A tool to integrate multiple clinical trials

A Capstone Project

Presented to

Department of Medical Informatics & Clinical Epidemiology

And

Oregon Health & Science University School of Medicine

In partial fulfillment of the requirements for the degree of

Masters of Biomedical Informatics

December, 2010

School of Medicine

Oregon Health & Science University

Master of Biomedical Informatics

CERTIFICATE OF APPROVAL

This is to certify that the Capstone Project of

Vandana Kapoor

"A tool to integrate multiple clinical trials"

Has been approved

Judith R. Logan, MD, MS

Date

Acknowledgements

I would like to place on record my deep appreciation and thanks for my project supervisor Dr. Judith Logan, who provided me valuable insight, guidance and inspiration during my graduate studies. It is a pleasure for me to extend my thanks also to Dr. Lois Declambre (Department of Computer Science, Portland State University) for guiding me in the project from the perspective of computer science. I would also like to thank Ms. Diane Doctor for her support in navigating the requirements for the curriculum.

Table of contents

Table of Figures

Table of Tables

ABSTRACT

A clinical trial is a research method for studying the risks and efficiencies of the medical products or procedures under analysis. The integration of data from multiple clinical trials can provide much-needed help in answering research questions, since the combined enrollment from the individual trials increases the significance of the research results.

Previous research by Logan et. al. [(1), (2), (3)] analyzed the usage, advantages, and effectiveness of what they call "dynamic data integration" in the context of multiple disparate electronic medical record databases. Both the theoretical framework and tools used in this project were developed as part of that work. This model of data integration presents many advantages for a data analysts. For example, it includes methods such that the query interface used by data analysts is essentially the user interface through which data was collected. This method provides maximal contextual information about that data that can improve query accuracy. In addition, by using these methods and tools, data analysts do not need to be cognizant of the underlying data schemas or to be database experts. Our current work extends the previous research by evaluating and enhancing the efficacy of the existing tools in the setting of clinical trials.

As a first step for enhancing the existing tools, we searched for data transformation operators used with clinical trials data. To do this, we repurposed results from a previously published study (*Cognitive load required for completion of case report forms* (4)) performed by the Duke Translational and Research Institute. These data transformation operators were then implemented in the query interface. Additionally, we created a new prototype user interface, to showcase the tools to different analysts for evaluating effectiveness in clinical trials data integration. We also created tools to assist

in pre-integration steps including a tool to transform data in REDCap (5) to a known format (g-tree (1)), so that the existing tools can directly consume the data, without any changes to the framework or existing tools.

This project demonstrates the use of the aforementioned enhancements, by integrating the data captured using two different systems (REDCap (5) and Cardiff TeleForm (6)) during the performance of a clinical trial. Integration of this data simulates integration of multiple clinical trials, since it presents similar complexity due to varied database and coding schemas.

1 INTRODUCTION

Analysis of integrated data from multiple clinical trials is important in obtaining statistically relevant conclusions. The main advantage of integrating clinical trials is to answer new questions or gain further confidence in a product or a health-care decision. Combining multiple clinical trials data also makes it possible for researchers to correlate various features of any clinical product or drug and highlight its positive and adverse effects on a larger demographic distribution. The larger sample set yields better understanding as compared to a single-focused trial.

Integrating clinical trial data from multiple clinical trials currently takes excessive resources at each step of the process. Partly due to this fact, drug companies generally use the information from a single trial or from multiple trials (without integration) to achieve quicker FDA approval (7).

One of the main reasons that the data integration problem is so complex is the lack of a standard format for case report forms (CRFs). CRFs are used to collect data during clinical trial procedures, and the lack of standard CRFs across clinical trials can result in disparate data schemas. Merging the data from disparate data schemas in an unambiguous fashion requires significant resources from investigators/ data analysts, since they are the subject matter experts who can resolve most ambiguities. Instead of focusing on the analysis of results, these investigators must spend time in understanding the underlying data or database schema and crafting often complex queries to obtain their data. Hughes (7) precisely lays out the problem, stating, "*Integrating data from multiple*

trials can be inordinately expensive. But eventually, it will provide a solid background for any research by reducing population in the trials as well as reducing calendar time."

The following example highlights the difference in the classification of data in a case of a data element known as "Presence of pain" and illustrates how a disparate data schema is generated. The values for this data element may vary from source to source. A data source 'A' with two values of pain, "*pain present"* and "*pain absent,"* cannot be fully integrated with a data source 'B' that has three related pain values, "*absent," "mild,"* and *"severe,"* without making any classification decision or without defining the different possibilities of data integration. Additionally, for creating a new study schema 'C', where an investigator needs four categories of pain ("*absent*," "*mild*," "*moderate*," and "*severe*") further granularity needs to be defined by the investigator during integration. As it is evident from this example, classification of such data elements unambiguously is challenging and may not be possible for all studies.

Different clinical data integration techniques have been proposed in the past that attempt to tackle the challenges discussed so far. For example, consider the CLIO Project (8), in which Miller et. al. developed a system that can deal with most of the complex tasks of data integration and transformation from multiple clinical data sources. Their approach is based on graphical user interface (GUI) mapping of data elements. In particular, the users can map various data elements using the drawing tool. However, their solution does not offer a solution for the aforementioned classification problem.

Similarly, at the University of Oxford, Calinescu et. al. (9) created a solution for sharing and integrating data between clinicians and researchers, allowing them to query multiple clinical trials by using a controlled vocabulary and common data elements

(CDEs). Their web-based interface, known as CancerGrid, allows researchers with little technical knowledge to query common data from different clinical trials but which use the same CDE repository. Their project is limited to cancer trials and its use requires a common language to design case report forms. It allows complex queries involving CDEs, with a simple way to access data across multiple trials. However, their project does not offer a solution for integrating multiple clinical trials that do not use the CDEs, so it cannot be generically used for multiple clinical studies that have heterogeneous data schema.

Logan et. al. [(1), (3), (2)] previously created a framework for tackling data integration problems. The approach is based on three principles that help to address various challenges posed during data integration.

1. The investigator/data analyst should not have to understand the database schema in which data is stored, since this is often a schema that benefits data collection and storage but not usability. Investigators should not have to be database experts.
2. The context in which research data is collected is important for its interpretation in studies. A significant amount of that contextual information is captured in the user interface through which the data is collected, and should be available to the investigators/data analysts asking questions of the data. An example of contextual information would be the type of field used (e.g. checkbox vs. radioarray), options for data values (e.g. the list of options in a menu), and which data values trigger other data collection (e.g. checking a field opens a sub-screen with other data

fields). The investigator should have access to a query interface that resembles the data collection user interface.

3. Data integration should be performed dynamically, not statically. When data integration occurs, decisions must be made about transformations to data so that all source data conforms to the same schema. When data is modified statically to fit the integrated schema, information may be lost. For a specific study, however, the data integration decisions might be different from the decision made at the time of static integration. Investigators should be able to make the decisions about how data from disparate sources is integrated with the data from each study performed.

These principles make the existing framework very flexible and form the foundation for our research in the area of multiple clinical trial data integration. For this project, modifications to the previously created tools and interfaces were proposed and evaluated with the objective of refining and evaluating the applicability of the framework and tools developed for clinical trial data integration.

2 BACKGROUND

2.1 Clinical Trials

Clinical trials are designed to intervene in and evaluate any health-related outcome. The objectives of any a clinical trial may include disease diagnosis, disease treatment, and disease prevention, as well as assessing other environmental factors that might affect health. Clinical trials answer specific questions about all these objectives and are also used for interventions in health sciences. Also, they help researchers in determining the effectiveness of any new treatment in comparison to previously established treatments and regimes. Thus, the outcome of these trials is an important factor in the progress of pharmaceutical companies. In most cases, clinical trials are very explicit about the questions that need to be answered and typically focus on a single characteristic of the product requirement.

> The National Institutes of Health (NIH) provides the following definition: (10)
>
> A clinical trial is a prospective biomedical or behavioral research study of human subjects that is designed to answer specific questions about biomedical or behavioral interventions (drugs, treatments, devices, or new ways of using known drugs, treatments, or devices). Clinical trials are used to determine whether new biomedical or behavioral interventions are safe, efficacious, and effective.

According to NIH, human subject research (11) involving an intervention to modify behavior (diet, physical activity, cognitive therapy, etc.) fits this definition of a clinical trial. Human subject research for the development or evaluation of clinical laboratory tests (e.g., imaging or molecular diagnostic tests) might be considered a clinical trial if

the test is to be used for medical decision making for the subject or if the test presents more than minimal risk for the subject.

2.2 Case Report Forms

All patient data are collected in forms for further review and analysis. Known as case report forms (12) (CRFs), these forms capture all the information related to the clinical trial in a specific format. CRFs are used to avoid duplication of data and cover all the questions that need to be answered for a trial. They are designed by study sponsors. The clinical trial protocol determines the type of data collected. A clinical protocol (13), defined as a document that explains the design of any study, covers all the questions that need to be answered by any particular trial. A clinical trial protocol also defines the background design and questions to develop for any given study, and is used by study sponsors and all the researchers involved directly in the study. CRFs also provide instructions so as to reduce any misinterpretation. Traditionally, these have been paper forms that are filled manually by investigators who collect and transcribe data from different sources. The data is then manually transcribed into an electronic database. These electronic databases are typically developed independently for each study, which can make it difficult to integrate and combine the studies for further analysis. But with the advent of the Internet and electronic medical records (EMRs), electronic versions of CRFs can reduce human error and also reduce the data entry time.

The electronic CRFs make it easier to track the progress of the clinical trial and can allow real time access to project data. The information collected in these forms must represent all the criteria in the clinical trial protocol. The ideal and standard CRF can

reduce cost and time for any trial and is also reusable for any future research. The data that are obtained by the sponsor are typically (or often but not always) de-identified by the investigators by removing patient name and other identifiers; a unique study ID is then given to each patient. To ensure the quality of the data, all the CRFs should be verified by some kind of automated method. These methods can highlight any incorrect information in the forms such as "Positive pregnancy with a gender as Male." These highlighted terms are then manually verified by study administrators.

Currently, there are no mandatory standards that can be used to design study protocols and CRFs. The Cancer Biomedical Informatics Grid (caBIG (14)), under the standardized case report work group, is focused on building a standard for designing CRFs to capture Phase II and Phase III cancer clinical data. The main advantage of building a core library for the design of standard CRFs is that it can provide a standard to compare and aggregate multiple cancer trials; however, as of now it does not provide any automated way to integrate cancer trials from multiple data sources.

2.3 Clinical Data Interchange Standards and Implications for Data Integration

Similar to caBIG, CDISC (Clinical Data Interchange Consortium (15)) builds some standards for clinical trial data sharing. These standards can provide a well-defined description of data—for example, their operational data model (ODM) includes CRF data, lab data, and events data. CDISC creates standards to describe study metadata, administration data, and clinical data from the recruited patients. The use of these standards can allow clinical research organizations (CROs) and vendor (pharmaceuticals)

applications to be shared in and out of enterprise or for CROs with Clinical Trial Management Systems (CTMS). However, still there is no automated tool that can integrate multiple clinical studies depending on the need of investigator.

These standards are also likely to reduce the risk of losing data at the time of integration and sharing data, thus providing a stable data exchange tool. ODM is in a form of XML for sharing and exchanging data efficiently between different applications. As these data follow a standard XML format, it would be easier to integrate multiple studies if all of them follow a common standard of CDISC.

2.4 Schema integration and role of operators

Database schema can be defined as a method of representing data elements or attributes in tables that also explain the relationships among the attributes. In addition, the schema also highlights the set of integrity constraints imposed on the database (16). The schema of a database is generally stored in a data dictionary of the particular database. Data dictionaries do not contain actual data of the database, but they include all the information about the tables, field by field. 'Data dictionary' is defined as a repository of information about a database that documents its data elements.

Different data base administrators usually derive different schemas for addressing similar problems, mostly driven by the users of the database. Even within the same organization, integration of the database schemas becomes a challenging problem to solve. The schema integration problem is well studied and documented. In 1984, Batini et. al. (17) presented a methodology for database schema integration in an entity relationship model. They divided the schema integration into three steps: *conflict*

analysis, *merging,* and *enrichment and restructuring.* Following this research, in 1986 Batini et. al. (18) wrote a very detailed survey paper comparing methodologies for database schema integration and concluded that more work is required on various directions including schema mapping.

The data integration problem can be broadly categorized into "static data integration" and "dynamic data integration". The basic premise of static data integration is that data is extracted, transformed and loaded into a single queriable schema. On the other hand, the dynamic data integration can be described as a virtual integration (19) where the actual data is present in the original database source, while the query executed against the virtual schema is translated back as multiple queries to the original sources. The result of these queries is then combined together to form an answer for the requestor. This form of virtual schema is also called 'mediated schema.'

The role of data operators has been important in various steps involved in integrating a data schema. A data operator is a symbol used for specifying an action to perform on one or more expressions. There are different types of operators that are used for data transformation. Some of the examples of operator types are comparison, arithmetic, assignment and logical operators. Several new operators for clinical data integration are defined in this project.

2.5 Overview of data integration terminologies

Researchers and investigators use various terms in integration of clinical trials data research. Often times these terms overlap and cause confusion about the activity performed in the research. It is essential to understand and distinguish these terms in

order to clearly identify the work in a particular project and to minimize the confusion within the research community. The next few sections explain the meanings and significance of different terms used in clinical research.

2.5.1 Clinical Trials Comparisons

Clinical trial comparison can be defined as a study in which the results and conclusions from one trial are compared with the results and conclusions of other trials. The results from these trials are compared with each other to determine the level of performance for a drug or procedure. The best example is the comparison of the efficiency and adverse effects of two different drugs for treatment of the same disease. Meta-analysis is defined as the combination of results from several studies or experiments that bear on similar questions or hypotheses. The meta-analysis involves integration of conclusions and results from multiple trials and does not combine raw data from various trials for a potential new study. For example, In the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET) (20) study, Hawkey et. al. found that patients who were not taking aspirin had a significant reduction in all ulcers and complications with Lumiracoxib. Effect took much less time (in days) compared to nonselective nonsteroidal anti-inflammatory drugs (NSAIDS), naproxen and ibuprofen. In this study, there was no data integration; the method involved only a comparison of the conclusions of the trials. The main aim of the current project differs from this method as our project is not related to meta-analysis. Our aim is to provide a tool for dynamic integration of clinical trials from multiple data sources based on investigator requirements.

