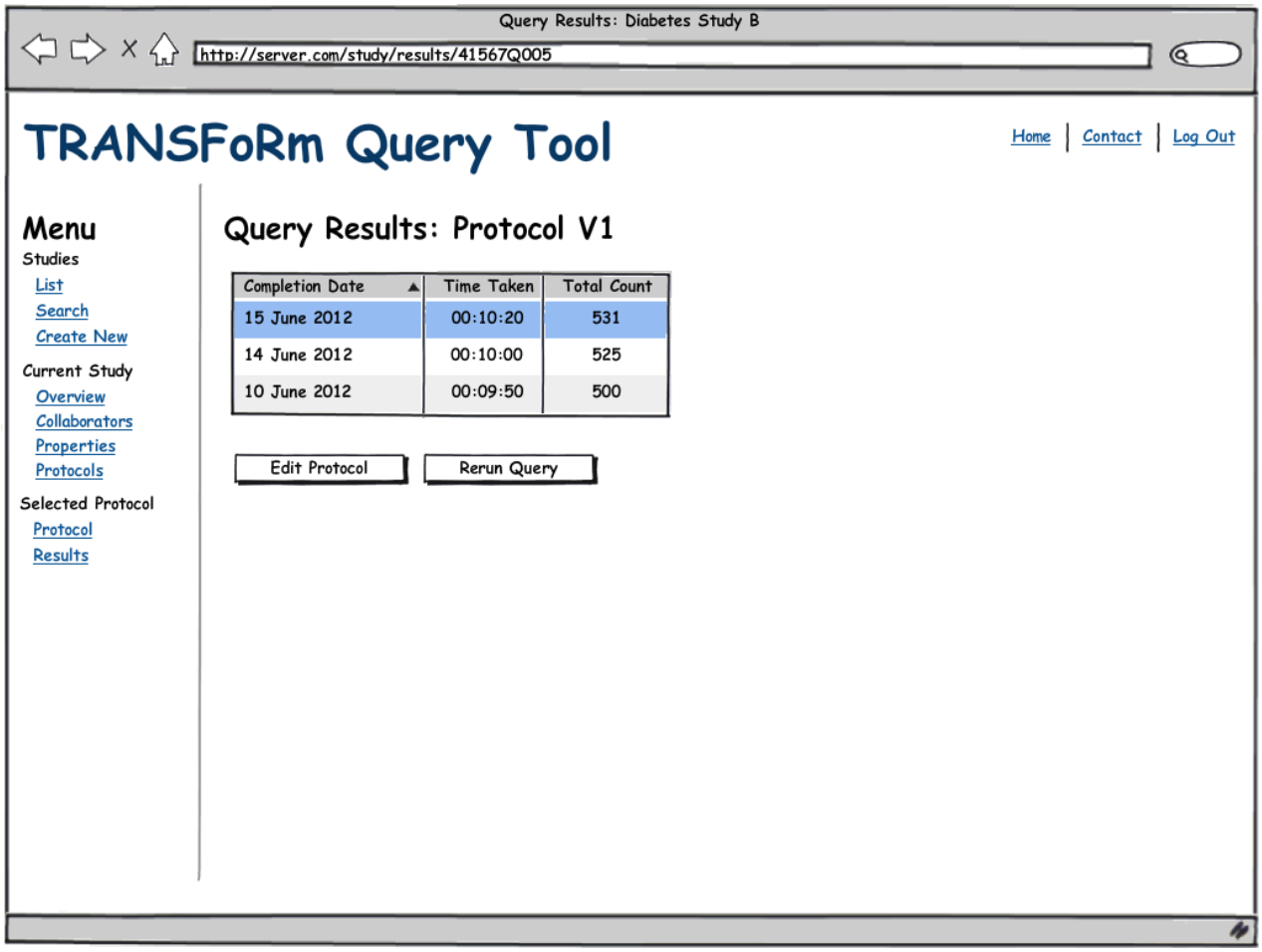

Figure 36: UI 8 Query Results

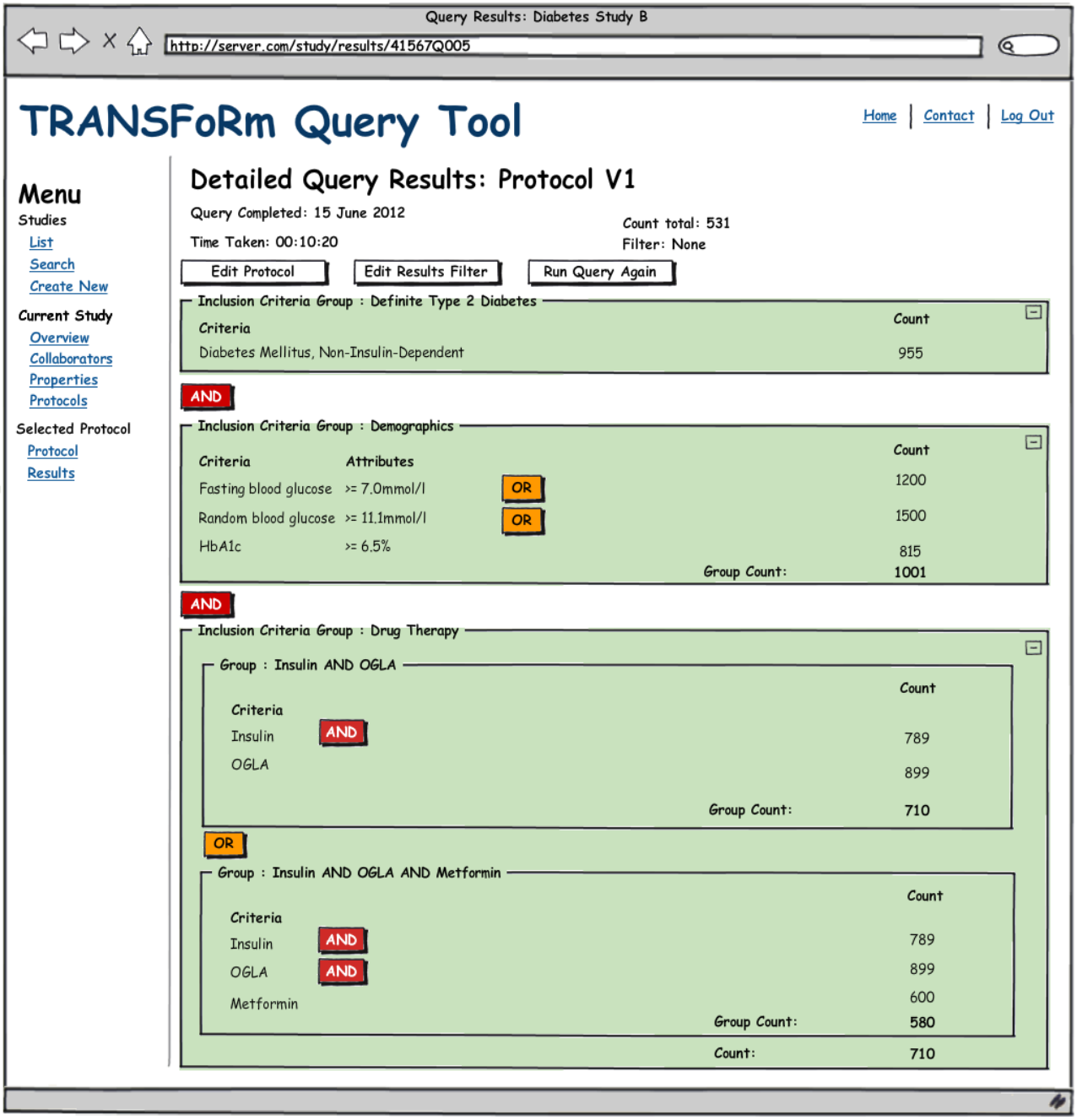

Figure 37: UI 9 Query Results – Detailed

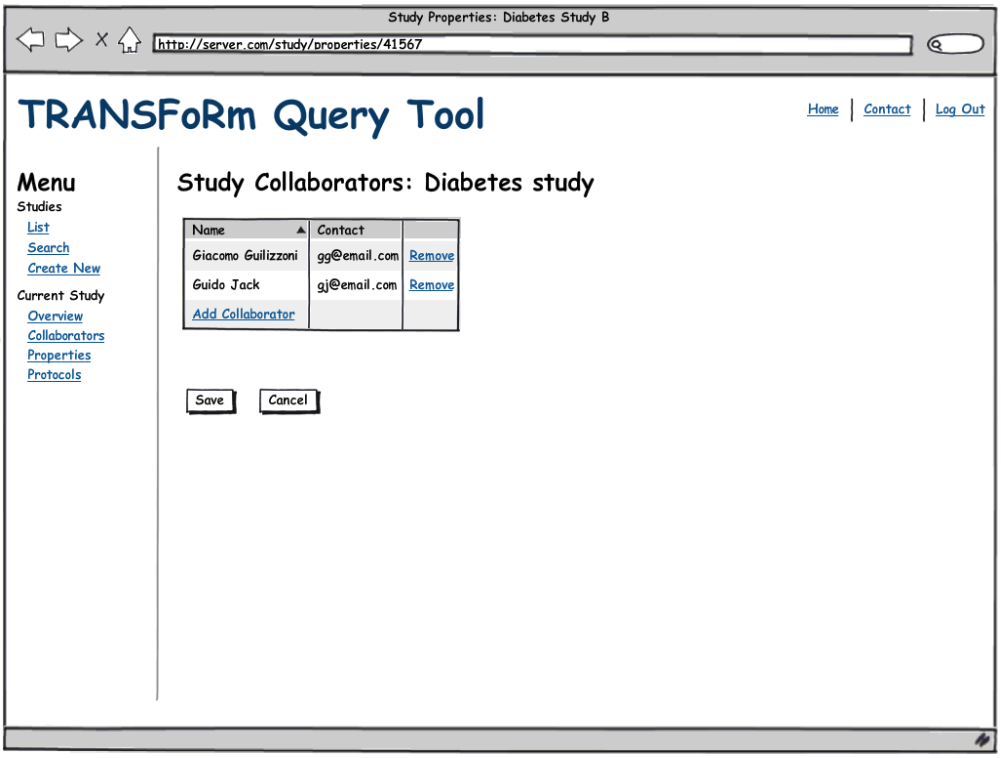

Figure 38: UI 10 Study Collaborators

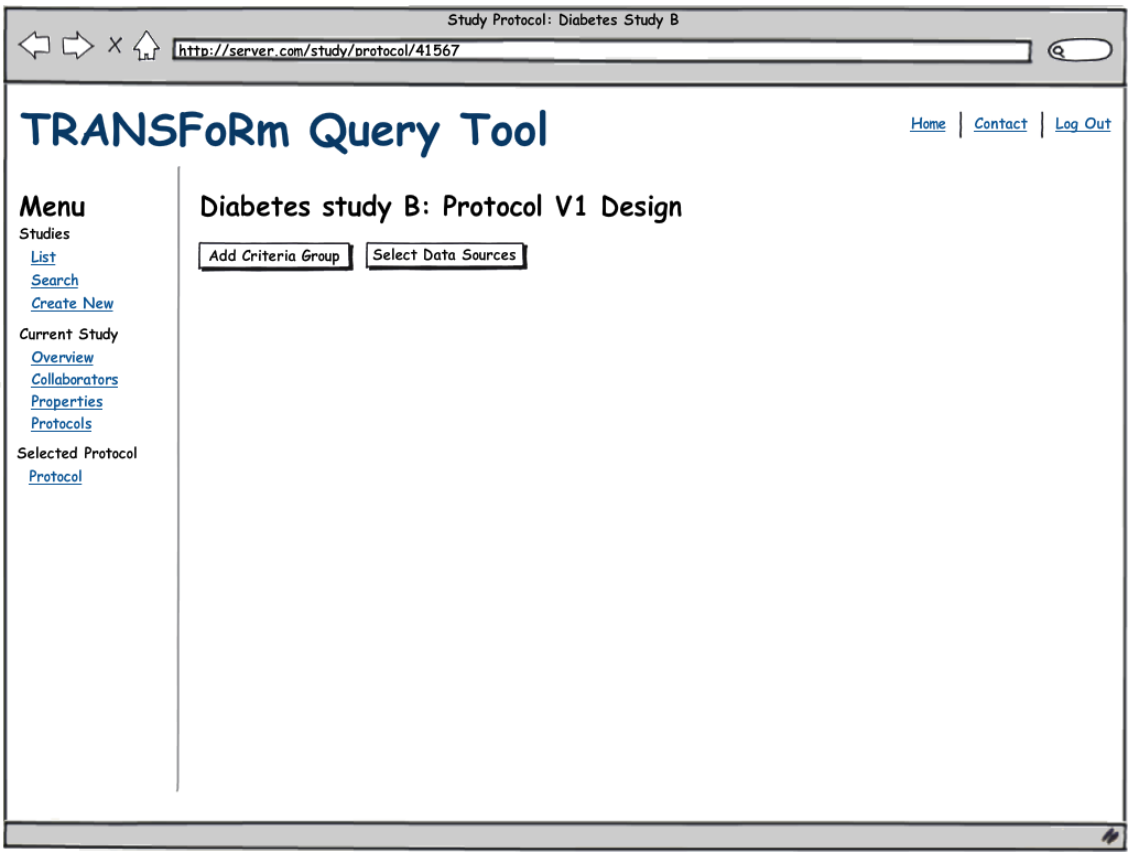

Figure 39: UI 11 Protocol Start

