

NCICB Clinical Trials Outcomes Project DCP Biometry Research Group Knowledge Acquisition Session Report

Session Date: September 16, 2003 **Session Time:** 1:30-3:00pm EST

Session Topic: User Requirements for a Clinical Trials Outcomes System (Statistician's Perspective)

Knowledge Analyst: Bill McCurry, ScenPro, Inc.

Session Location: NCI Offices - Rockville, MD

Type of Session:

Interview

Task Analysis

Scenario Analysis

Concept Analysis

Observation

Structured Interview

General Topic Area

The NCI Center for BioInformatics (NCICB) is funding an effort to develop a technical solution to the problem of providing complete and reliable clinical trial outcomes data to the cancer research community. Due to the large scope of developing a solution to collect, manage, report and analyze clinical trial outcomes data, the project has been divided into multiple phases. The focus of this Phase I effort is on gathering specific user data requirements and desired system functionality.

Report Summary

This report documents information gathered during a Knowledge Acquisition session with experts for the Division of Cancer Prevention (DCP), Biometry Research Group (BRG). The DCP designs, develops, implements and monitors cancer prevention trials. The BRG provides statistical analysis and study design expertise, support, consultation and advice to the DCP, NCI divisions and external organizations. BRG is mainly comprised of Mathematical Statisticians who:

- Plan and conduct studies on cancer epidemiology, prevention, screening and diagnosis using statistical methods
- Conduct statistical analyses of clinical and pre-clinical prevention trial data sets to identify relationships between patient baseline, interventions and outcomes data
- Develop and validate models for statistical research in cancer prevention activities

Dr. Vance Berger, Ph.D. and Donald Corle, M.S. are Mathematical Statisticians in the Biometry Research Group. Dr. Berger has expertise in ensuring validity through design and analysis of randomized clinical trials. He develops novel between-group analyses to compare the outcomes experienced by the subjects in one treatment group to the outcomes experienced by the subjects in the other treatment group. Dr. Berger specializes in selection

bias for detection, prevention and correction in randomized clinical trials. Mr. Corle specializes in statistical methods, interactive statistical data analysis, sample size and power calculations for experimental designs in prevention trials.

Summary of Findings

Information obtained during this session includes:

- BRG requires detailed patient-level data to perform the necessary statistical functions
- BRG statisticians see the outcomes system as a way to retrieve the data they need to conduct their in-depth analyses of cancer prevention activities
- Types and levels of data required for their analyses.
- Potential issues and challenges that should be addressed when implementing the outcomes system
- Documentation of an example as-is scenario of how statisticians currently access outcomes data as well as a to-be scenario of how statisticians would use the outcomes system.

Desired Data Requirements

Data Needed for Analysis

Figure 1 is a representation of the primary data types BRG statisticians need to support their analysis. Study events and measurement timing are not as critical as the actual patient and study endpoint data for statistical analysis.