2.5.2 Multi-site Data Integration

Multi-site data integration can be defined as integration of data collected from multiple sites (research centers in different regions or different hospitals) for a single trial with a common protocol. This type of integration involves combining data that is collected in the same format (same CRFs) and having the same trial protocol. These data may have similar schemas for which data integration is less of a challenge.

Drozd et. al. (21) developed a relational XML schema to integrate real-time patient data from electronic medical records at six centers for Acute Immunodeficiency Syndrome (AIDS) research (CFAR). This system facilitates the use of multi-site data for clinical research to improve the treatment of AIDS. The CFAR Network of Integrated Clinical Systems (CNICS) is the first electronic records-based network capable of integrating data from multiple sites from the large HIV-infected population. Mugavero et. al. (22) performed a retrospective analysis of HIV-infected patients through this platform to determine the racial disparities in HIV virology failure. This study used the same data schema and standardized CRFs throughout CFAR network.

Another large-scale multi-site data integration example is that of the North American AIDS Cohorts Collaboration on Research and Design (NA-ACCORD) (23). These investigators used international epidemiological database for the evaluation of an AIDS project that integrated AIDS data from U.S. and Canadian hospitals. This collaboration consisted of 22 research groups representing more than 60 sites. Again in this case, the data schema is the same, so it is relatively easy to conduct the data integration.

2.5.3 Multi-source Data Collection

Multi- source data integration is the integration of data from multiple sources usually within the same research center to fulfill the requirements of a single clinical trial protocol. Figure 1 explains the integration of data from four different sources within a research center for a single clinical trial. The data integration in this system is not manually driven and can be done using a simple data integration tool. The data from these multiple sources are generally linked with each other by an identifier that allows the integration of these data. Some research centers and hospitals build a central repository to store the integrated data, and this repository can easily be accessed by multiple users for further research and analysis.

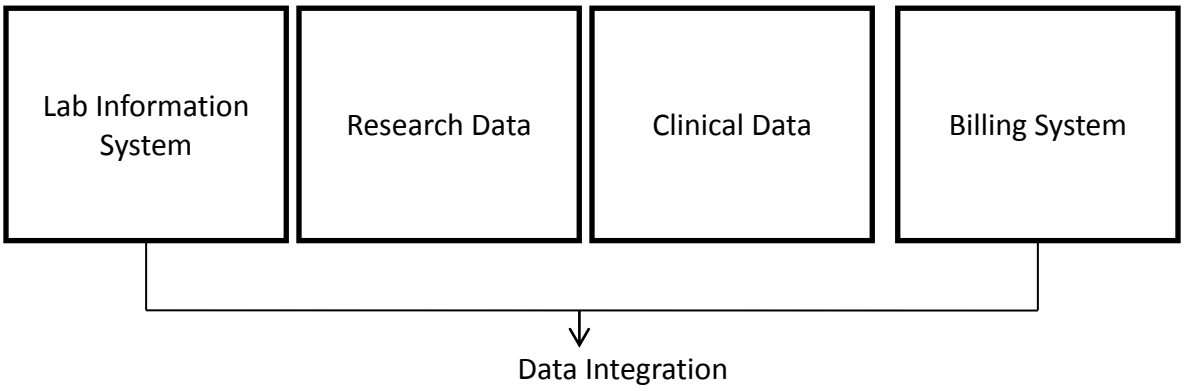


Figure 1: Multi-Source Data Integration

To exemplify this type of data integration, the Andrology Research Information System (ARIS) (24) is the best fit. Timers et. al. developed and implemented an information system to integrate data from the multiple sources of the hospital into a well-structured database for clinical research. ARIS integrated the data from Health Information System (HIS) data (including patient history, diagnosis and therapy) from clinical laboratory and

data collected from the Department of Andrology. This centralized system allows researchers to find and analyze all the related data for any patient in order to build clinical trials. This system does not, however, allow integration of multiple trial data; instead, it allows integration of patient data to build clinical trials.

2.5.4 Crossover Trial Integration

Crossover trial integration (25) can be defined as integration of data from a single trial in which data is collected more than once under different circumstances, for example, in a trial where subjects receive more than one treatment. These types of clinical trials provide statistically and clinically relevant results using only a small population.

This method of integration involves integration of data for a single clinical trial at different points in time. Therefore, this trial has the same clinical protocols and CRFs, and the data have similar underlying schema, therefore facilitating the use of common data integration tools. Our tool for clinical trials integration is different in the sense that it promotes integration of data from multiple sources with different data schema.

2.5.5 Clinical Trials Integration

Clinical trials integration is the integration of data from more than one clinical trial where the underlying data schemas are different. Figure 2 shows this method of integration comprised of clinical trial data from multiple trials including different CRFs and trial protocols. As the trial protocols are different, this type of data integration can be labor intensive, requiring data transformation and mapping elements from the different trial sources.

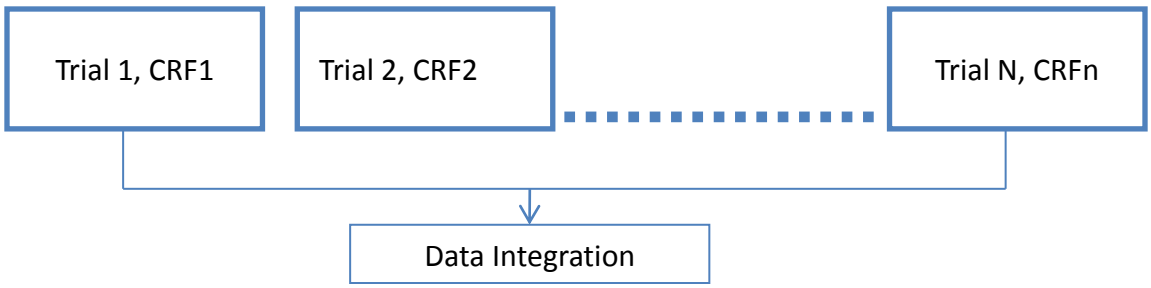


Figure 2: Clinical Trial Integration

3 ARCHITECTURAL OVERVIEW OF CURRENT DATA INTEGRATION TOOLS

This section presents an overview of the existing data integration tools, which were used and extended in our current work. These data integration tools were developed by the research team headed by Drs. Logan (OHSU (26)) and Declambre (PSU (27)). These tools provide users with a query interface through which various queries can be built, including queries which integrate data from multiple data sources.

Figure 3 and Figure 4 (1) provide a general comparison between the traditional approach to querying data and the current approach as explained by Terwilliger et. al. Traditionally a user inputs data using a forms-based user interface (Figure 3) and the data is stored in the database, which an analyst must query using a query language such as SQL. The current approach (Figure 4) improves on the traditional approach by building a "natural schema" for the data that can be connected to the query interface. The natural schema is a format that mimics the user interface (UI) through which the data was collected. An associated structure, the "g-tree", provides the contextual information from that interface and a "channel" provides the connection to the database. The data analyst now can use a graphical interface to query what appears to be the data collection UI or similar representation of the data. The analyst has no need to know either the database schema or a database query language.

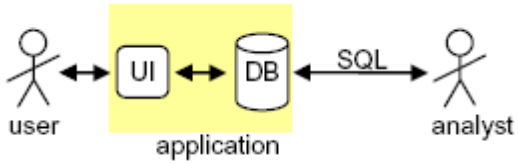


Figure 3: Traditional Approach Of Extracting Information By Writing Queries Against Physical Database (1). UI = user interface, DB = database, SQL = query written in a database languages such as SQL.

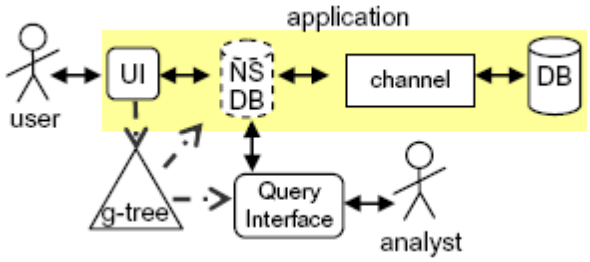


Figure 4: Approach Used By This Tool With Development Of Query Interface To Extract Data From Physical Database(1). NSDB = the "natural schema" of the data, DB = database, UI = user interface

In the paper "Querying through user interface," Terwilliger et. al. (Terwilliger, Delcambre, & Logan, *Querying through a user interface*, 2007) described and developed three artifacts of this platform as follows:

- *GUAVA* (Graphical User interface As View) *trees*, also known as "g-trees," represent the UIs with all associated contextual information in XML format.
- *Study schemas* contain information about the data elements and their nature that a data analyst wants in the new studies.
- *Classifiers* generate a relationship between the selected data elements of g-trees with the study schema.

Using these artifacts, a database query can be generated automatically from the query interface. Terwilliger also described the use of forms-based UIs to generate a conceptual model that represents the information in the query interface. In this case, a g-tree represents the information present on a forms-based UI along with the relationships between the various forms. The user can express queries using a query interface against this conceptual model. The context elements for the controls, such as the control's type

(e.g., textbox, checkbox, or drop-down list), its default value, and its text are also represented in the g-tree. All these context values are especially important for an analyst using the query interface while integrating data from multiple sources. The data integration tools are broadly divided into two categories, the “database administrator dashboard” and “analyst dashboard”. These tools are further explained in the next sections.

3.1 Database Administrator Dashboard (DBA dashboard)

DBA dashboard was developed for database experts who will create a channel for a dataset, bridging the database and the natural schema. This dashboard allows a database expert to easily combine several computational functions for any set of data. This process should be performed only once for any dataset and can then be reused any time the data is used for a new study. The DBA dashboard consists of three main components, the channel building, the GUAVA server, and the annotator.

3.1.1 Channel Builder

This component allows a database expert to transform the data from the database schema in which the data was collected to a *natural schema* that investigators can easily understand or, conversely, from the natural schema to any arbitrary database schema where the actual data resides. The channel acts as a database view in that the data always remains in the storage schema, but queries are issued to the view. A *channel* transforms all the queries from the UI or query interface using the natural schema to the physical database at run-time. This component also allows updates to the underlying data irrespective of complexity of data transformations. The DBA builds the channel using

database transformation operators, which can be "stacked" as needed for the appropriate data transformation. The DBA dashboard provides a tool that lets the DBA perform this task.

3.1.2 GUAVA (GUI as a view) Server

The channel and g-tree for any study are combined using the GUAVA server, which acts as a linkage between the study database and the investigator's query interface. The GUAVA server is software that must be available to the data analysts and acts as if it is the datasource.

3.1.3 G-tree Annotator

The g-tree annotator is required when the UI is not built *de novo* using GUAVA components. The GUAVA components automatically capture contextual information for a g-tree. In a non-*de novo* application, this contextual information may need to be added. The G-tree annotator was developed to help the DBA perform this task once for any dataset.

3.2 Analyst Dashboard

The analyst dashboard is defined as a query interface that allows investigators to query a single dataset or to create a single dataset from multiple datasets using GUAVA servers. Figure 5: Relationship Between Data Sources and Analyst Dashboarddemonstrates the flow of information from the data source to the analyst dashboard.

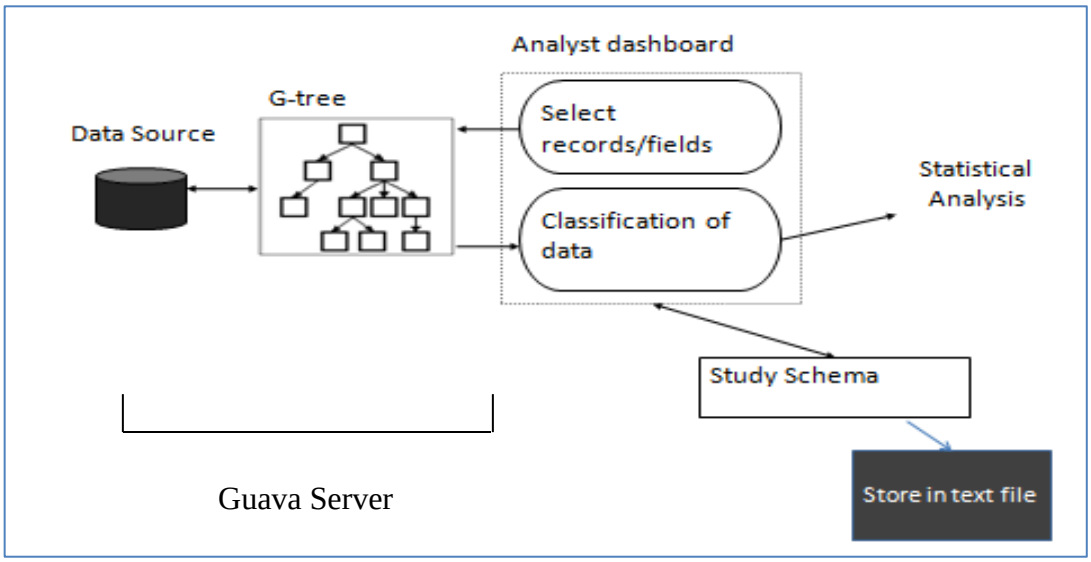


Figure 5: Relationship Between Data Sources and Analyst Dashboard. The analyst dashboard creates or imports a study schema, issues queries against the data source, and outputs a data file for statistical analysis.

To use the analyst dashboard, the data analyst can select the data sources that are to be integrated and define the final output for the new study. The analyst then identifies each data element in the underlying data in its natural and annotated schema, and applies any data transformations that are necessary to build the new data set. There is no need to know either the underlying database schema or a database query language.