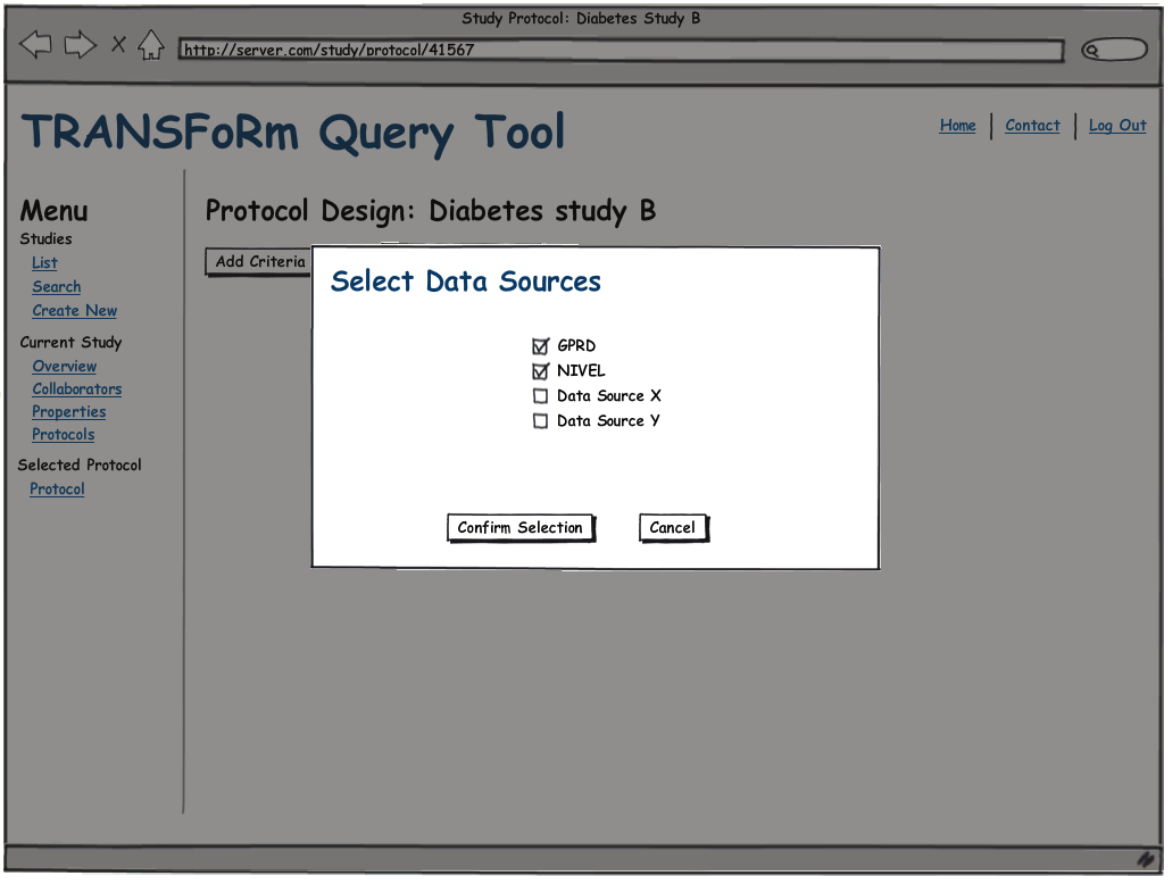

Figure 40: UI 12 Data Source Selection

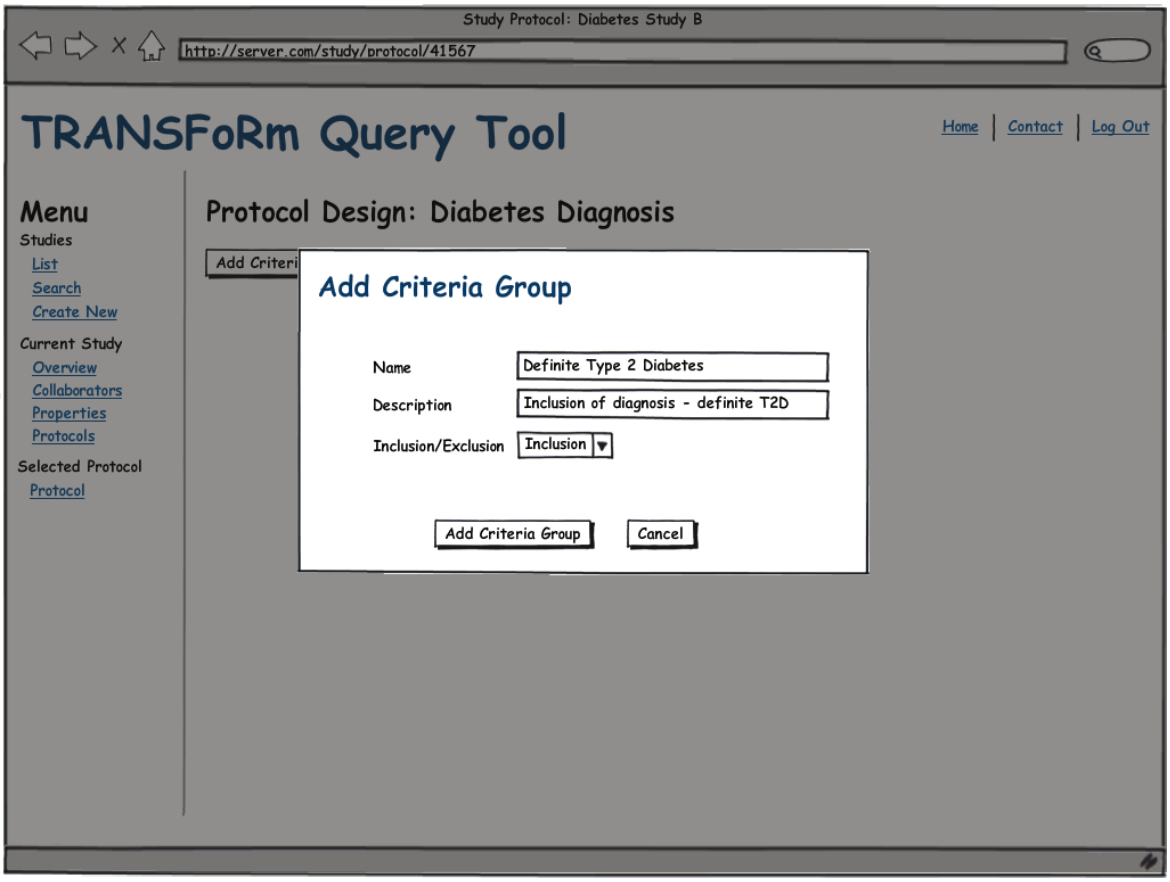

Figure 41: UI 13 Add Criteria Group

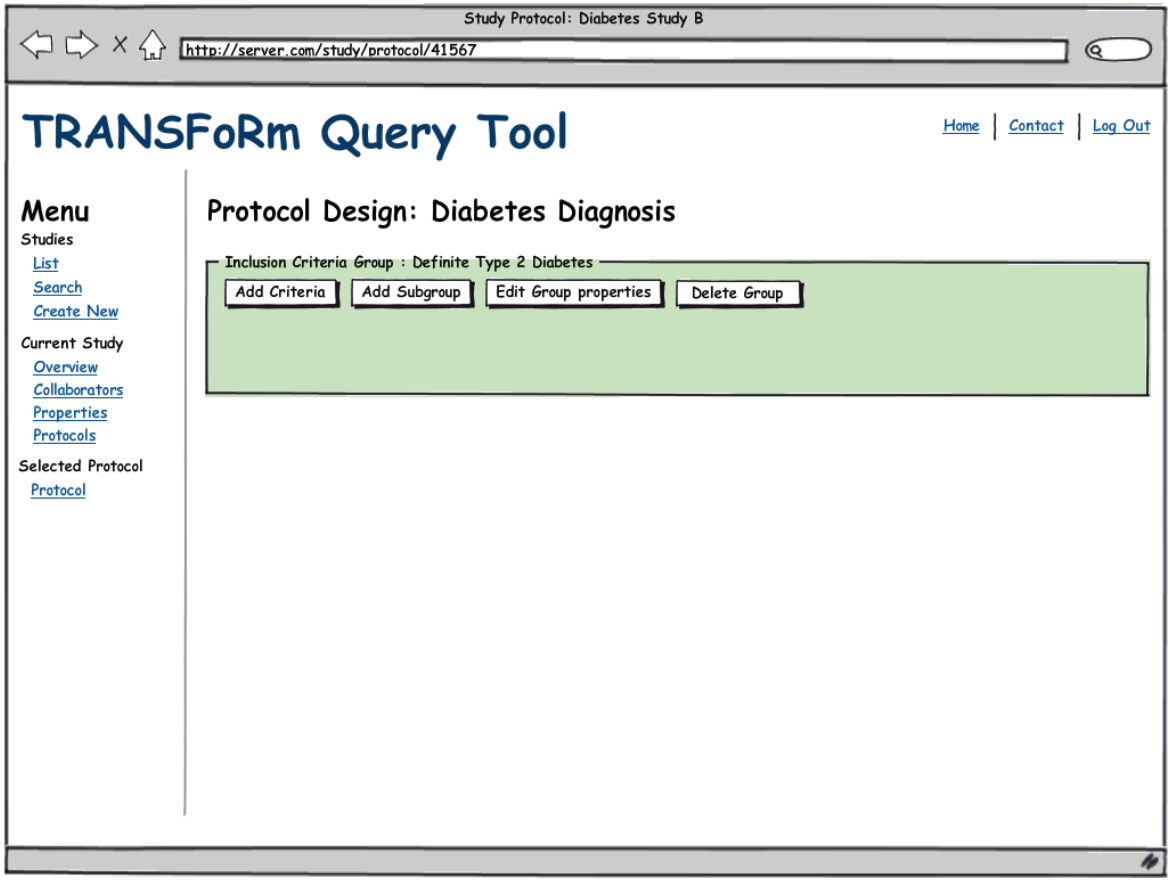

Figure 42: UI 14 Criteria Group Added

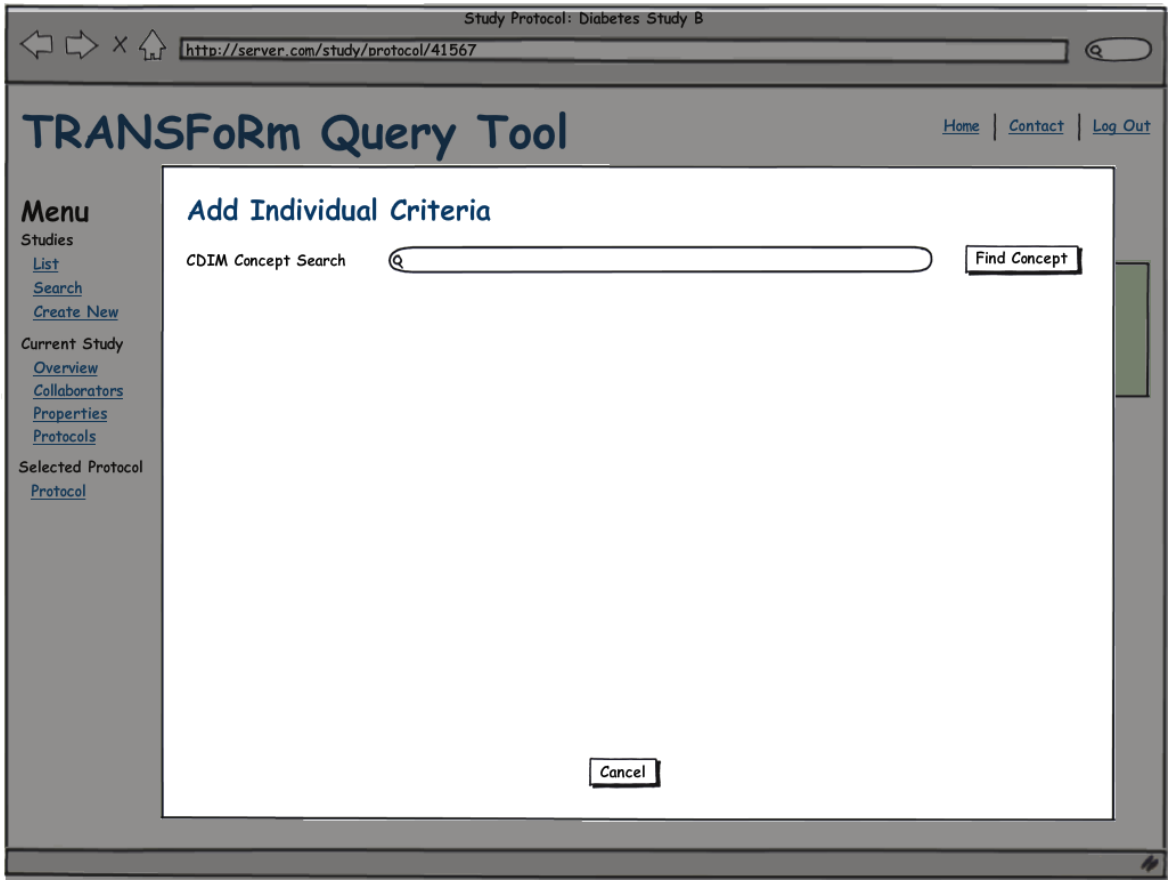

Figure 43: UI 15 Choose CDIM Concept