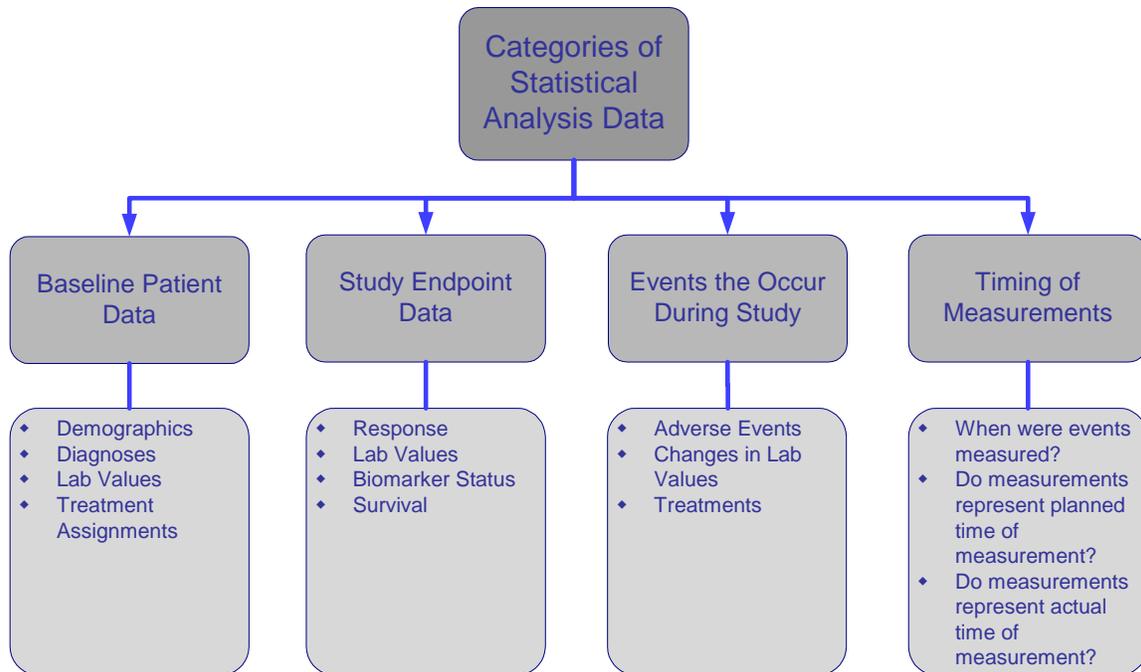


Figure 1. Data Needed for Analysis

Desired Features to Support Data Quality

Statisticians want detailed data in a standardized and raw form to perform advanced statistical analysis. Automated conversions of data are of limited use to statisticians. For example, a basic conversion from Fahrenheit to Celsius might yield a level of impreciseness that is unacceptable to statisticians. They prefer to obtain data in its original unit of measure and then perform the necessary data conversions after download.

Statisticians need to see simple descriptive measures of data such as mean, median, distribution measure, and extreme measures to perform their analysis. They are also interested in additional detail such as definition, units and codes for a data element. Implemented in a system, this data could be accessed by simply clicking on the data element. The name and descriptor are not enough information for the statistician; the metadata about the set is as important as the data themselves. Understanding the values will require explanation of any codes. The explanation may be in the data set itself or attached in a data dictionary. The statistician uses these methods to evaluate the reasonableness, quality and accuracy of the data.

Another method to ensure data quality is to have the system indicate that data is missing. Statisticians will then code or recode missing values for the analysis.

Data quality would be acceptable to statisticians with the implementation of a 'QA Pedigree' feature. This pedigree would be an indicator of whether the data had been reviewed and approved and who conducted the review.

Aggregated Vs. Patient-Level Data

Statisticians want covariant data at the individual participant level in order to perform their statistical analysis. Any data aggregated above the participant level is of limited use for statisticians. Data aggregations at the treatment assignment or randomization group level are useful in conjunction with participant level data for statistician's analytical tasks. Data aggregated at a higher level might be useful for the following types of analysis:

- Metadata analysis involving questions that are addressed by evaluating a number of studies
- Political questions such as showing how many studies were done involving particular sorts of outcomes

Search Parameters

Statisticians are interested in obtaining study and trial participant information by querying any and all available data and attributes. The specific query criterion depends on the statistician's research question. Figure 2.0 below represents examples of the types of information that would be used to conduct queries.