On completing the study definition for each selected data source, the decisions can be saved to a text file. For future studies or revisiting same study, this file can be read back into the analyst dashboard and modified as needed. After running the query, the results (the data) can be output in a format appropriate for further statistical analysis.. All of the aforementioned tools are used in the current project and were tested in the setting of clinical trials. This project is designed to determine any new peculiarities in clinical trials data that can inform extensions and improvements on the current tools.

4 METHODS: ADAPTING DATA INTEGRATION TOOLS FOR CLINICAL TRIALS

This section explains the various steps involved in improving, verifying, and testing the existing data integration tools (which are explained in section 3) in the domain of clinical trials data integration. The steps include searching and defining the applicable data transformation operators from a preexisting clinical trial study, searching for a representative clinical study that can measure the applicability and results of the existing tools, preparing g-trees for analyst dashboard querying, unit testing the tools, and demonstrating the success of the analyst dashboard.

4.1 Defining Data Transformation Operators

First, we searched for data operators that were applicable for clinical trial data, in order to test our existing tools in this domain. The data collected by Nahm et. al. (4) at Duke's Translational and Clinical Research Institute in a study titled *Cognitive load required for completion of case report forms* (see abstract) was deemed appropriate for this purpose. We chose this particular study both because it used data from multiple clinical trials and because the researchers found that almost 61% of data elements required data transformation; this made the dataset rich for understanding data transformations. Additionally, the research provided guidance on various data elements that are needed together with data operators to create a study element. As an example, Table 1 shows the multiple dimensions (which can exist in single or multiple sources) that are needed by the study personnel in order to generate a Yes or No answer to the study element, "*Did the patient have an unplanned early discontinuation from Tirofiban therapy*?" We analyzed

the data along with the dimensions to find the applicable operators such as "subtraction" (an arithmetic operator) and "less than" (a comparison operator). Most of the identified operators (excluding the arithmetic operators) were implemented in the data integration tools and later employed in the usability study conducted by our project on a different set of clinical trials data. In total, we identified fifteen such operators that are further discussed in Section 5.1.

Form Element	**List each dimension required to abstract the data element (result set)**	**# dimensions**
Did the patient have an unplanned early discontinuation from tirofiban therapy? (yes/no)	enrollment date, tirofiban planned duration, tirofiban start date, tirofiban stop date, termination reason	6

Table 1: Duke Study Example Result.

4.2 Analysis of OCTRI Study Data Sets

The Oregon Clinical and Translational Research Institute (OCTRI) is a partnership between Oregon Health & Science University and the Kaiser Permanente Center for Health Research, funded by the National Center for Research Resources through the Clinical and Translational Science Awards (CTSA). Five different studies conducted at OCTRI (28) were selected for further review. Table 2 shows the study titles with total number of individually analyzed questionnaire forms.

OCTRI Study ID	Study Titles	Number of Forms

722	Diet and Prostate Cancer Risk	3
814	Catechins and Fatty Acids Impact on Fatty Acid Synthase Activity in the Prostate: A Randomized Controlled Trial	9
1037	Omega-3 Fatty Acids and Prevention of DCIS and/or ADH: A Translational Approach	10
1051	Genetic Susceptibility, Environment and Prostate Cancer Risk	13
10156	Sulforaphane: A Dietary HDAC Inhibitor in DCIS	13

Table 2: Studies Considered for the Present Project

With IRB approval, we selected Study #1037 to use in our project for the usability testing and implementation of the data integration tools. The reason for choosing this particular study was that two different data collection methods were used for the study, which mimics multiple clinical trials data. Study #1037 was conducted by Dr. Jacklien Shannon, Principal Investigator at OCTRI, to determine the relationship between omega-3 fatty acids and prevention of breast cancer. The data from this study was first collected on paper, and then a portion of it was scanned using OCTRI TeleForm services. The remaining data was entered through the OCTRI REDCap system, which differed enough from TeleForm in the database schema and naming conventions to mimic a separate clinical trial. Integrating the data from these sources using manual data mapping and data transformation would be time-intensive. Instead, we used our g-tree-based query-builder tool to integrate these two sets of data into a single dataset. Table 3 shows all the forms used in study #1037. From the forms listed, we decided that the integration of data from just two forms, the "Family History Questionnaire" and the "Risk Factor Questionnaire," would be sufficient for these experiments.

SNo	Forms name
1	Risk factor Questionnaire
2	Family History Questionnaire
3	Diet Changes, supplemental and Herbal
4	Specimen Collection Form
5	Diet history Questionnaire (DHQ) Addendum2
6	DHQ Part I
7	DHQ Part II
8	DHQ Part III
9	DHQ Supplemental Fish and Game addendum
10	Adverse Event Questionnaire

Table 3: Forms in Study #1037 "Omega-3 Fatty Acids And Prevention of DCIS And ADH: A Translational Approach"

4.3 Data Sources and Data Schema

Once the study sources were selected, the next step was to create g-trees for each dataset. The TeleForm data, which was collected on scannable paper forms, was scanned into the electronic TeleForm database; the REDCap data was entered into that system using electronic CRFs. As with TeleForm, these CRFs were designed using the original data dictionary and trial protocol of the study. The data in REDCap was collected into a single lengthy form in which all the forms were merged as opposed to the multiple questionnaires used by TeleForm, which created a difference between the data schemas of TeleForm and REDCap.

As a precursor to data integration, the data from TeleForm was imported directly into a study database. With this database and using the DBA dashboard, the channel was defined and g-tree created for this dataset. Similarly, the REDCap data related to the "Family History Questionnaire" and the "Risk Factor Questionnaire" was exported from

REDCap in the form of comma-separated files, which were then imported into a new database. Using the DBA dashboard again, the channel was defined and a g-tree created for this data. Figure 6 is a screenshot of the current channel builder tool within the DBA dashboard.

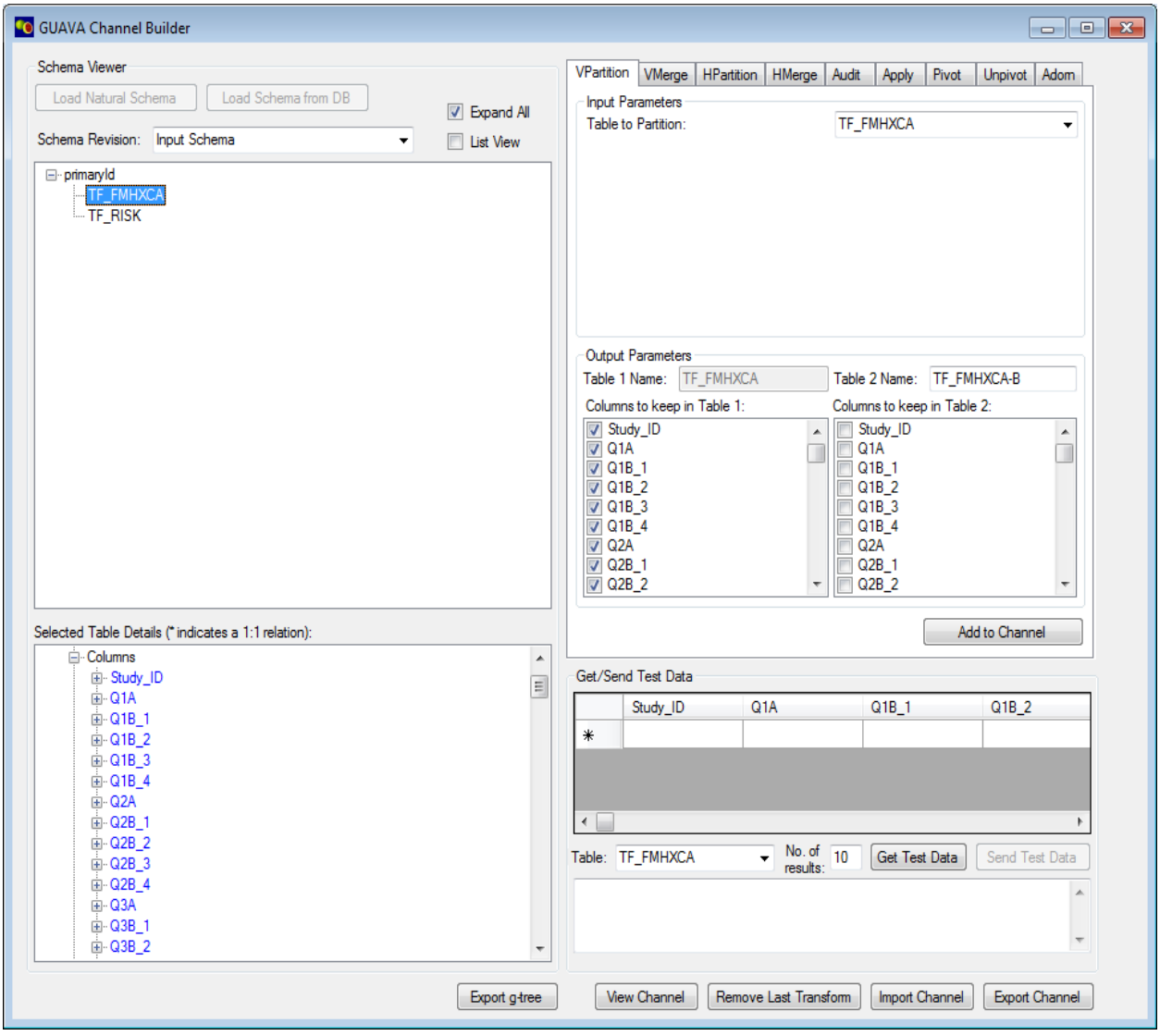


Figure 6: Screenshot of the Channel Builder Tool Within the DBA Dashboard.

4.4 Annotation of G-trees

Before experimenting with the selected data in the analyst dashboard, it was necessary to build a g-tree for the TeleForm data from the source data using the channel builder tool. The g-tree we created contains information about the database schema including all the tables in the database but it does not include many of the contextual elements from the UI. For the purpose of this project, the g-tree with database schema of "Family history" and "Risk factors" questionnaires were annotated manually in order to add all the contextual information. In the future, the DBA dashboard will need to be extended to allow a more automated method of annotating the g-tree, which is an XML representation with all the contextual values present in the questionnaires.

The following figures show the difference between the "raw" g-tree, created using the channel builder in the DBA dashboard, and the manually annotated g-tree. Figure 7 demonstrates the g-tree generated by channel builder with empty text values. Figure 8 demonstrates the manually annotated g-tree with all the contextual information and classification of all the attributes with enumerated values in database schema.

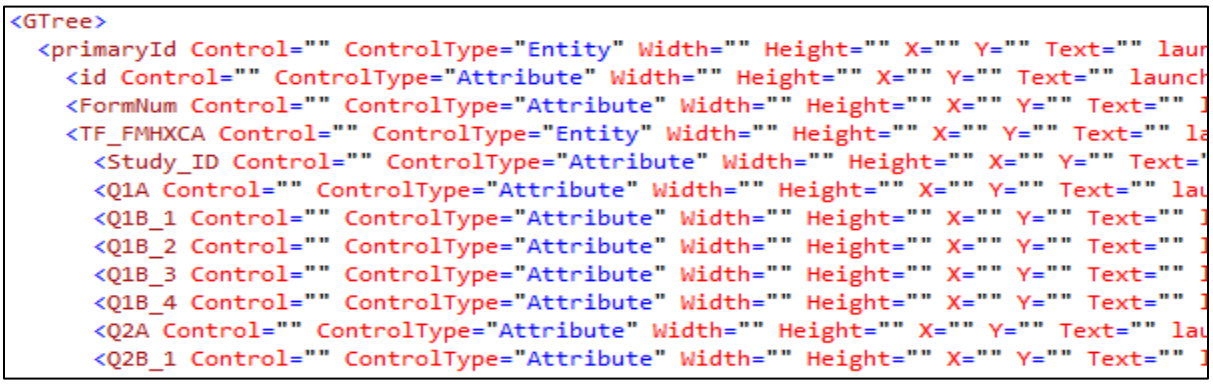


Figure 7: g-Tree For Teleform Data Source Without Contextual Information

```
<TF_FMHXCA Control="" ControlType="Entity" Width="" Height="" X="" Y="" Text="Family History Q
 <Form_Id Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="Form_Id" launc
 <BatchNo Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="BatchNo" launc
 <BatchRDate Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="Batch date"
 <FORMNUM Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="FORMNUM" launc
 <Study_ID Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="Study_ID" lau
 <Q1A Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="Mother?" launchTyp
   <Domain Type="Enumerated">
       <Item Text ="Yes">1</Item>
       <Item Text ="No">0</Item>
   </Domain>
  </Q1A>
 <Q1B_1 Control="" ControlType="Attribute" Width="" Height="" X="" Y="" Text="Types?" launchTy
   <Domain Type="Enumerated">
       <Item Text="Bone/Joint">01</Item>
       <Item Text="Brain/NervousSystem">02</Item>
       <Item Text="Breast">03</Item>
       <Item Text="Colon/Rectum">04</Item>
```

Figure 8: g-Tree For Teleform Data Source With Contextual Information

4.5 Functional Testing and Quality Assurance

Before presenting the tools to reviewers and analysts, the UI was functionally tested to measure the quality of functional components of the system. The goal of functional testing was to determine whether each component of the UI behaves correctly under all conditions that may be required by data analysts. We built and performed multiple studies using REDCap and TeleForm data to test the functionality of all the specified requirements of this tool. The following parameters were used during testing of the Analyst Dashboard, and the Analyst Dashboard was iteratively improved based on this testing.

1. Data source import verification: We tested the ability of the Analyst Dashboard to import (i.e. to use in studies) a variety of data sources.
2. Study creation verification: We tested building and saving new studies as well as the proper integration of imported data sources.

2.1 Data operator verification: We tested and evaluated the behavior of all operators.

2.2 Data mapping verification: We tested whether the mapping tool satisfied all the user requirements under different scenarios.

2.3 Classifiers verification: We tested the behavior of classifiers to ensure that they provided all the contextual information needed.