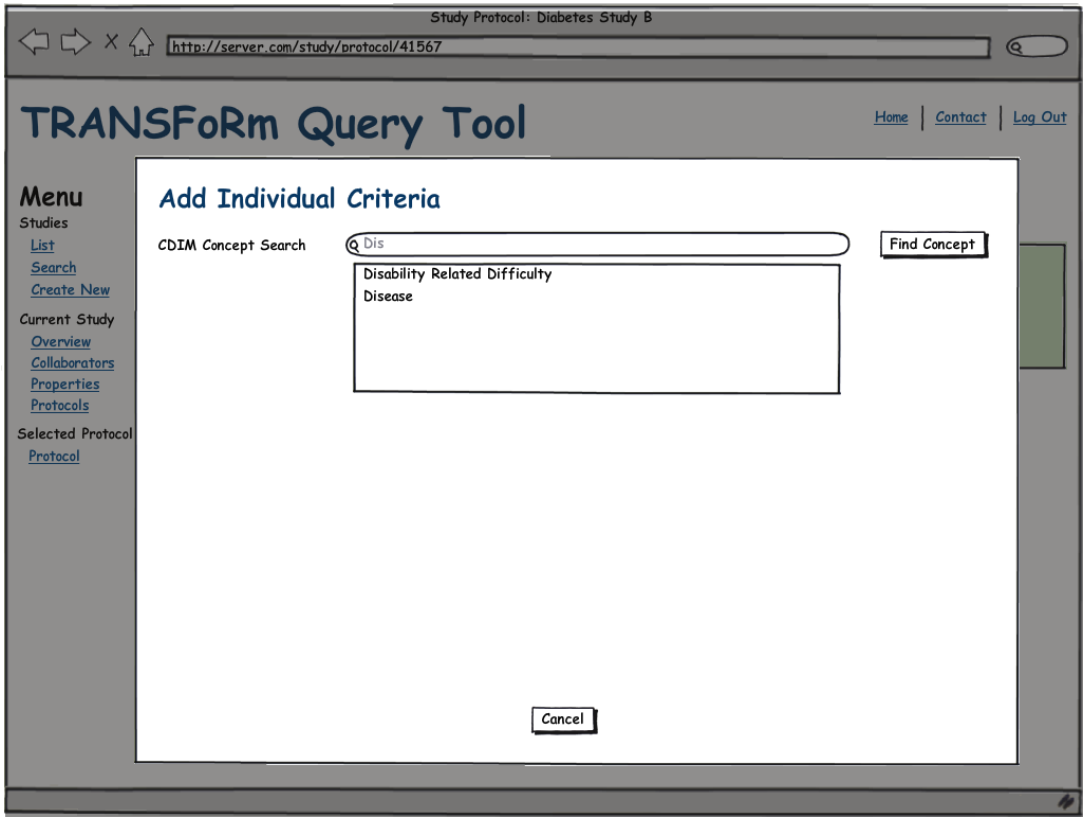

Figure 44: UI 16 CDIM Concept Dynamic Filtering

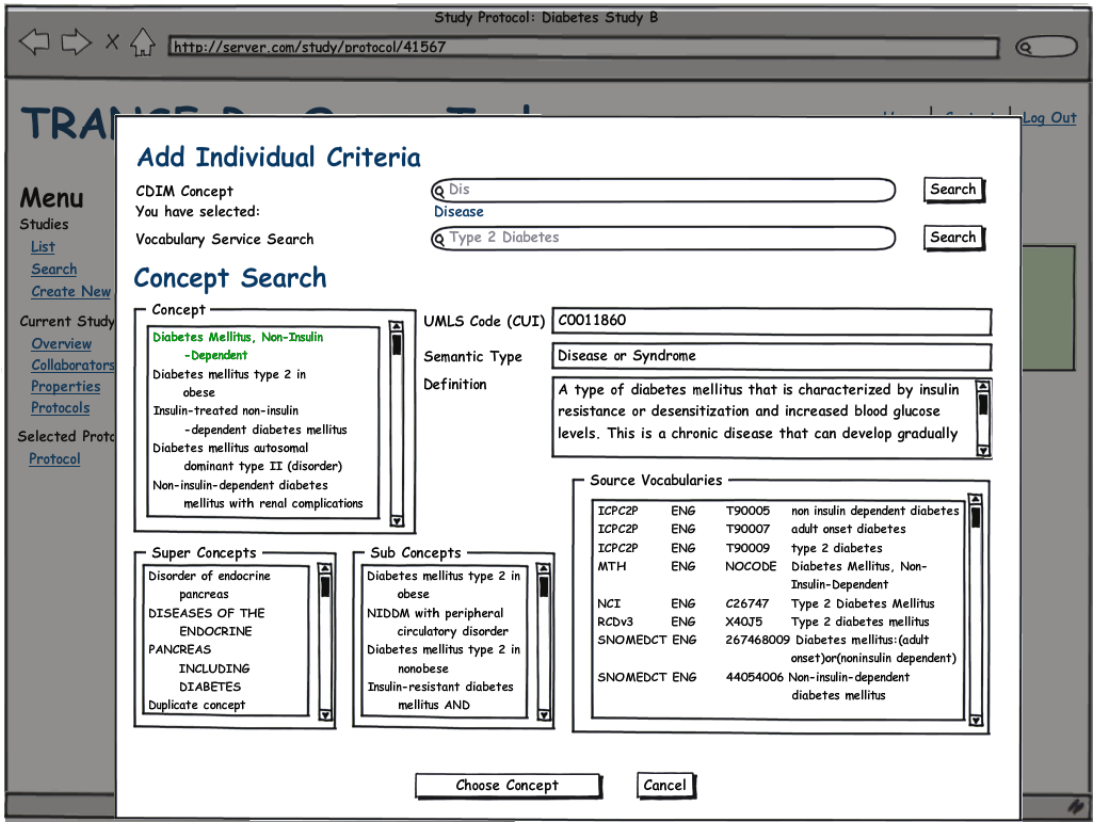

Figure 45: UI 17 Select Vocabulary Service Concept

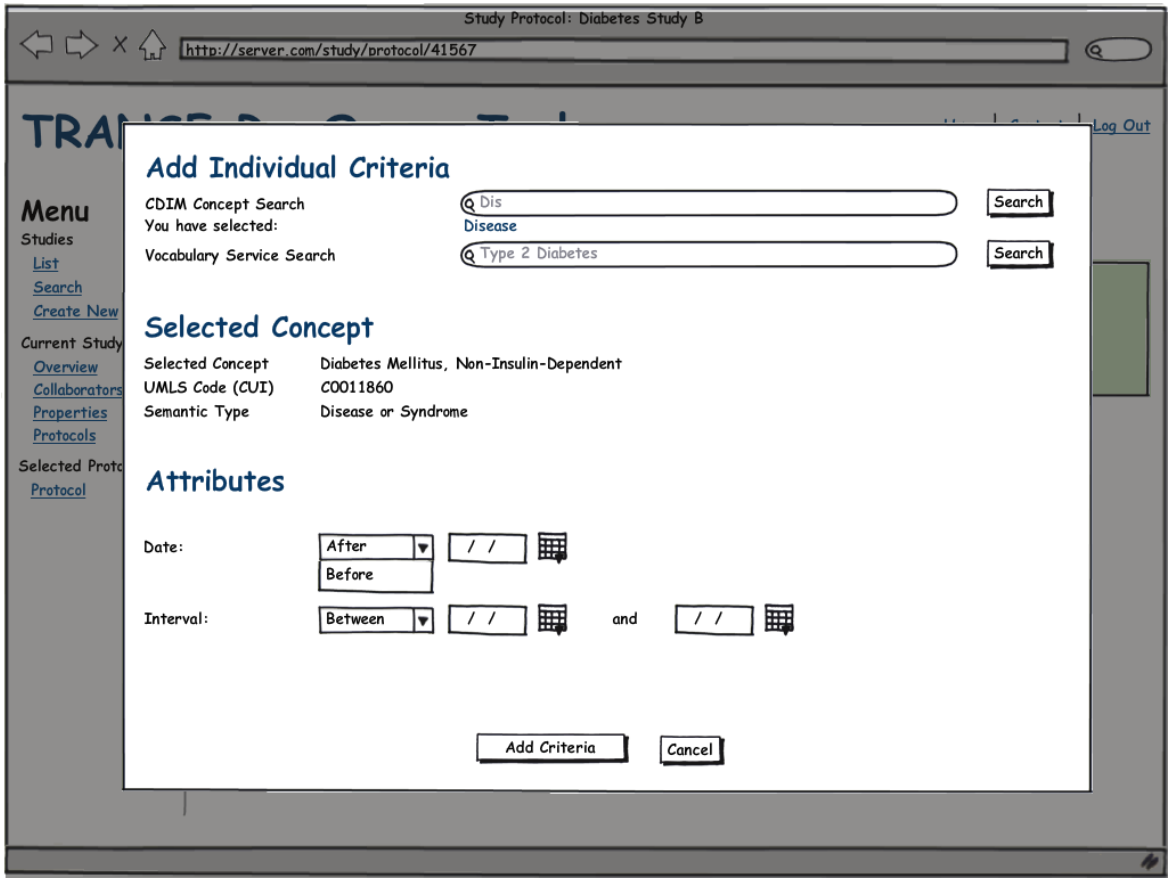

Figure 46: UI 18 Concept Attributes – Diagnosis

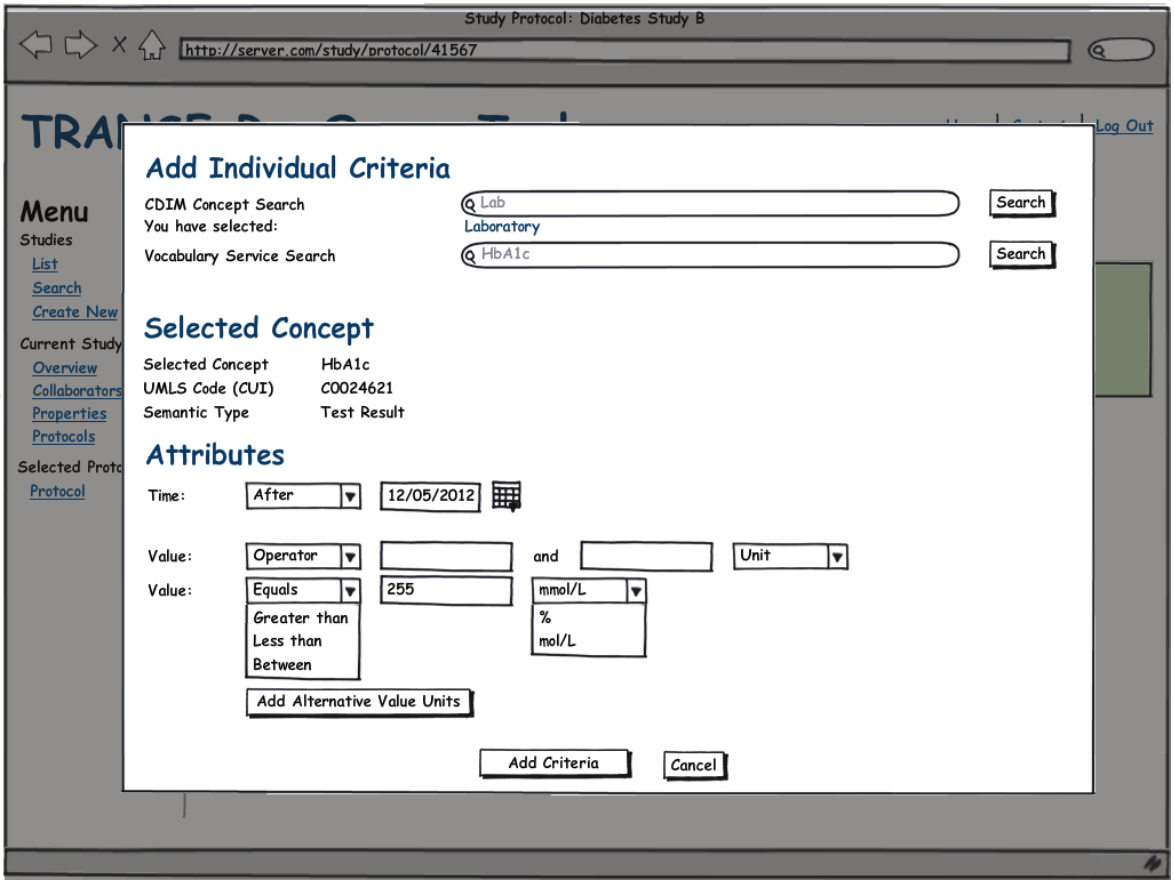