2.4 Result verification: We verified result sets by manually comparing the data from both data sources. We found various errors in this phase, which were fixed.

4.6 Using the Analyst Dashboard

Once the data sources were ready, the g-trees created, and the tools tested, the final step was to integrate the data using the analyst dashboard and to verify the usability of the tools. Figure 9 shows the flow of data from the two data sources to the analyst dashboard via g-trees. It also demonstrates that an analyst can build new studies by integrating and classifying data from these sources and save the study schema for further analysis and modifications.

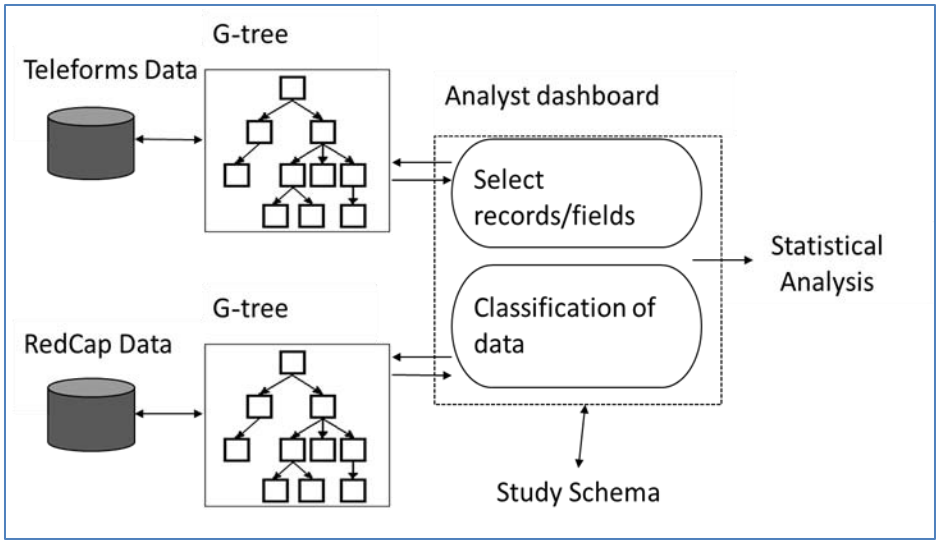
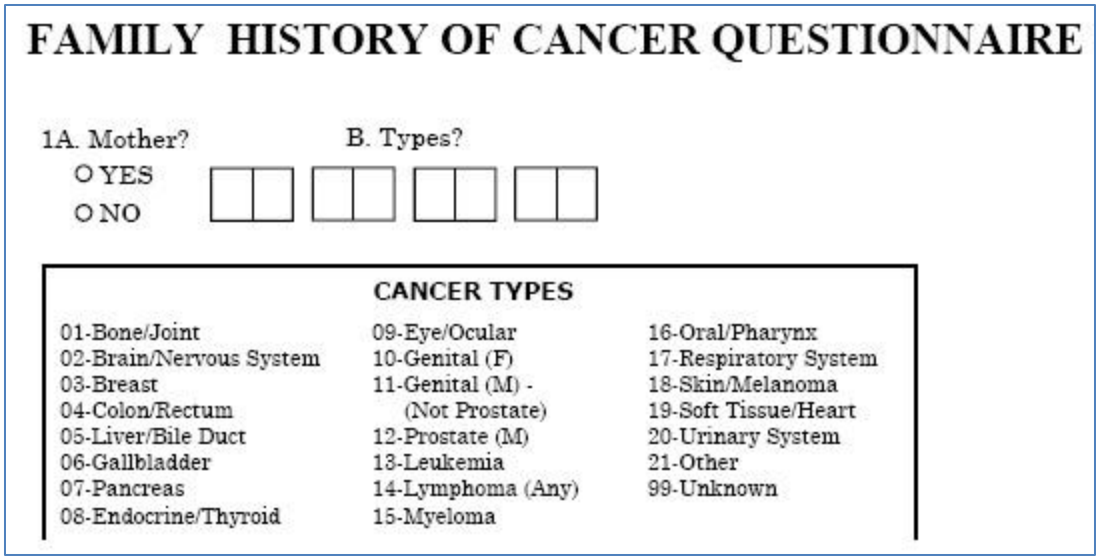


Figure 9: Represents The Flow Of Data From Data Sources To Analyst Dashboard For New Study Schema

There are various criteria defined by analysts for building a new study. These criteria are termed “dictionary elements.” These dictionary elements are presented as the result set of the study and may have categorical values, so the Analyst Dashboard allows classifying and mapping these elements from the underlying data sources according to the needs of the new study.

Figure 11 presents a sample form from the TeleForm questionnaire that illustrates some of the difficulty in integrating the chosen data sets. If the answer is "Yes" to question 1A, then up to 4 values may be entered in question 1B. For example, in the case of a patient whose mother has had multiple cancer types, example answers to 1B might be “01, 03, 13,” or “03, 10.” It should be noted that the order of types is not enforced, so an analyst will have to consider this limitation while mapping the data source.



FAMILY HISTORY OF CANCER QUESTIONNAIRE

1A. Mother?
O YES
O NO

B. Types? ☐☐ ☐☐ ☐☐ ☐☐

CANCER TYPES

01-Bone/Joint	09-Eye/Ocular	16-Oral/Pharynx
02-Brain/Nervous System	10-Genital (F)	17-Respiratory System
03-Breast	11-Genital (M) - (Not Prostate)	18-Skin/Melanoma
04-Colon/Rectum	12-Prostate (M)	19-Soft Tissue/Heart
05-Liver/Bile Duct	13-Leukemia	20-Urinary System
06-Gallbladder	14-Lymphoma (Any)	21-Other
07-Pancreas	15-Myeloma	99-Unknown
08-Endocrine/Thyroid		

Figure 10: An example from TeleForm Family History Questionnaire

Figure 11 also demonstrates how an analyst could map this as the dictionary element "Mother Breast Cancer," with categorical values in the new study "yes" and "no." The Analyst Dashboard allows analysts to map categorical values from multiple data sources (again, the two sources in this example are TeleForm and REDCap).

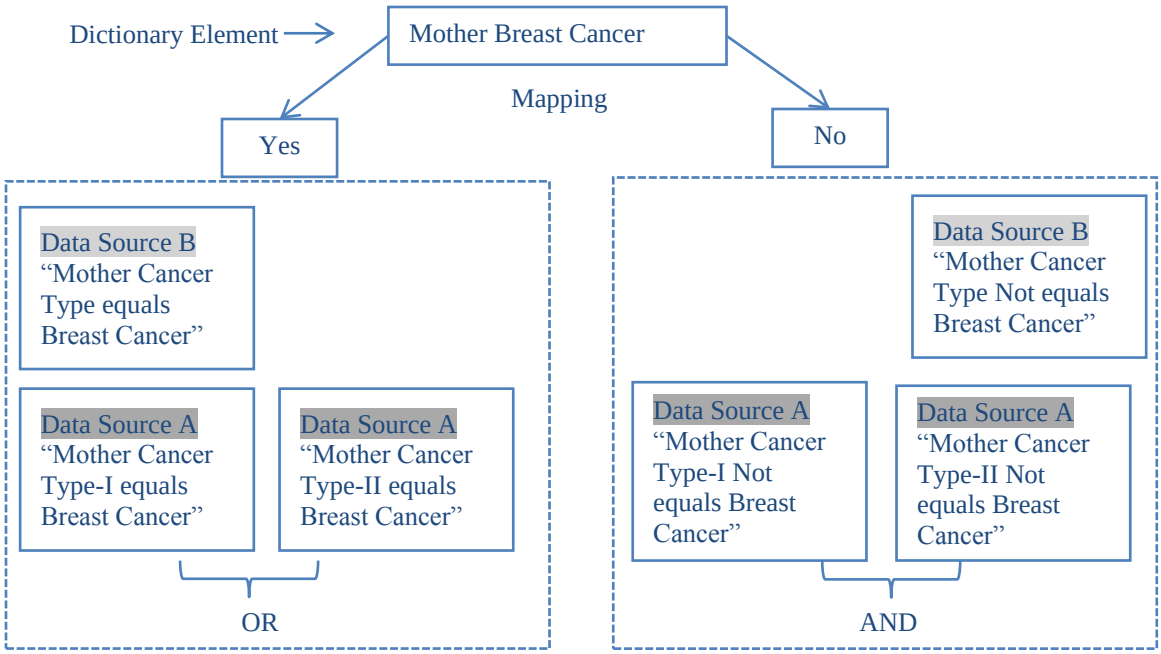


Figure 11: Relationship Between Two Categorical Values Of the Dictionary Element And Data Sources A&B

In addition to defining dictionary elements, analyst dashboard allows data analysts to filter the results by defining inclusion and exclusion criteria. For example, for including all patients with data element "Age" greater than value "55", an analyst has to define it in the dashboard as an inclusion criteria for the study. Similarly, for excluding patients with "no smoking history", the analyst can define an excluding criteria in the dashboard.

To summarize, here are the steps that can be used in the analyst dashboard to create a new study.

1. Add data sources/trial data sources (here, its REDCap and Teleform)
2. Specify and define Dictionary elements along with the categorical values if required.
3. Specify and define inclusion criteria.
4. Specify and define exclusion criteria.

5. Map all the criteria to each data source using the mapping tool.

    5.1 Select criteria to be mapped.

    5.2 Select data sources one by one for mapping to the criteria.

    5.3 For each data source, define one or more conditions using the data operators.

    5.4 Specify the Boolean relations between the conditions.

6. Run the query to get the results.

7. Save results.

The Analyst Dashboard was demonstrated to a convenience sample of subjects who are familiar with clinical trials research. In addition to demonstration of creation of a new study using this data, prototypes of suggested Analyst Dashboard UI modifications were shown to these subjects. Feedback was obtained about the interface, features and functionality of the Analyst Dashboard prototypes.

5 RESULTS

This experimentation using the existing platform (DBA dashboard and Analyst Dashboard) with clinical trials data suggests that not many architectural changes are required for full usability. With some improvements, the tools can be extended and made useful for clinical trials data analysts.

5.1 Data transformation operators

Similar to the example in Table 1, we analyzed all 250 elements in the results data set from the Duke study to find all the operators that study used, with the assumption that these operators would suffice for our next test. All the operators found were added to our Data Integration tool for its testing and implementation in clinical trials integration. Table 4 presents all fifteen unique operators that were identified. They are shown in the column titled "SQL Operator Defined." The rest of the columns are specific to the data from

OUPUT DATA (Form Element)	**INPUT DATA (Dimensions required to abstract the data)**	**Data Transformation**	**SQL Operator Defined**
Protocol study number	protocol number, site number	protocol number +site number	Combine.String
Patient's initials	Patient FirstName, Patient LastName	Patient FirstName+ Patient LastName	LTRIM/RTRIM
Recurrent Angina(Yes/No)	Discharge date, Documented MI	Yes = (If documented MI -discharge date)< 30 days	DATEDIFF
Prior discharge (Yes/No)	EnrolledDate, Discharge date	Yes = (enrolled date – discharge date) < 30 days	< (comparison)
Documented EligibilityViolation (check box if TRUE)	date of violation **,** enrollment date	Yes = date of violation > enrollment date	> (comparison)
Symptoms attributed to	documented	**Any:**	'ANY'

chronotropic incompetence	symptoms	fatigue/dyspnea/dizziness/effort intolerance/lead problem/other	
Patient withdrew consent (Yes/No)	Documented withdraw of consent date	If **Exists**'patient withdrew consent' = Yes	'EXISTS'
Revascularization planned within 2 weeks of entry into Z-Phase (YES/NO)	Z-phaseEnrollment, datePlanConceived	**If** (Z-phaseEnrollment – datePlanConceived)>14 days **then** Revascularization = 'Yes' **ELSE** Revascularization = 'No'	If-then-else
Reason for Patient discontinued, Other specify	Presence of other reason for discontinuation , other reason description	**IN** (Presence of other reason for discontinuation)	'IN'
CholesterolLevel< 150 or >250 mg/dl (check box if any of them TRUE)	total cholesterol value	total cholesterol value < 150 **OR** >250 mg/dl	OR
Bolus 0.4 mcg/kg/min for 30 minutes (check box if TRUE)	dose type, dose rate, dose duration, 30 minute criterion	If dose type = 'Bolus' **AND** dose rate= '0.4 mcg/kg/min' **AND** dose duration= '30 min'	AND
Date Tirofiban First given (day/month/year)	date patient received first dose (mm/dd/yyyy)	Change date formatting	DateFormat (Cast and Convert)
Time Tirofiban First given (00:00 to 23:59) (hh:mm)	time patient received first dose (hh:mm:ss)	Change time formatting	TimeFormat
Initial dose (Enoxaparin): amount given in ml	dosage amount, required units	Change Units	UnitConversion = (- , +, \, *) Arithmetic Operators
Other event yes/no	documented other event, event date, discharge date, 30-	**!=** stroke AND **!=** CABG AND **!=** MI AND **!=** PTCA	'!='

Table 4: Operators Identified From Duke Study Result Set. Form_ID represents the name of the form in which the element was found; OUPUT DATA (Form Element) represents the format or name of the element needed by researchers in the results; INPUT DATA shows the data elements required to abstract the element from the forms; Data Transformation shows the type of transformation or conversion; and SQL Operator Defined defines the type of operator.

Duke. These operators are used by clinical analysts for data mapping and transformation in the Analyst Dashboard.

5.2 G-tree annotator

As previously discussed, for certain data sources, the DBA dashboard does not provide functionality to build a g-tree containing rich contextual information. However if the source data can provide a data dictionary, as was found with the REDCap data, a tool can be built to automatically generate g-trees with contextual information. Figure 13 shows the interface of the tool that we built for the purpose of automatically generating a g-tree with contextual information using the REDCap data dictionary.