Figure 47: UI 19 Concept Attributes - Lab Result

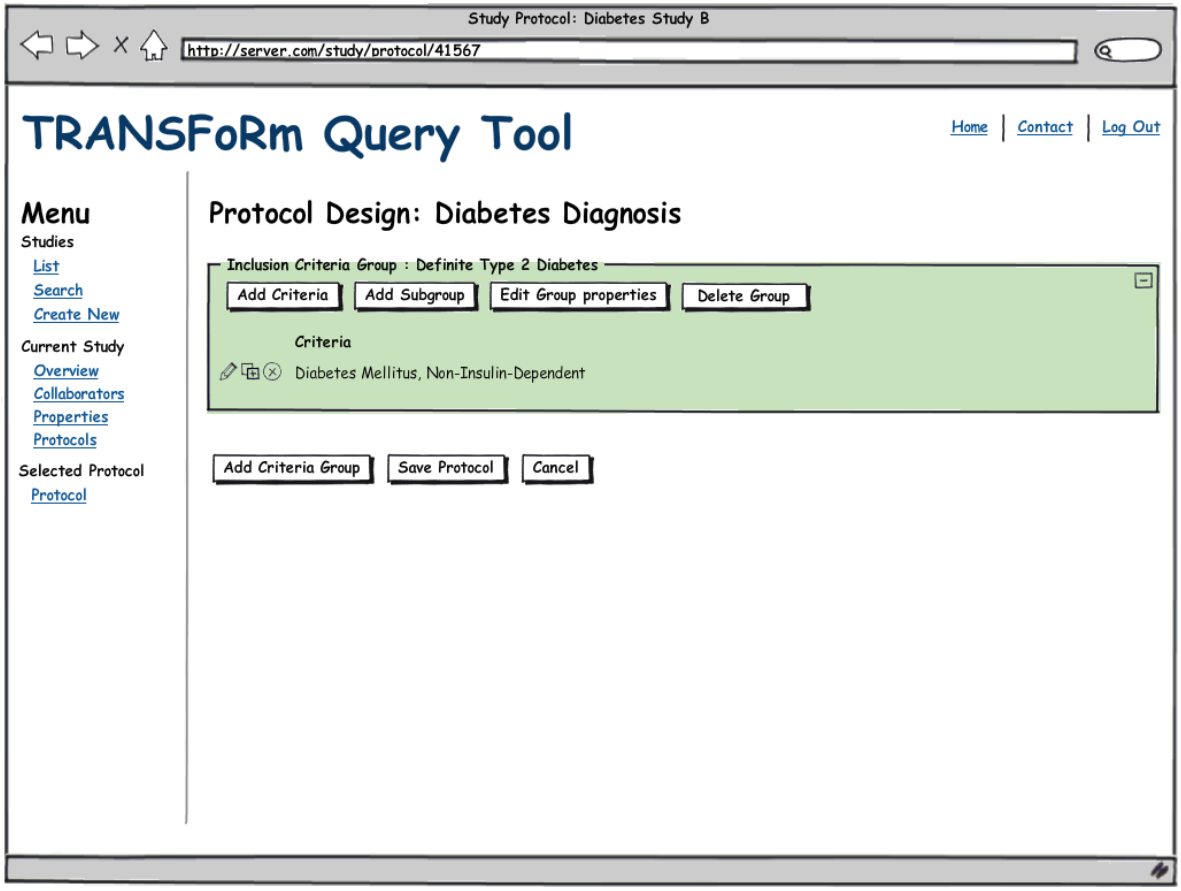

Figure 48: UI 20 Inclusion Criteria Group Added

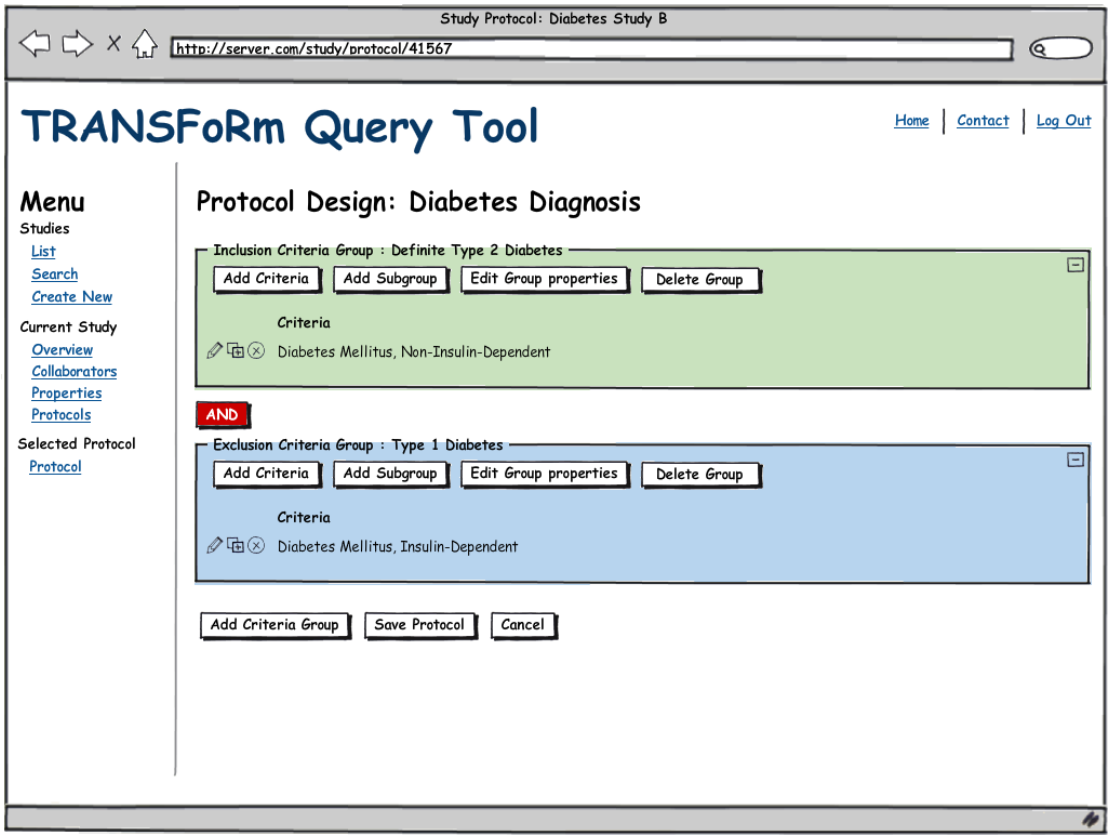

Figure 49: UI 21 Exclusion Criteria Group Added

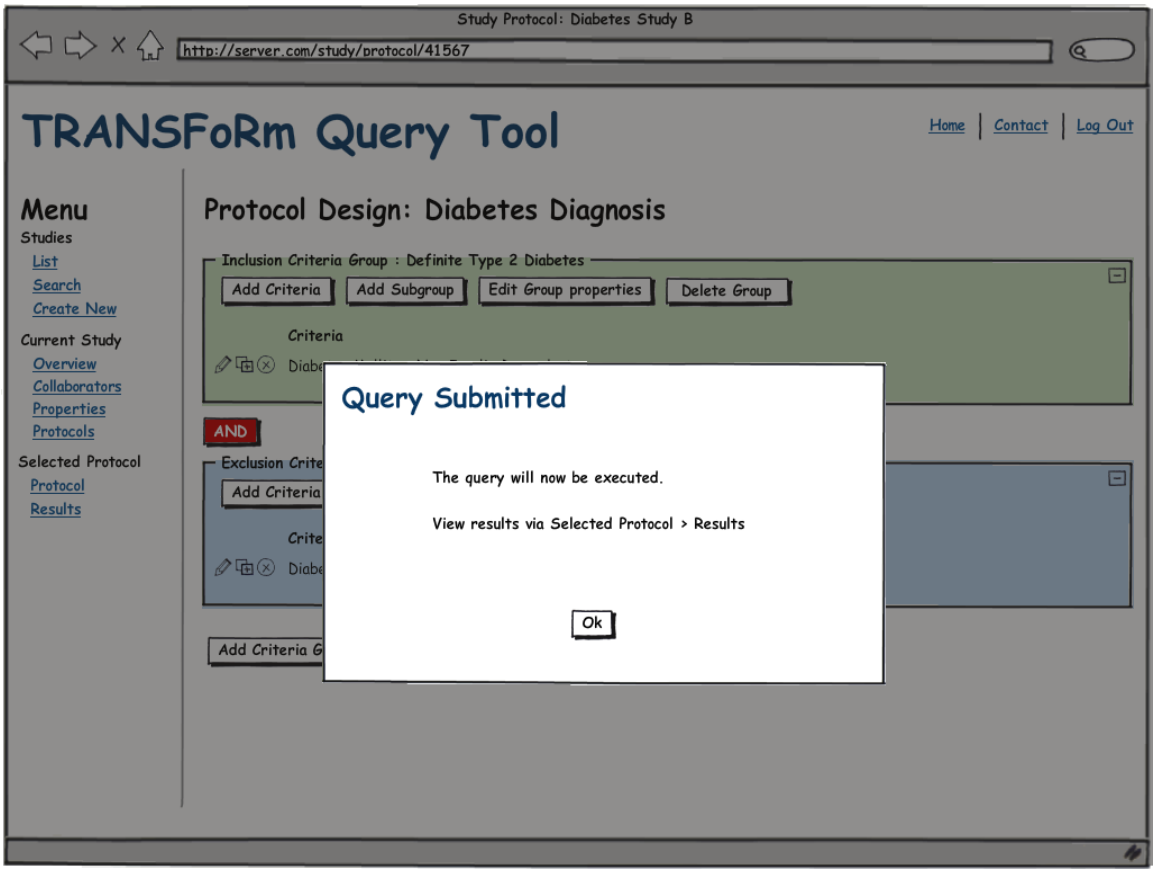

Figure 50: UI 22 Query Submitted

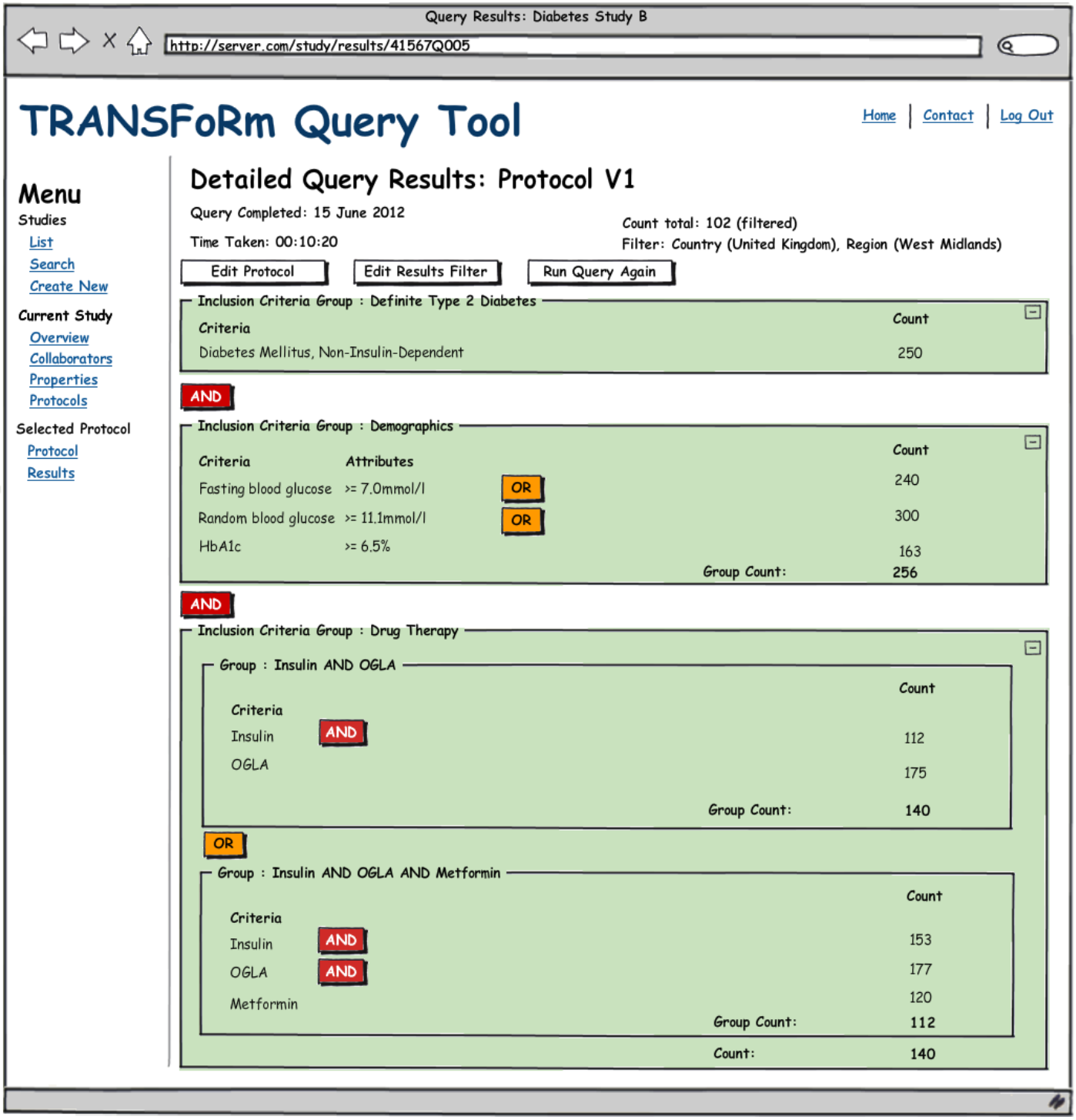