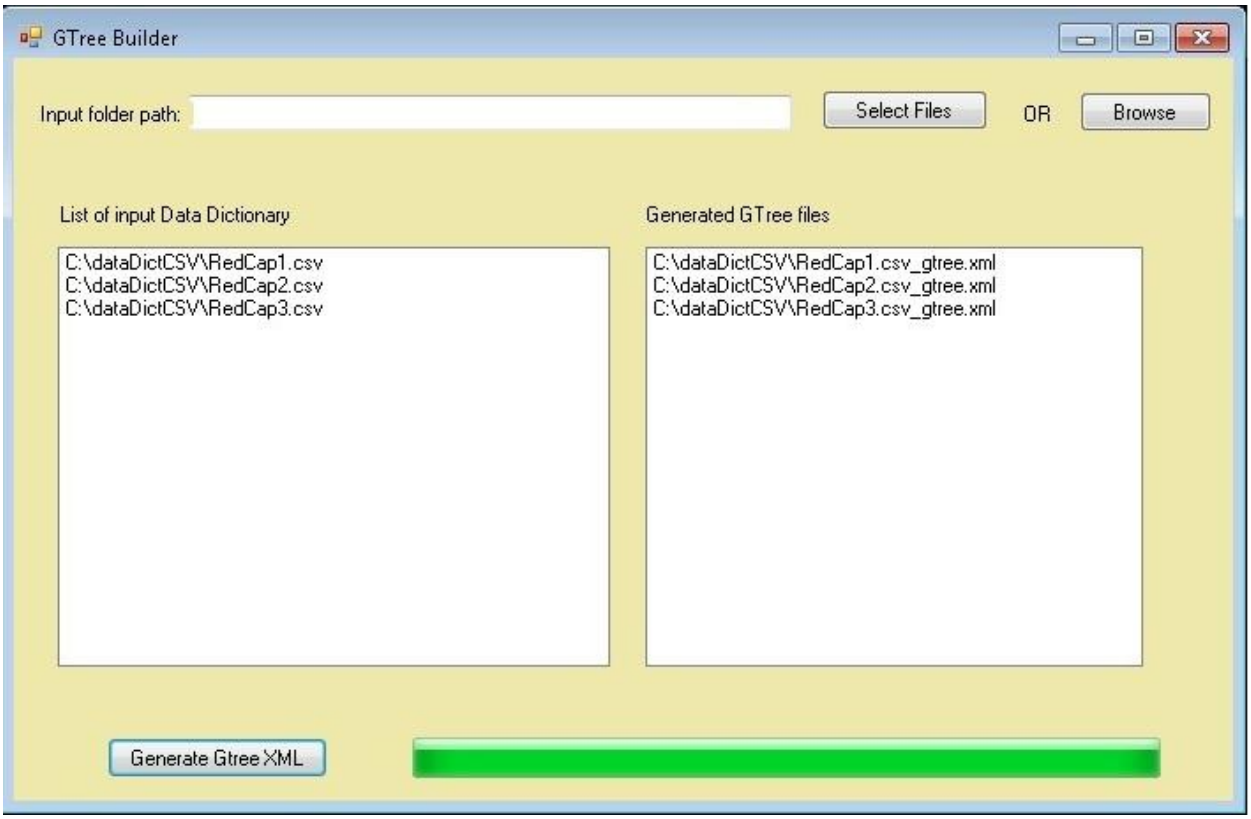


Figure 12: Screenshot of an Application to Generate g-Trees From Redcap Data Dictionary

5.3 Analyst Dashboard Prototype

The current Analyst Dashboard is functional; however, an analyst using it would not intuit how to operate the tool. Although this version makes the process of data integration simpler as compared to manually integrating data, it does not provide any guidance to the analyst during the data integration process, thus making the process non-intuitive and error prone. For example, as shown in Figure 13, the first step of the process is to select the data sources that the analyst needs to build the new study. The current version of the Dashboard requires the analyst to have knowledge of GUAVA server location and port number for each data source. A more intuitive approach would be to allow the data analyst to browse to a file that represents the GUAVA server for the specific data source.

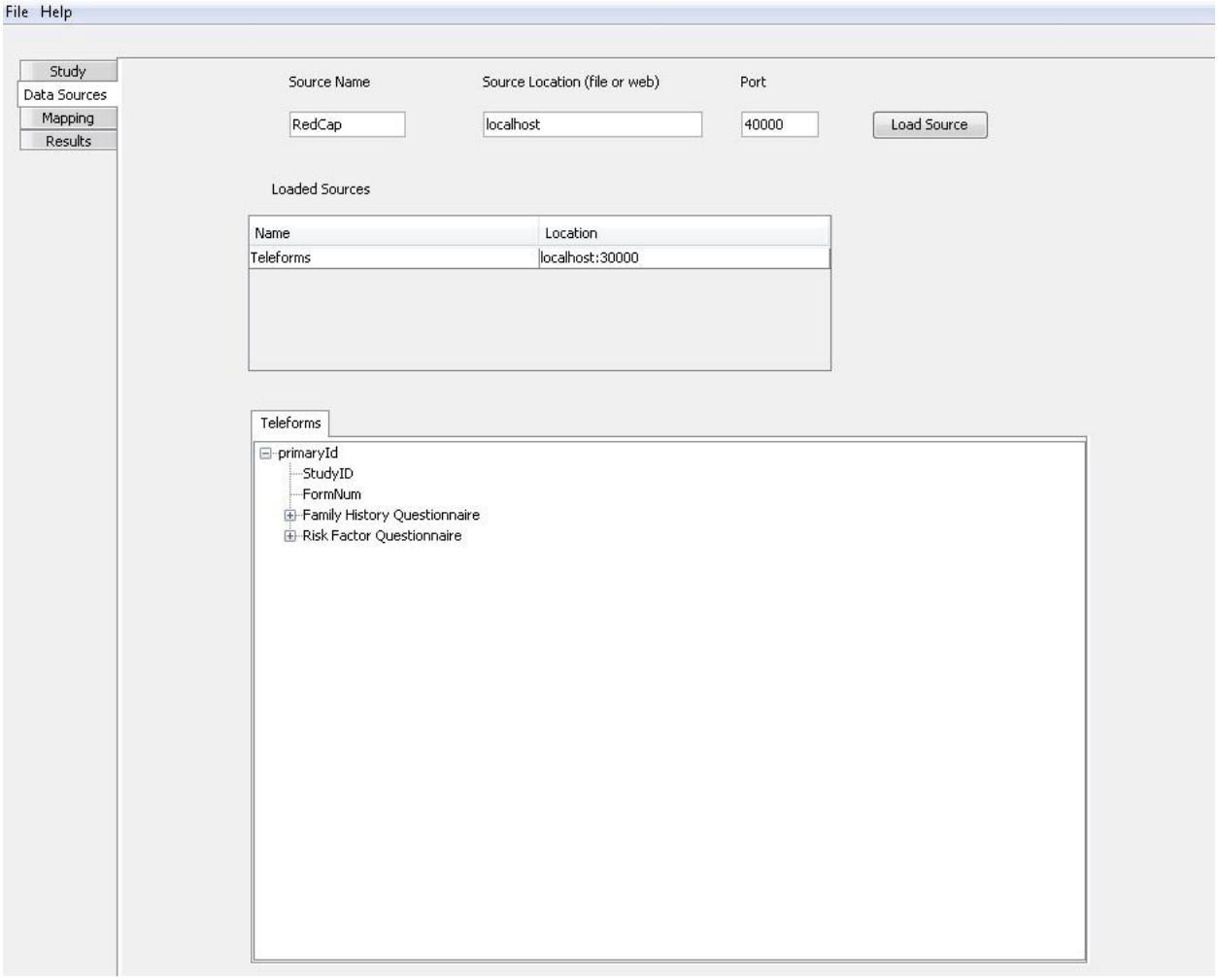


Figure 13: Older Version Of Analyst Dashboard

In preparation for modifying the Analyst Dashboard UI, a prototype was created using Visual Studio in the C# programming language. The prototype was designed to be more intuitive for users. One of the requirements of this tool is to reduce the load of technical learning for the analyst. The step of selecting data source was modified and a simple selection of sources was added, as shown in Figure 14, which can be selected using 'Browse' button; alternatively, it can be selected from a drop-down menu presented as "user data sources." The database port connection is transparent to the user. The step-

by-step functionality of this improved dashboard is described via a case study in section 6.

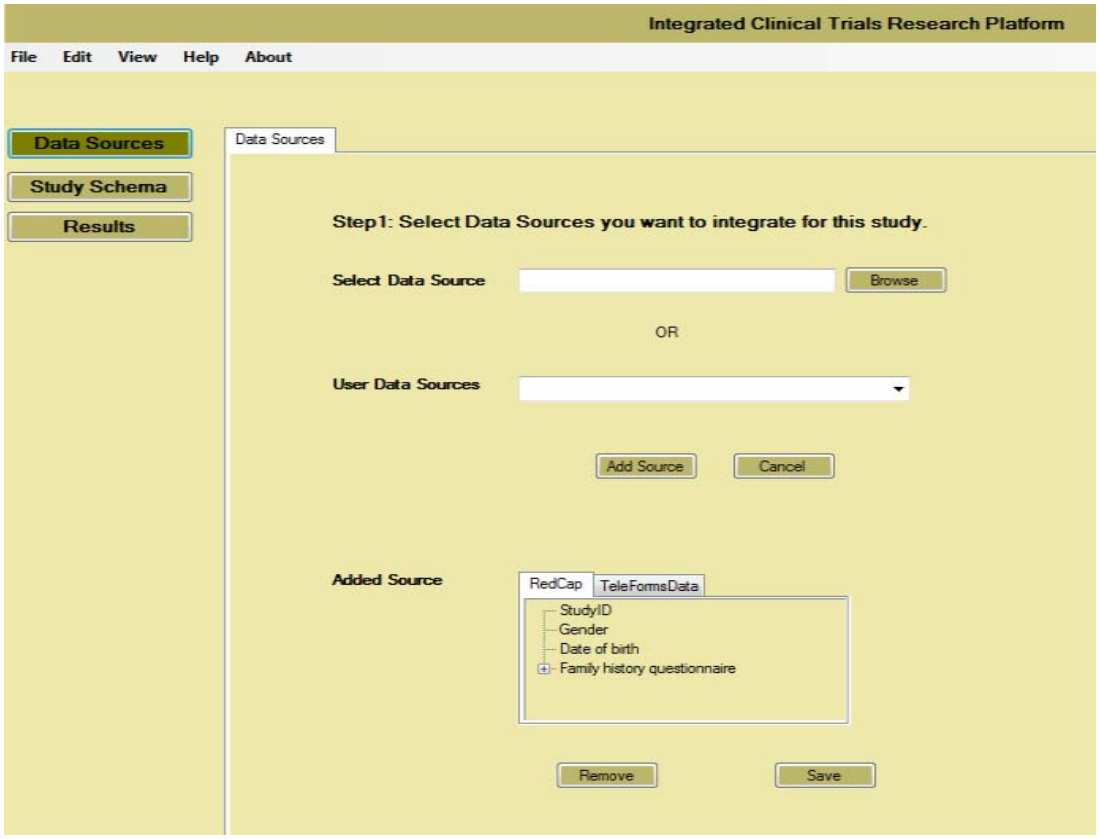


Figure 14: New UI Prototype Data Source Selection Screen

The appendix lists and explains all of the prototype screens that were developed during this project.

5.4 Analyst Interviews and Outcome

Usability testing of the current tool and prototype was performed with three subjects. This testing highlights the areas that need further improvements from an analyst's perspective. It also ensures that the tool is working efficiently with all different

combinations, is easy to use, and provides all the information needed by the analysts. A new study was performed using the current Analyst Dashboard. Then, subjects were asked to repeat that study using the prototypes and to comment on the interfaces as they proceeded. If not already mentioned by them, we asked specifically about the following metrics:

- Ease of the process navigation
- Effectiveness of presentation of user interface and
- Success rate of the task.

The study performed was designed to answer the following question:

What is the incidence of maternal family history (in the subject's mother, maternal grandmother or aunts) with breast cancer?

The result set needed to have four columns: Participant ID, Mother (options: Yes/No), Grandmother (options: Yes/No), and Aunts (options: Yes/No). In addition, the study was to include patients with date of birth greater than 12/31/1946 from TeleForm and no age limit from REDCap. We also illustrated the use of exclusion criteria by excluding patients who are not married. Table 5 and Table 6 show all the criteria specified for this study:

Criteria	**Name**	**Definition**
Inclusion	Date of Birth	Include Patients with DOB greater than 12/31/1946 for TeleForm and REDCap
Exclusion	NotMarried	Exclude patients who are not married

Table 5: Inclusion and Exclusion Criteria

Dictionary elements	Definition	Categorical values	Allow Nulls
ID	Patient reference ID	None	No
MotherBC	Mother Breast Cancer	Yes, No	No
MotherSisBC	Mother's sister with breast cancer	Yes, No	No
Mother's mother BC	Mother's mother has breast cancer	Yes, No	No

Table 6: Dictionary Elements And Definitions

After the completion of steps involved in building a new study and review of the prototypes, various suggestions were made by the subjects to enhance the functionality of the process. One such suggestion was to modify some of the clinical and data integration terms, which would increase the usability of UI. Notes were taken while demonstrating the study and new prototype dashboard. Below are a few suggestions made by the analysts:

1. Instead of two inclusion and exclusion criteria, use one field of "Eligibility Criteria."
2. Add check boxes that will allow users to mark the fields that are to be included or excluded in the final data set.
3. Add Boolean operators (AND/OR) to specify multiple inclusion/exclusion criteria.
4. Change the naming of study schema to "Query builder."
5. Opt to display raw data from the chosen data sources on the result screen.

6 DISCUSSION

The data integration problem has been around for many decades, although there has not been a silver bullet to solve all the integration problems encountered thus far. One of the major reasons the problem of data integration has been so challenging is semantic heterogeneity of the data. The semantic heterogeneity exists because the interpretation of the data is different among different people. For example, database schema created separately by two people can be recognized as different although representing the same data. The problem worsens when the data schema is designed for different organizations by different people even though the problem domain is the same. For example, as previously illustrated, the term "pain" to denote a data element is not sufficient enough to unambiguously determine if the term means the presence and absence of pain or the severity of pain. These ambiguous states are unavoidable, since it is not possible to completely define the meaning of a data element in the database. To overcome this challenge, a possible solution can be to present the data to the expert consuming it and let that expert do the interpretation of the meaning of the data. The premise is that since they are the domain experts, if given a single, unified view of multiple data sources and necessary tools to merge the data they would be able to solve the semantic heterogeneity problem.