Figure 51: UI 23 Query Execution in Progress

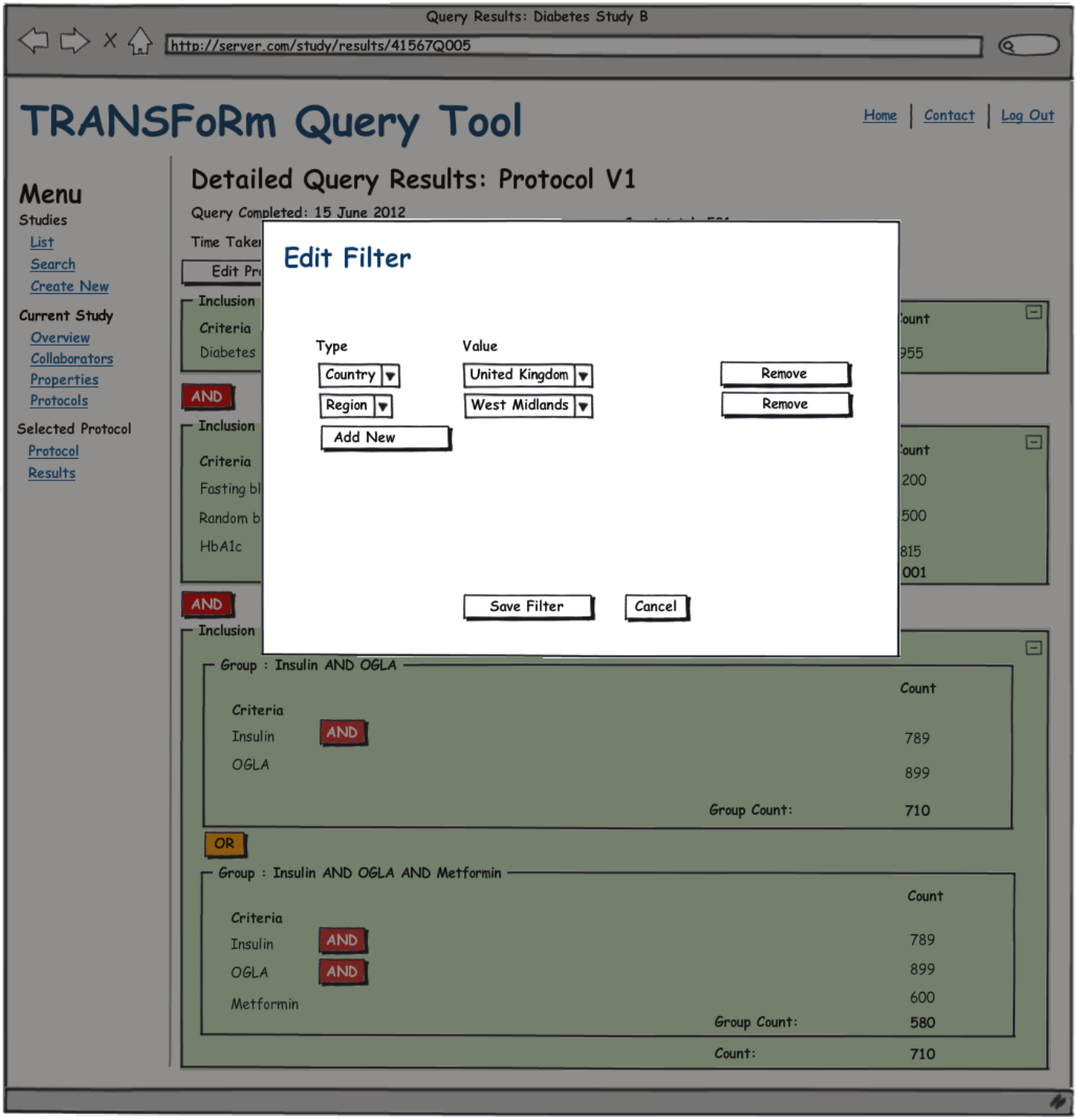

Figure 52: UI 24 Query Results - Edit Filter

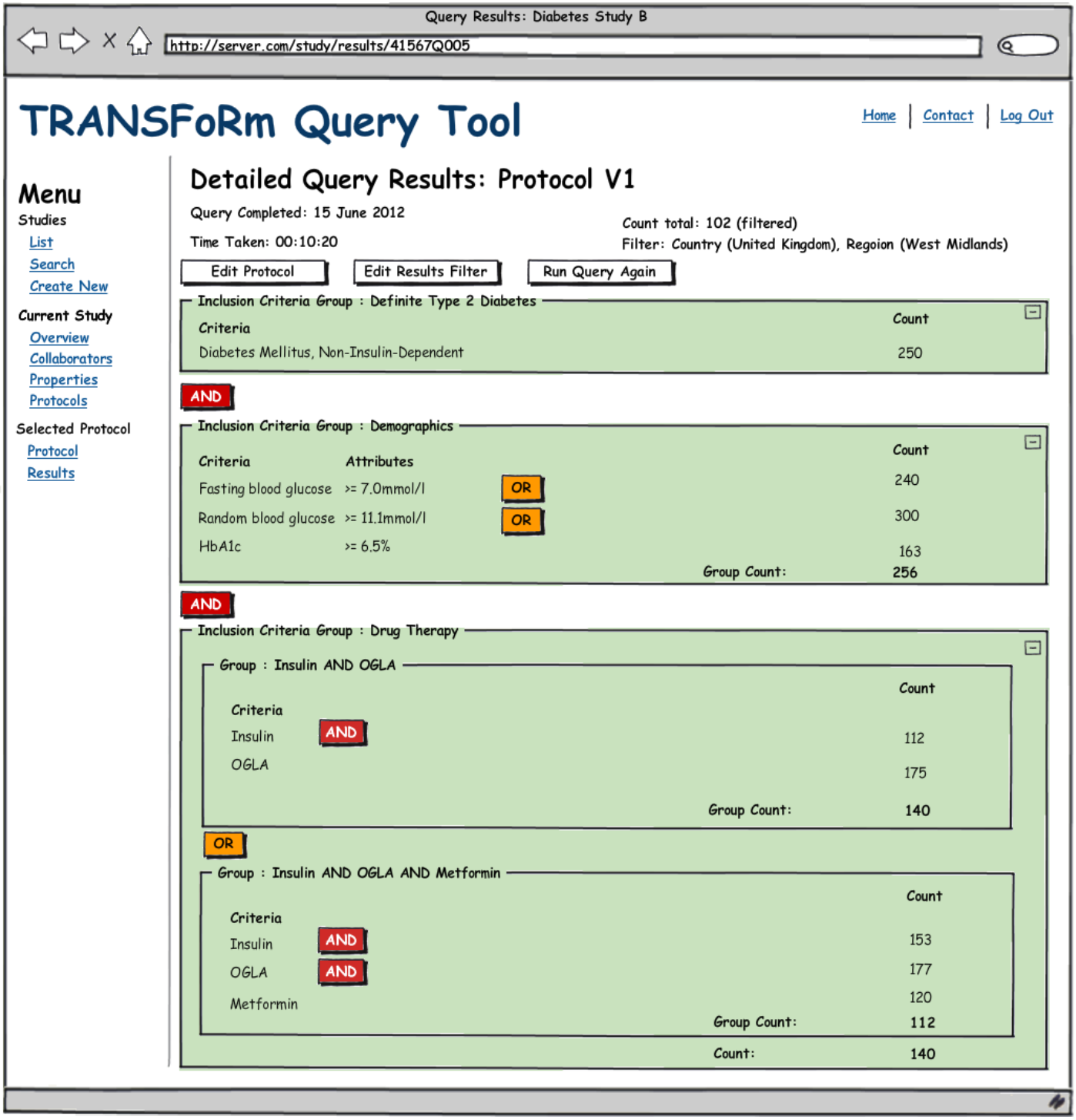

Figure 53: UI 25 Query Results Filtered

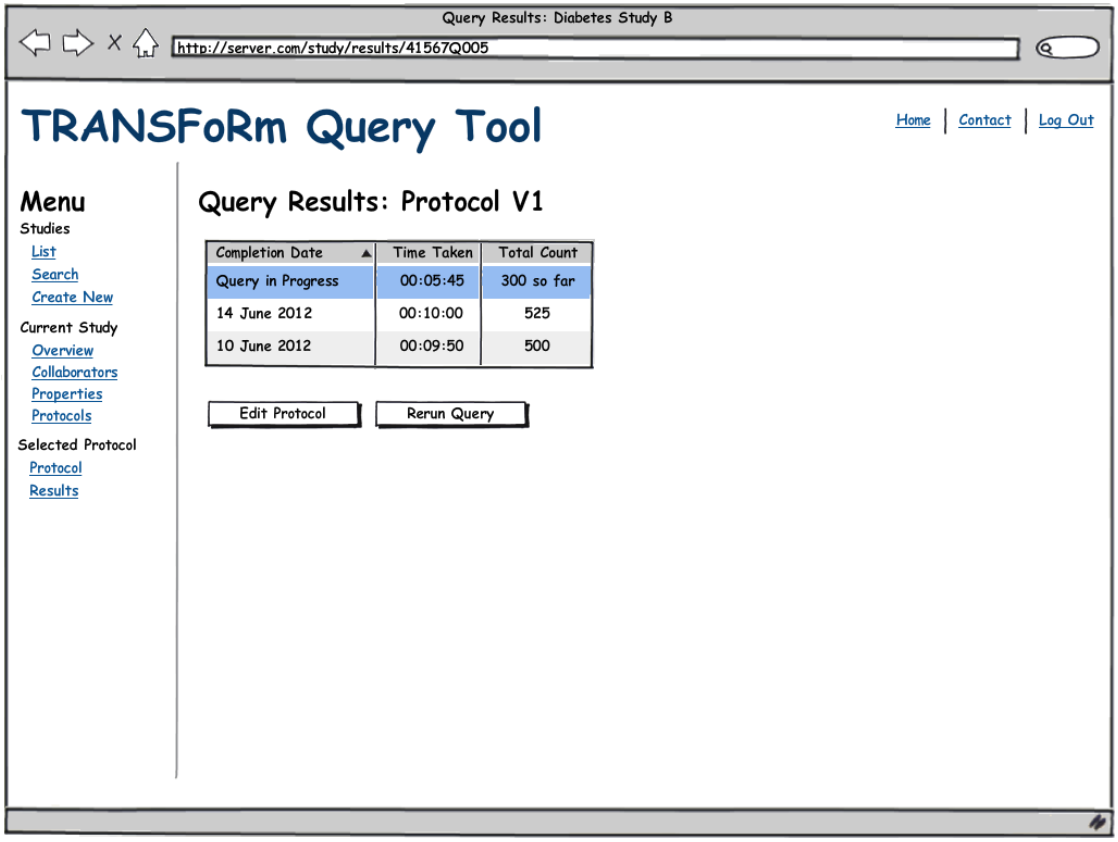

Figure 54: Query Results With Incomplete Query