The previous works by Logan, Delcambre, and Terwilliger built a data integration framework that presents the domain experts with a view of multiple heterogeneous data sources and tools to define the rules of merging this data. This was a significant step where the database experts would 'connect' the data to the framework's *DBA dashboard*,

and the domain experts (clinical analysts, in our case) would query it using *Analyst Dashboard*. Additionally, they could create their rules and views of data from these heterogeneous sources.

The flexibility of this framework motivated our current work. We wanted to analyze the strengths and weaknesses of the existing tools and their applicability in the area of clinical trials data integration, which could benefit the general medical field. For this purpose, we analyzed various clinical trial studies and selected one study that could set us on the right path for further exploration. We found that the previous framework needed more data transformation operators to meet the needs of the clinical analysts. We also found that the Analyst Dashboard was too cumbersome to be used by clinical analysts and was not providing them with the high-level view of the data that is needed for integration and analysis.

In order to meet the needs of clinical data integration, we added more data transformation operators to the framework, which will help the clinical analyst during the data mapping process. Additionally, we improved on the Analyst Dashboard user interface so that its use is more intuitive. We conducted three interviews with the clinical analysts and presented them with the enhanced framework to see if the tools were well suited for their needs and to locate other areas of potential improvement. The outcome of the interviews was encouraging, with the framework proving useful to the analysts during the data integration exercises.

In this project we added incremental values to the existing framework by enhancing its components and demonstrated, with the help of interviews, that the framework is feasible for use in the clinical trials domain. However, due to limited time allotted to this

project, we were not able to test our improvements on more clinical trial datasets that would help us to further improve the framework. More follow-up interviews need to be conducted to make this system stronger to tackle wider problems that may arise during clinical trials data integration.

7 CONCLUSION

The data collection methodologies used during clinical trials do not currently have common standards, which leads to complexities during clinical trials data integration. We found that the current data integration tools implemented in the DBA dashboard and Analyst Dashboard provide a robust framework and can be customized for clinical trials integration without significant changes to the existing architecture. The Analyst Dashboard received positive reviews from the investigators who viewed it and made suggestions for improving the UI. Additionally, more interview sessions with analysts would be beneficial for understanding their needs and further improving the Dashboard.

8 BIBLIOGRAPHY

1. Terwilliger JF, Delcambre LM, Logan J. Querying through a user interface. Data & Knowledge Engineering 2007 Dec; 63(3).

2. Terwilliger JF, Delcambre LM, Logan J. Context-Sensitive Clinical Data Integration. Proceedings of the Workshop on Information Integration in Healthcare Applications (IIHA), in conjunction with the Conference on Extending Database Technology; 2006; Munich.

3. Terwilliger JF, Delcambre LM, Logan J. The User Interface is the Conceptual Model. Proceedings of the 25th International Conference on Conceptual Modeling; 2006; Tucson.

4. Nahm M, Nguyen VD, Razzouk E, Zhu M, Zhang J. Distributed Cognition Artifacts on Clinical Research Data Collection Forms. 2010. CRI Summit.

5. Research Electronic Data Capture [Internet] [cited 2010 December 5]. Available from: http://project-redcap.org/.

6. TeleForms and eForms [Internet] [cited 2010 December 5]. Available from: https://www.ohsu.edu/cbcibm/isr/teleforms.cfm.

7. Hughes R. Applied Clinical Trials Online: Multi-Trial Data Integration [Internet]. Available from: http://appliedclinicaltrialsonline.findpharma.com/appliedclinicaltrials/article/articleDetail.jsp?id=145640.

8. Miller RJ, Hernandez MA, Haas LM, Yan LL, Ho CTH, Fagin R, et al. The Clio Project: Managing Heterogeneity. Proceedings of SIGMOD; 2001; 30(1):78-83.

9. Brenton J, Caldas C, Davies J, Harris S, Maccallum P. CancerGrid: developing open standards for clinical cancer informatics. Proceedings of the UK e-Science All Hands Meeting; 2005;:678-681.

10. NIH. http://grants.nih.gov/grants/glossary.htm#C [Inernet] [cited 2010 December 10]. Available from: http://grants.nih.gov/grants/glossary.htm#C.

11. Glossary and Acronyms List, Office of Exramural Research [Internet] [cited 2010 December 10]. Available from: http://grants.nih.gov/grants/funding/phs398/instructions2/p2_human_subjects_definitions.htm.

12. Mailhot DW [Internet] 2006 [cited 2010 December 10]. Available from: http://www.nihtraining.com/cc/ippcr/current/downloads/Mailhot020706.ppt.

13 Cohen FJ. Can You Handle the Truth? Journal of clinical research best practices. 2007 . Sep; 3(9).

14 CaBIG NIH [Internet] 12/15/2010 [cited 2010 December]. Available from: . https://cabig.nci.nih.gov/.

15 David IH. CDISC - An operational data Model - Ready to Roll? Applied Clinical . Trials. 2004 July.

16 Wikipedia [Internet] [cited 2010 December 10]. Available from: . http://en.wikipedia.org/wiki/Database_schema.

17 Batini C, Lenzerin M. A Methodology for Data Schema Integration in the Entity . Relationship Model. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING. 1984 Nov; 10(6).

18 Batini C, Lenzerini M, Navathe SB. A Comparative Analysis of Methodologies for . Database Schema Integration. ACM Computing Surveys 1986; 18(4):323-364.

19 Wikipedia [Internet] [cited 2010 December 10]. Available from: . http://en.wikipedia.org/wiki/Data_integration.

20 Hawkey CJ, Weinstein WM, Stricker K, Murphy V, Richard D, Krammer G, et al. . Clinical trial: comparison of the gastrointestinal safety of lumiracoxib with traditional nonselective nonsteroidal anti-inflammatory drugs early after the initiation of treatment. International Aliment Pharmacol Therapy 2008 May; 27(9): p. 838-845.

21 Drozd DR, Lober WB, Kitahata MM, Smith KIRSV. Developing a Relational XML . Schema for Sharing HIV Clinical Data. Proceedings of the AMIA Annual Symposium; 2005.

22 Mugavero MJ, Lin HY, Allison JJ, Giordano TP, Willig JH, Raper JL, et al. Racial . Disparities in HIV Virologic Failure: Do Missed Visits Matter? Journal of Acquired Immune Deficiency Syndromes 2009; 50(1).

23 Sterne JA, May M, Costagliola D, Wolf FD, Phillips AN, Harris R, et al. Timing of . initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. Lancet. 2009 April;:1352-1363.

24 Timmers T, Pierik F, Steenbergen M, Stam H, Ginneken AMv, Mulligen EMv, et al. . ARIS: integrating multi-source data for research in andrology. Proceedings of the Annual Symposium on Computer Application in Medical Care; 1995:445–448.

25 Louis T, Lavori P, Bailar J3, Polansky M. Crossover and self-controlled designs in . clinical research. The New England journal of medicine 1984 Jan; 310(1).

26 OHSU Graduate Studies Faculty [Online] [cited 2010 December 10]. Available from:

. http://www.ohsu.edu/xd/education/schools/school-of-medicine/academic-programs/graduate-studies/faculty/grad-studies-faculty.cfm?facultyid=215.

27 Lois M.L. Delcambre, Ph.D [Online] [cited 2010 December 10]. Available from:
. http://web.cecs.pdx.edu/~lmd/.

28 Oregon Clinical and Translational Research Institute [Internet] [cited 2010 December
. 5]. Available from: http://www.ohsu.edu/xd/research/centers-institutes/octri/index.cfm.

29 Sheth AP, Gala SK, Navathe SB. On Automatic Reasoning for Schema Integration.
. International Journal of Intelligent and Cooperative Information Systems. 1993; 2(1).

APPENDIX: PROTOTYPE SCREENS FOR THE ANALYST DASHBOARD

This section illustrates (using screenshots) the functionality of the new analyst dashboard via the process of building a new study. The main aim for this exercise is to show the ease and usability of data integration process. For this demonstration, we selected the same trial as in the older version of the UI prototype (Refer to the study trial referenced in section 5.4, "Maternal Breast Cancer History").

Step 1: Build new study.

As a first step, the option *build new study* is selected, as seen in Figure 15. (Note that a previously saved study can also be opened from this menu). This selection changes the view of the screen by presenting the next steps of building the trial including *data sources*, *study schema* and *query results*.

Figure 15: Build New Study Selected: UI Prototype

Step 2: Select data sources

Figure 16 demonstrates the selection of data sources and its various options. The analyst can add multiple data sources by browsing through all the available options or, alternatively, selecting data sources from user's data library. The user data library stores all the data sources previously used by analyst.

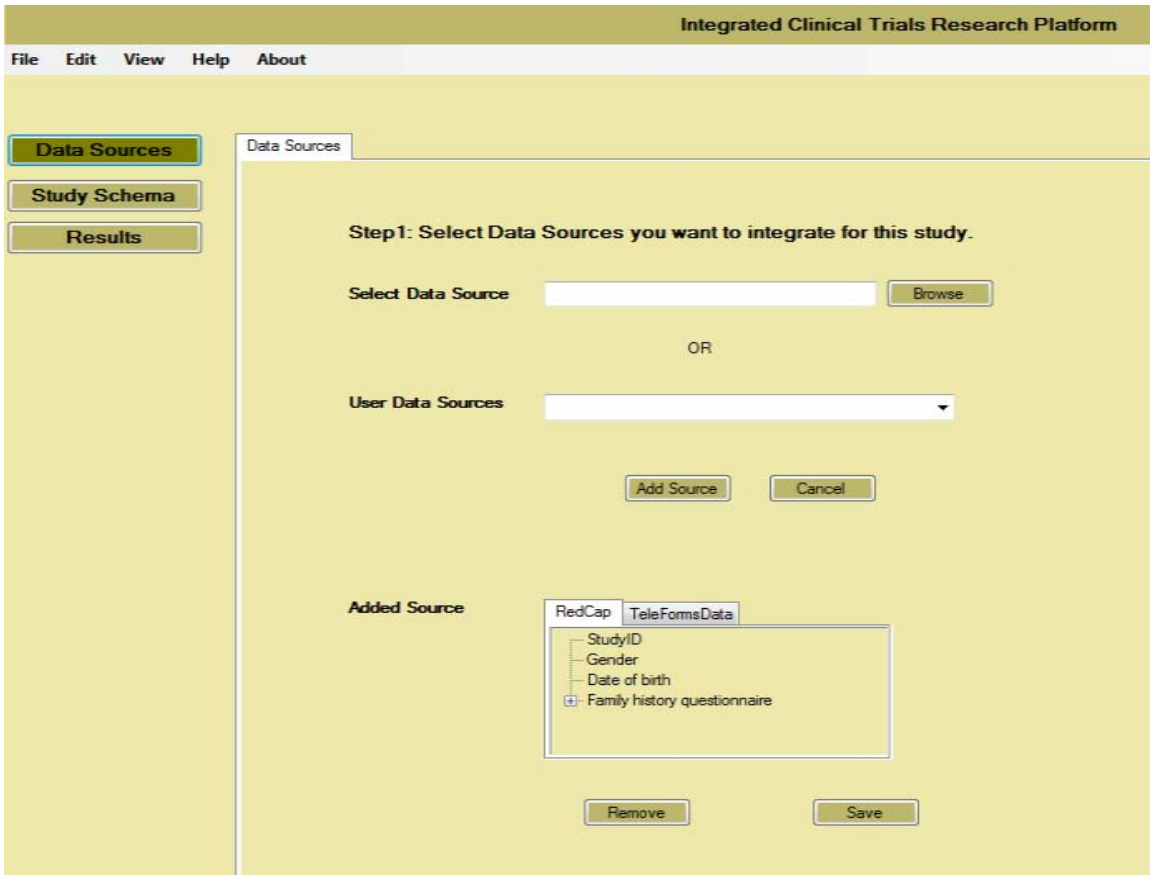


Figure 16: Data Sources Already Added By The Analyst

Step 3: Build Study Schema.

Building the study schema includes various steps of defining criteria and mapping the criteria to different data sources. This step is divided into multiple small steps as follows:

Step 3.1: Define Inclusion Criteria

In this study, we define 'Age' as our inclusion criteria. For maternal breast cancer, we want all the patients with date of birth greater than 12/31/1946. Figure 17 shows the inclusion criteria screen of new UI prototype. The analyst can define any number of inclusion criteria for a single study.

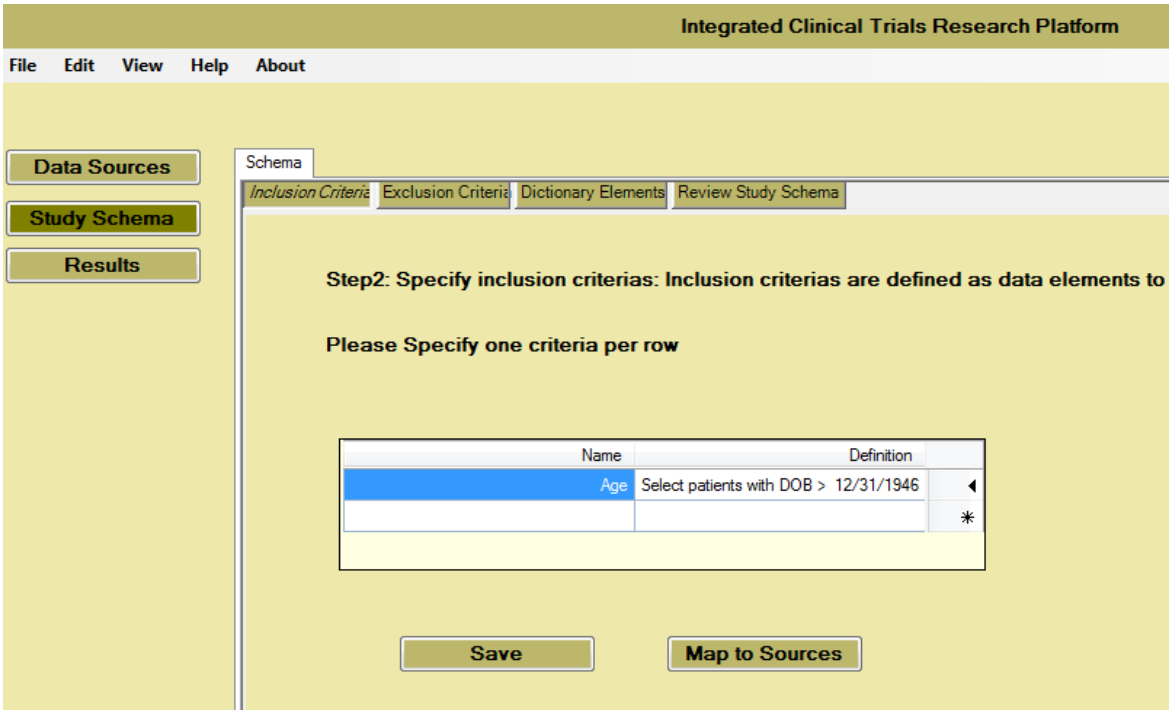


Figure 17: Define Inclusion Criteria

Step 3.2: Map Inclusion Criteria

After defining inclusion criteria, the analyst needs to map this to all the added data sources. Figure 18 shows stepwise guidance of mapping criteria. The first step is to select the data source to map these criteria. Next, the analyst can add various conditions by selecting the data element from the tree view in the selected data source. The following step includes the selection of the operator from the drop-down list ('Greater Than'). The analyst can then specify any particular value for the criteria (12/31/1946, in this

example). The *build mapping* button control creates the condition. If there are multiple conditions, an analyst can add relations between these conditions. In this example, there are no relations as there is only one data element.

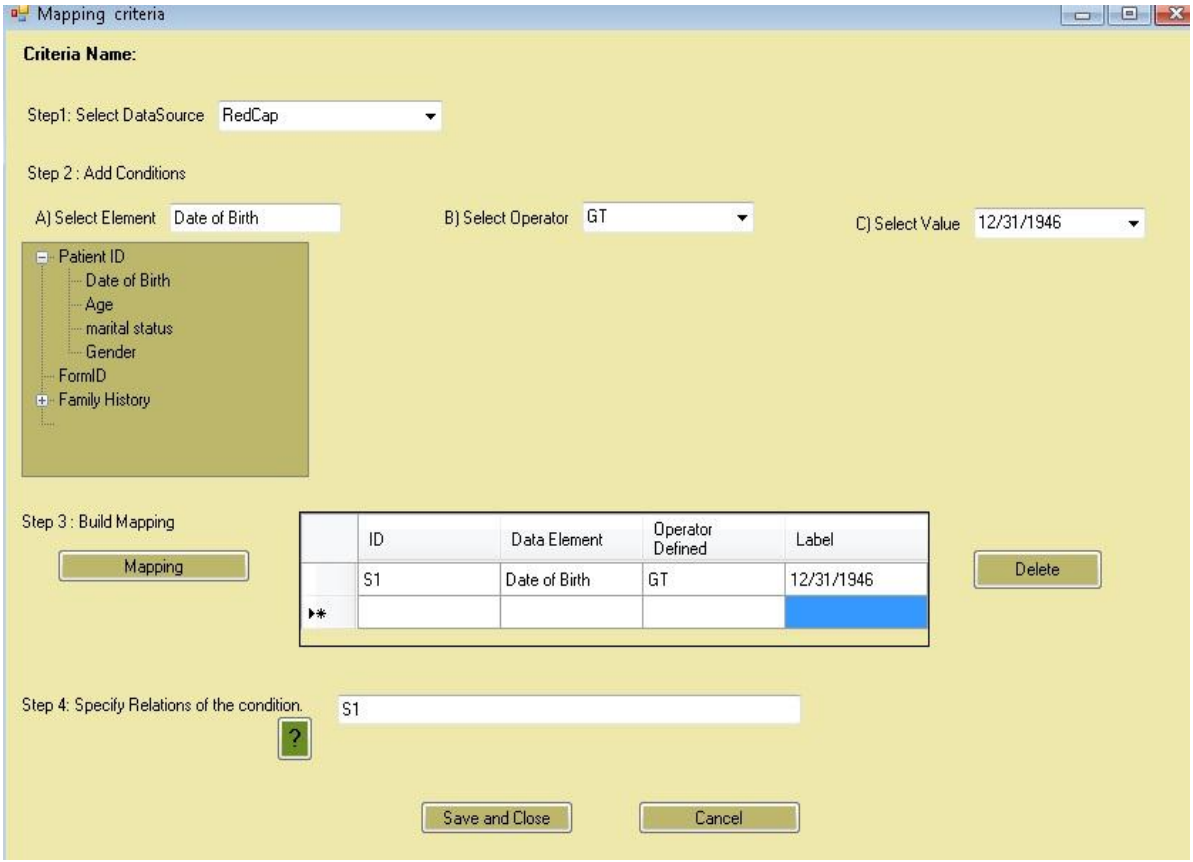


Figure 18: Mapping Inclusion Criteria

Step 3.3: Define Exclusion Criteria

An analyst can then choose 'Marital Status' as an exclusion criterion of this study, which excludes all patients who are not married. Figure 19 shows the selected and defined exclusion criteria.

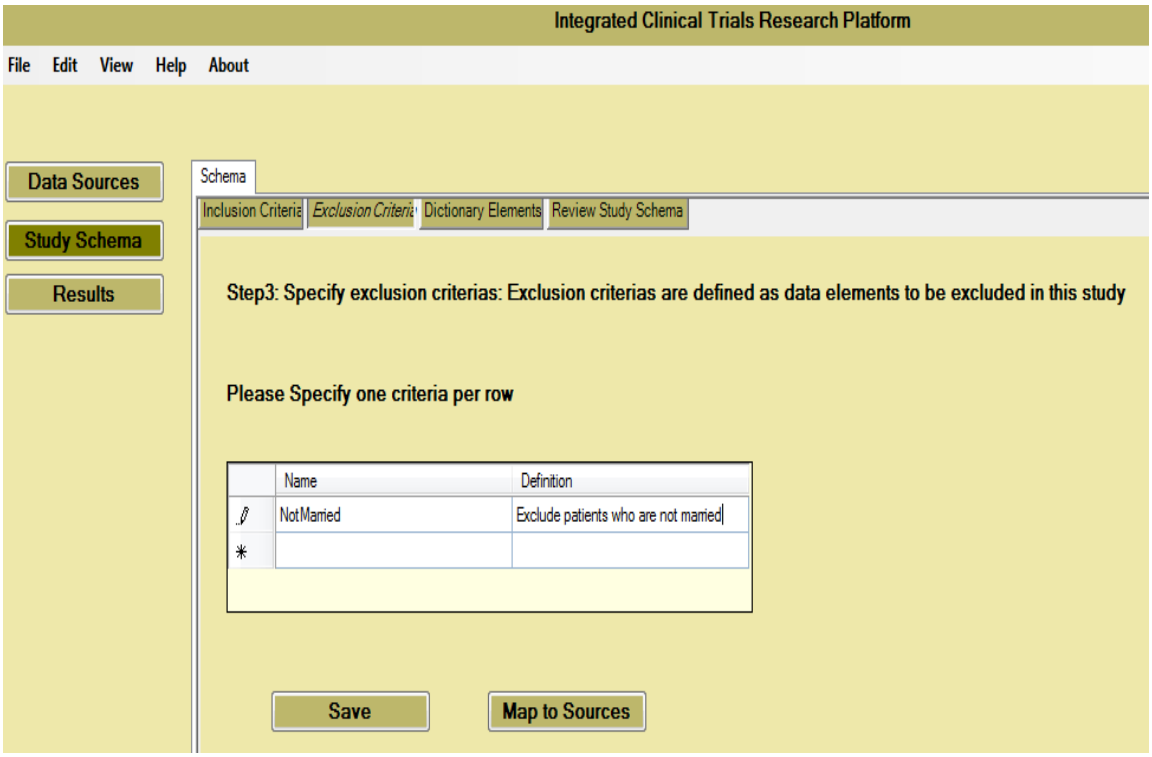


Figure 19: Define Exclusion Criteria

Step 3.4: Map Exclusion Criteria

After defining exclusion criteria, the analyst needs to map this to all the added data sources. Figure 20 shows stepwise guidance of mapping criteria. Similar to the process of mapping inclusion criteria, the first step is to select the data source to map these criteria. An analyst can add various conditions by selecting the data element from the tree view in the selected data source. Next step includes selection of operator from the drop-down list ('equal to'). Then, the analyst can select the value from the drop-down list, which includes all the possible values from the data sources. In this case, the drop-down includes *Married*, *Divorced*, *Separated* etc. The choice *Married* is selected for this criterion. *Build mapping* creates the condition. In this criterion, there is no relation as there is only one data element.

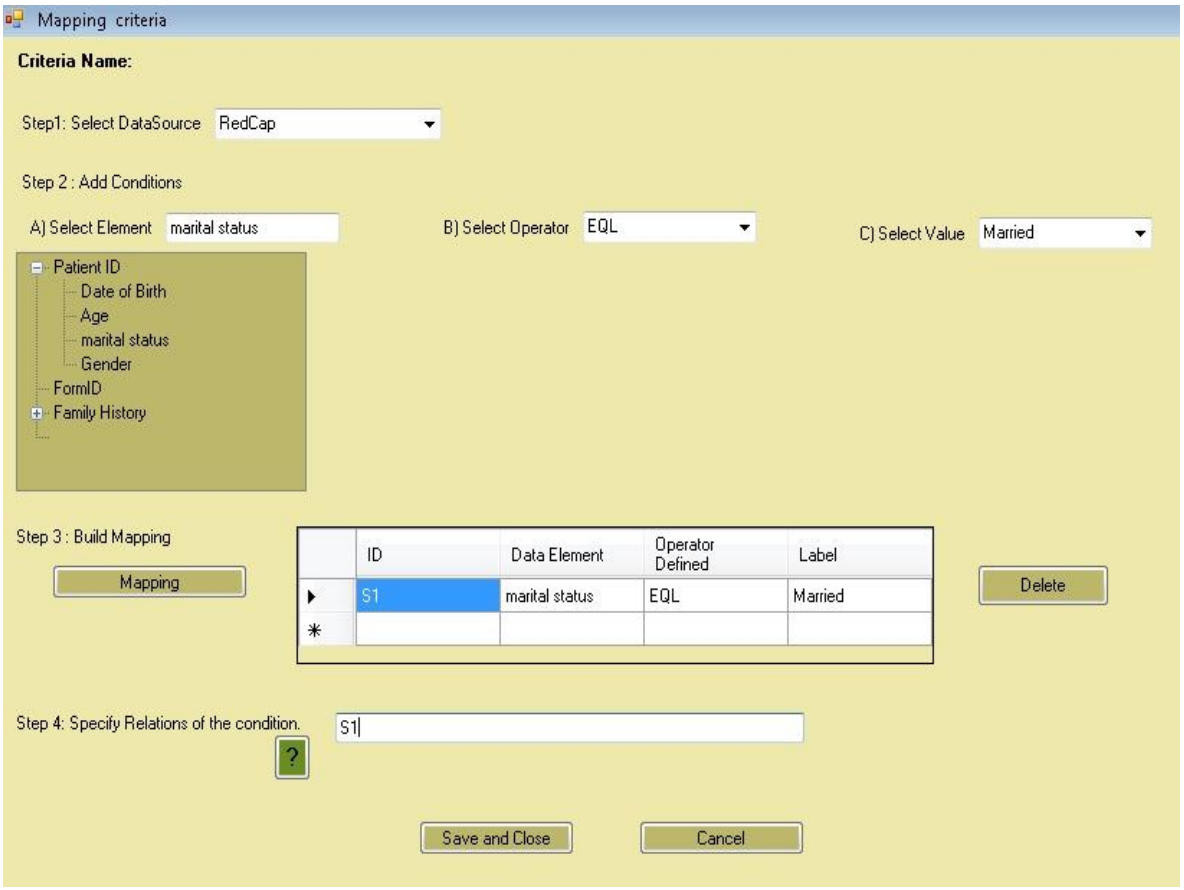


Figure 20: Mapping Exclusion Criteria

Step 3.5: Define Dictionary Elements

In this step, an analyst can define all the elements that need to be included in the result set. Refer to section 5.4 (Table 6). The analyst can define multiple dictionary elements for any given study. She can also define categorical values if needed for these elements. Elements can be defined to be NULL or not NULL.

In Figure 21, four different dictionary elements are shown to be selected, ID with no categorical values, Mother Breast cancer, Mother's Mother Breast Cancer and Mother's Sister Breast Cancer all three with Yes and No categories.