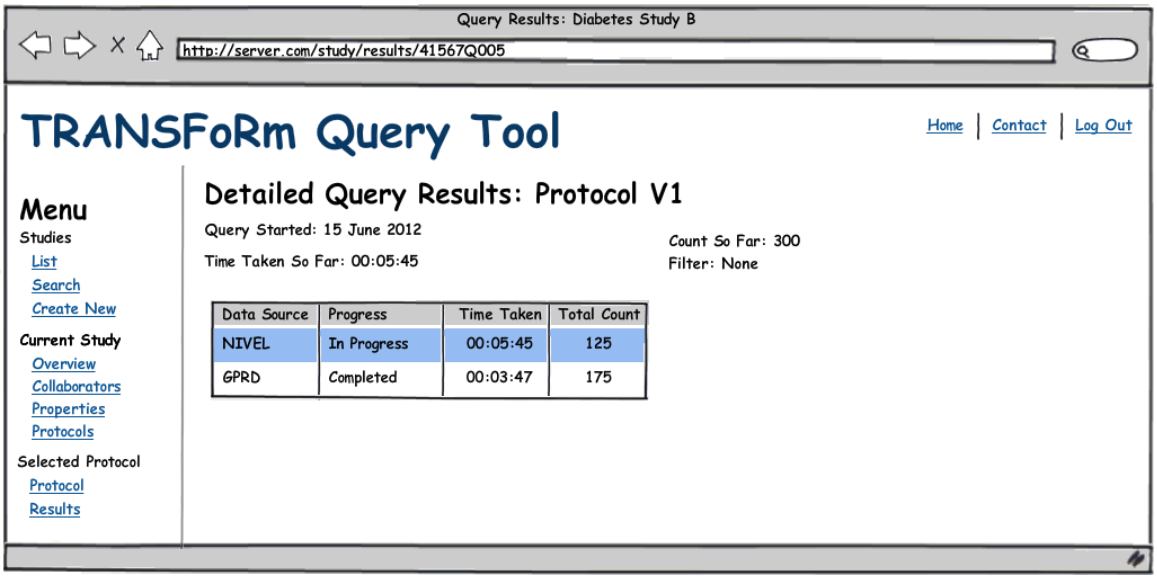

Figure 55: Incomplete Query Results

Appendix B

Several example user interface instances for the query formulation workbench have been developed by using simple HTML and the jQuery framework. These medium fidelity mock ups support some user interaction and are useful in understanding user requirements for rich interfaces. Although they are running on a web server, the underlying web application necessary for data processing and integration with other software such as the TRANSFoRm distributed infrastructure is not present. The supporting code for the interface is developed entirely client side.