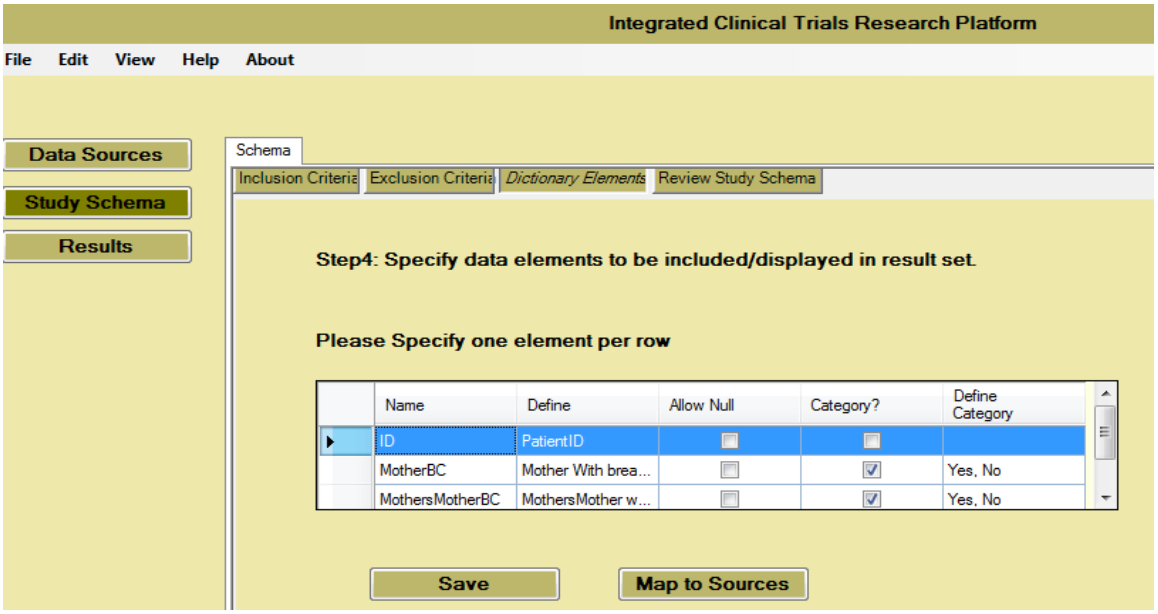


Figure 21: Define Dictionary Elements

Step 3.6: Mapping Dictionary Elements

Mapping dictionary elements needs requires more steps than mapping inclusion and exclusion criteria. The analyst may need to individually map each category of elements to the selected data sources. The analyst can select one element at a time for its mapping to all the sources. The mapping of ID was done using the same steps as inclusion/exclusion criteria, since in this case it does not have any categorical values.

"Mother's Breast Cancer" has two categories. Figure 22 explains the stepwise procedure of mapping this data element. The first step is the selection of data sources; next the analyst can select the categorical value (in this case, *yes*). Following that the analyst can select the element from the tree view of selected data source. The operator 'Equals To' is selected in this case. The selected value in step 3(C) lists all the possible values of that data element from the selected data source (for example, *Colon cancer, Breast Cancer, Abdominal Cancer, Lung Cancer,* etc.). It is also possible that the selected

data source has multiple elements listed that are needed for a dictionary element, so the analyst can build multiple conditions. As shown in Figure 22, two *types* are selected from the tree view of data source, and the relationship between these two conditions is defined as 'S1 OR S2'. The reason there can be multiple *types* is because of the way data is collected as explained in section 4.6 (Figure 10).

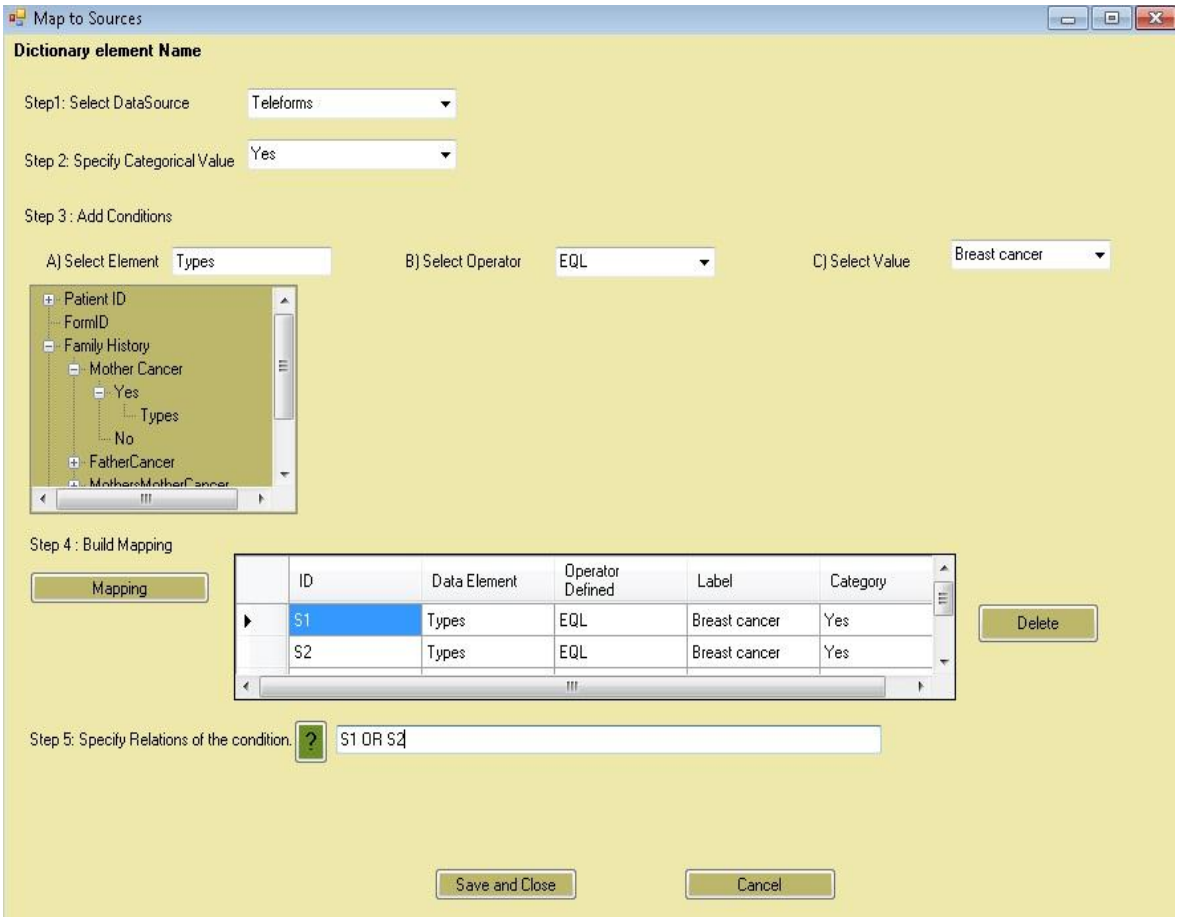


Figure 22: Mapping Dictionary Elements

Step 3.7: Review Schema

In this step, the analyst can verify the study schema and finalize all the elements and mapping for data integration. Figure 23 shows all the defined criteria and dictionary elements along with check-boxes that define if the criteria and/or elements are mapped.

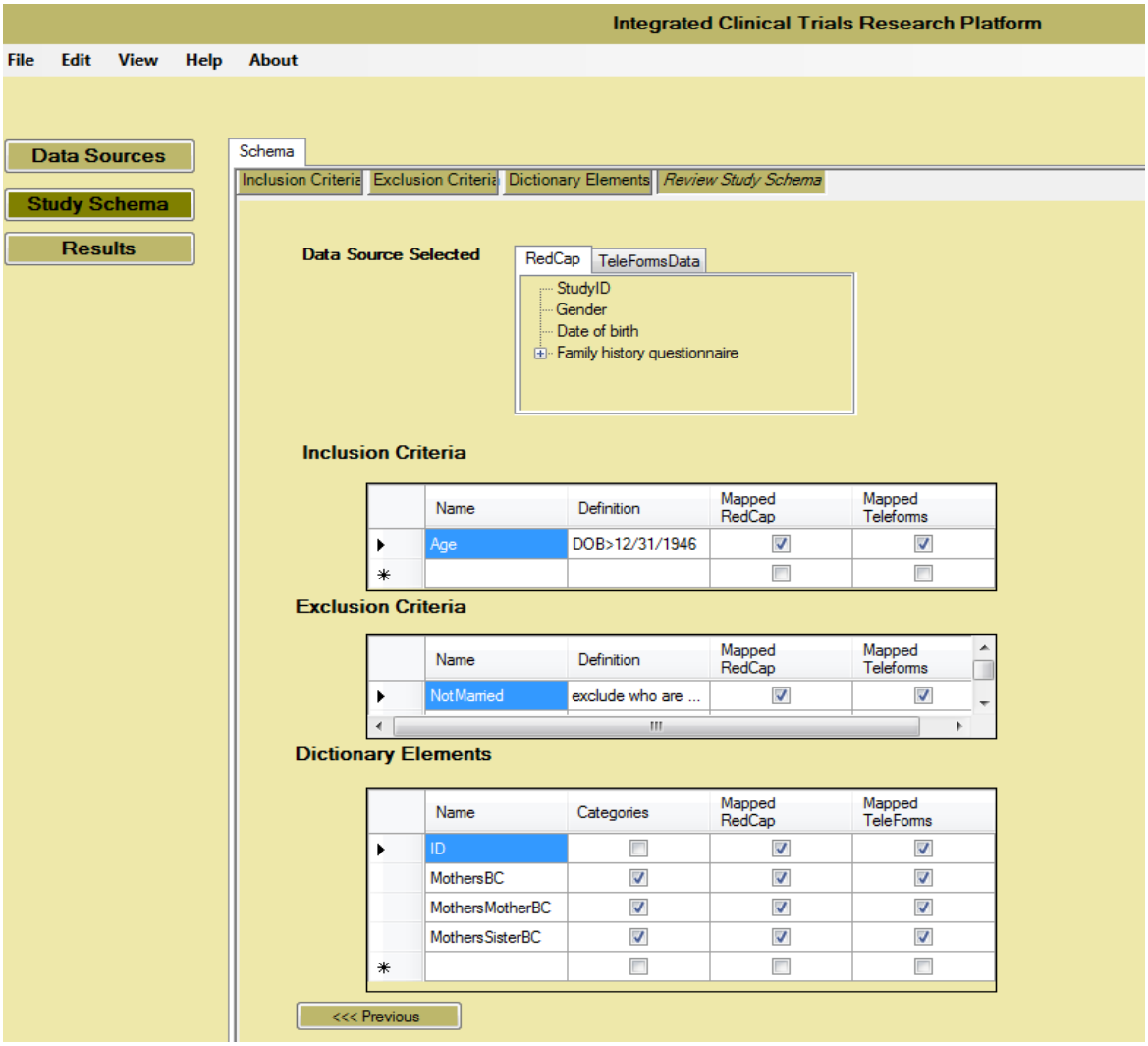


Figure 23: Review Study Schema

Step 4: Run Query

This is the last step that generates a dataset representing the study according to the criteria defined by the analyst. Figure 24 shows the result dataset that is generated for this case study. The analyst can run the query against both of the selected data sources and integrate the data from all sources. The output data is displayed in the form of a spreadsheet and the result can also be exported to a comma-separated file. The analyst can save the query or study for any future modification and research.

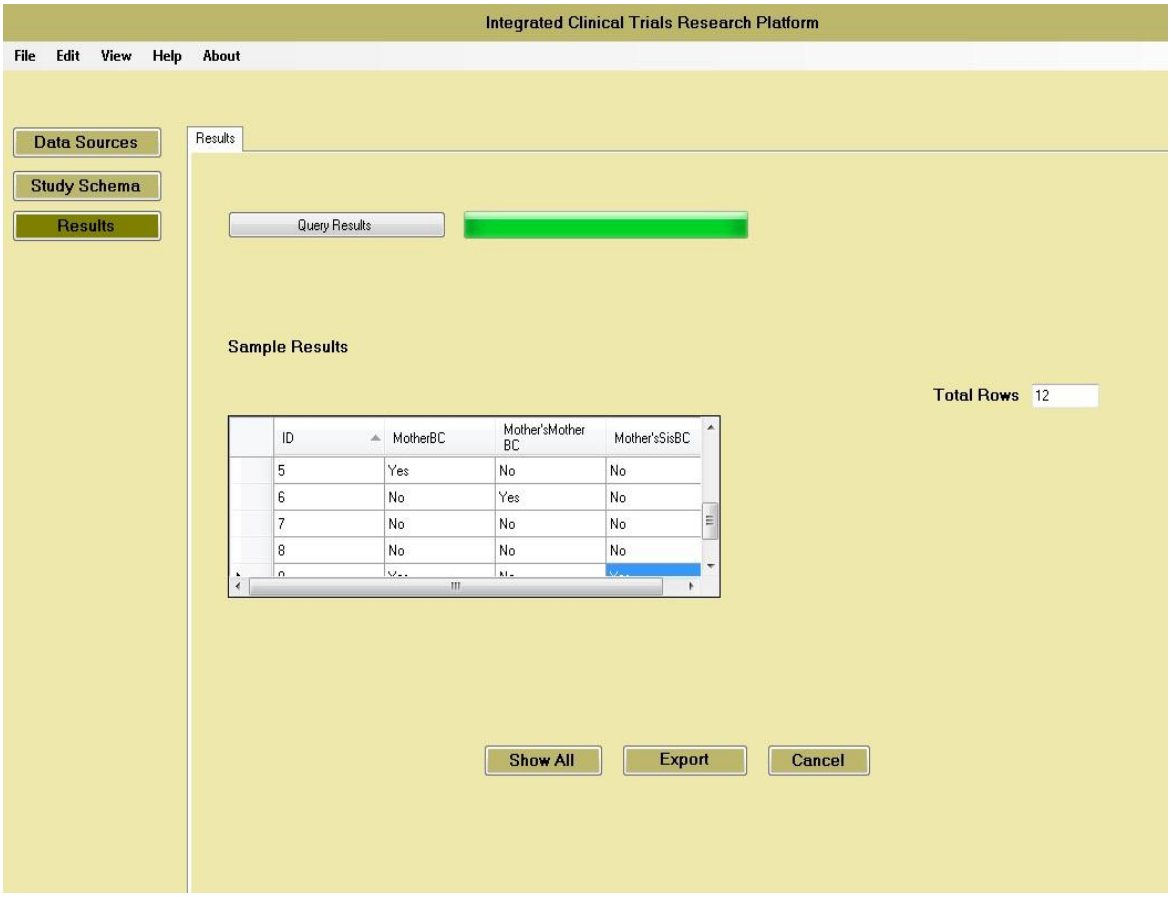


Figure 24: Run Query Screen With Results Displayed