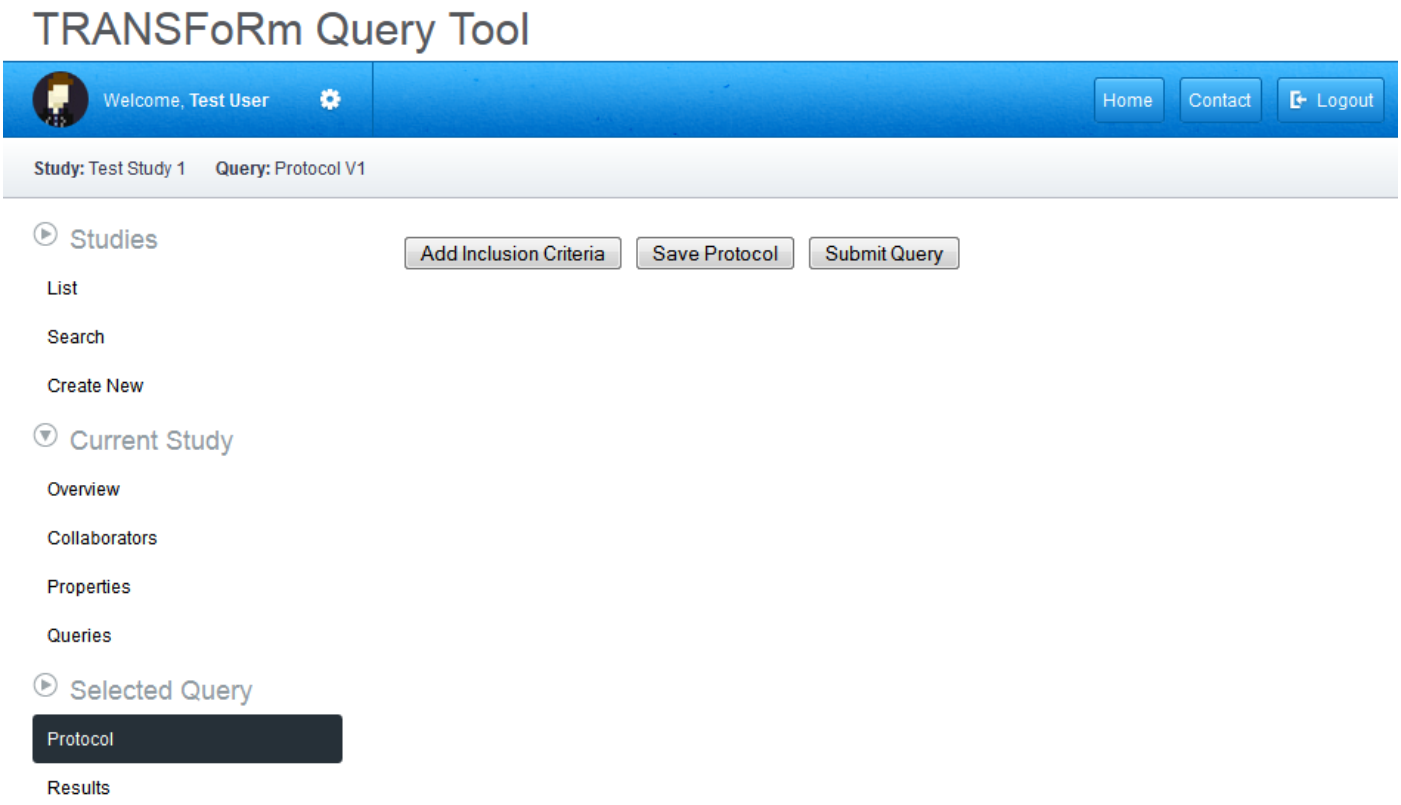

Figure 56: Medium fidelity mock up - initial loading of the query tool

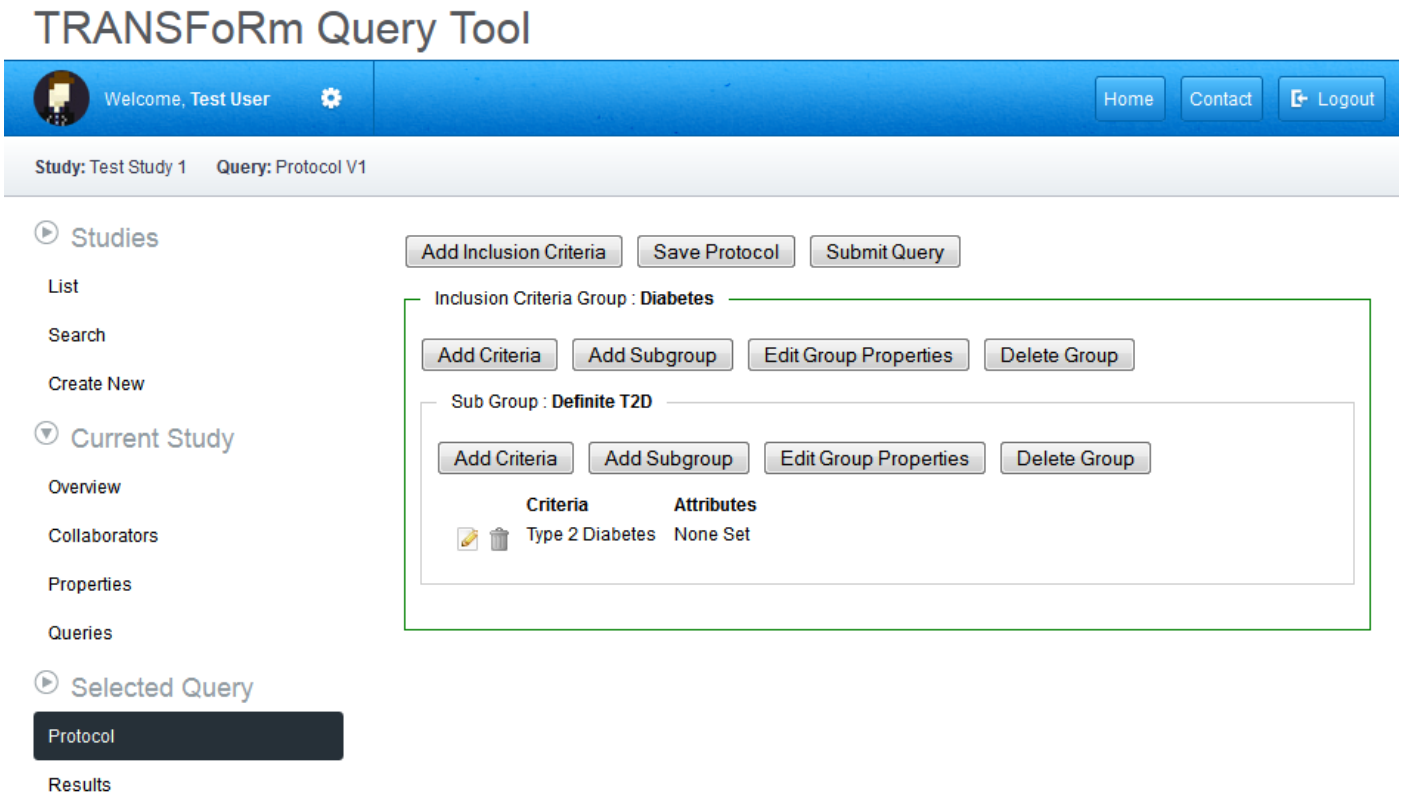

Figure 57: Medium fidelity mock up – after addition of an inclusion criteria group with nested sub group and criterion.

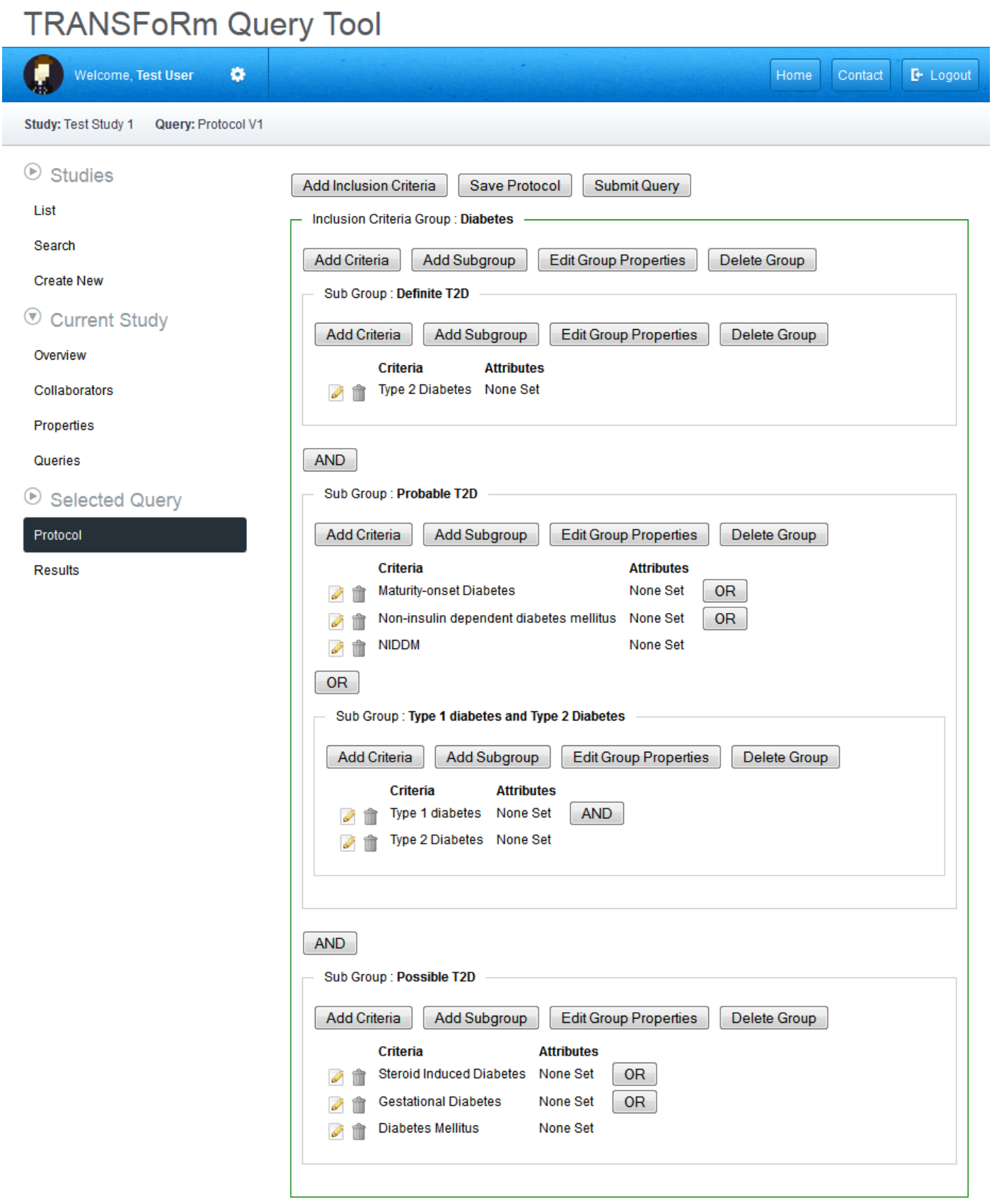

Figure 58: Medium fidelity mock up – after creation of an inclusion criteria group containing more complex query criteria

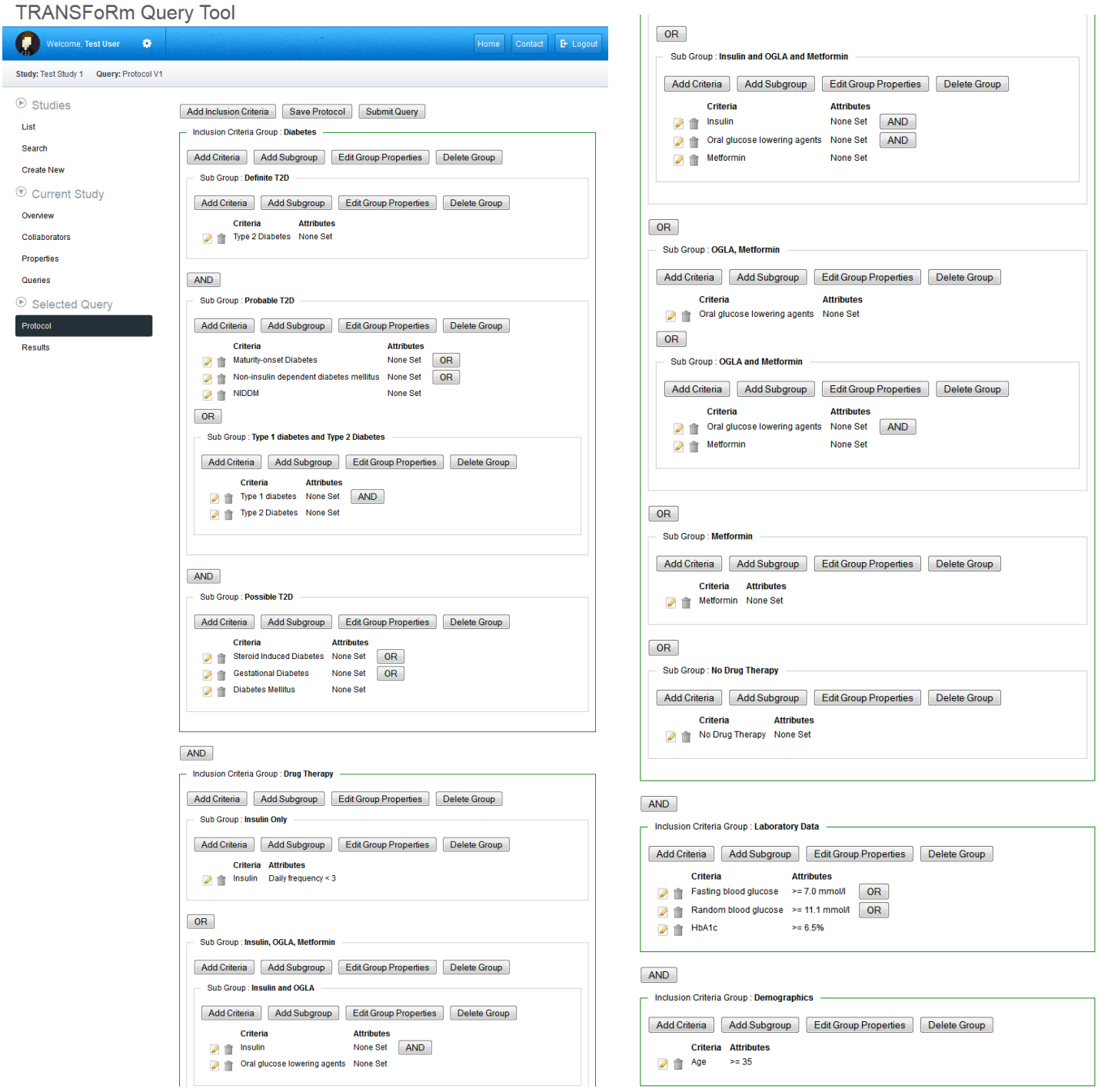

Figure 59: Medium fidelity mock up – completed query design based on diabetes use case.