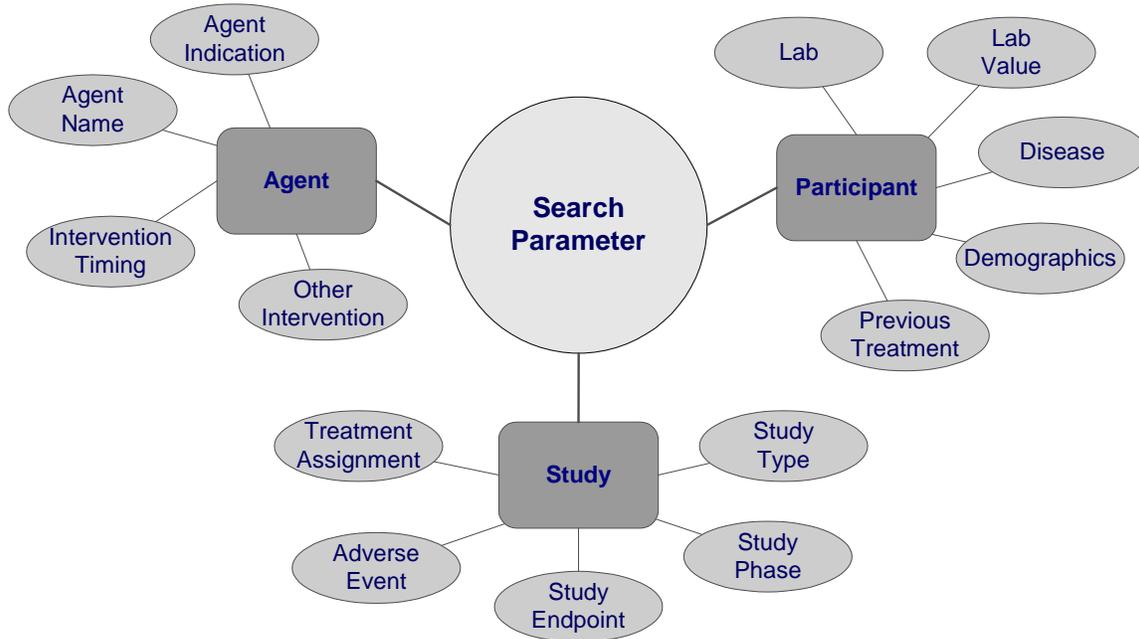


Figure 2. Search Parameter

Refining Search Parameters

One desired query feature for the outcomes system is the ability to refine the search parameters after the initial search is conducted. For example, a user should be able to find a set of studies involving early stage prostate patients over age 50 (500 studies), then refine the search by limiting it to phase III studies (100 studies), then further refine the search by selecting studies only involving agent X. This yields a subset of studies that meet all the search criteria.

Saving Search Parameters

Another desired function of the outcomes system is the ability to save and download search parameters with the selected data set. This feature would provide users with the ability to document exactly how a data set was selected. This is important if the user wants to use the saved parameter as a basis for editing and refining the query. This would also enable users to share queries between users and to extend research into other cancer areas. Statisticians may learn more about what data are needed as the analysis progresses. It would be useful for a statistician to return to the system after the initial download, change the original parameters or add new data elements, and then download the data set again.

Desired System Functions

Search Functions

A comprehensive, effective and easy to use search interface is one of the most important system functions for the statisticians. They are interested in accessing study and patient information by querying on a variety of data. Statisticians would like to query by any parameter in the data set. Statisticians would like the following functions to be included when querying data:

- Access study abstracts, data dictionaries and other detailed information with the trial data
- Ability to refine search parameters
- Pop-Up menu of searchable items similar to NIH library website
- Display and ability to save search parameters

Displaying Data

When the Outcomes System returns data to the user, Statisticians would like the ability to perform the following functions with that displayed data:

- View descriptive measures of data (mean, median, distribution measures, extreme values)
- Execute pre-processing functions such as: sort data elements, select specific data elements, and recode data elements
- Obtain information about metadata about data by clicking on the screen (definitions and descriptions of data elements, units of measurement and their values, codes and their meanings)
- View universal unique identifier for each study and each piece of data that conveys meaning in and of itself rather than just being a sequential indicator. For example, at the study level, a certain portion of the identifier could indicate the type of the study, another could indicate the disease/condition, and another could indicate whether the trial was randomized.

Data Retrieval Process

Data Sources

The DCP Monitoring Contractor (Westat) is collecting study data using Oracle Clinical. The Biometry Research Group statisticians can access electronic data sets from this database. If the trial is sponsored by DCP, a tape of the trial data may be included with the final report. However, this is not standard for every study. There are significant challenges due to non-standardized formats and lack of a data dictionary that includes the definitions and detail of the data.

Downloading the Data

After getting the data, the statisticians will spend time getting familiar with the data set, reviewing descriptive measures of the data, and resolving problems in the data such as outliers or missing data. The statistician must be able to use the outcomes system to sort,

select, reject and recode data to prepare a data set for download. Recoding includes collapsing data into more general categories (such as taking age in years and collapsing it into several age categories). More extensive processing and analysis of the data should not be performed in the outcomes system, but should be done by the statistician after downloading the data.

Each statistician has their own data format, analysis tools, and analysis platform preferences. Delimited flat files and MS Excel files are commonly used and are useful for taking an initial look at the data. The download should take no longer than 30 minutes.

The following information must be downloaded in conjunction with the data set itself:

- Metadata about the data set (data element descriptions and meanings of codes)
- Query details by which the data was selected

After downloading the data, statisticians use a variety of software and hardware to perform their analysis. Software might be SAS or similar commercial analysis packages, however many statisticians prefer to write their own analysis programs. Hardware may range from personal computers to mainframes.

The statisticians carry out involved analyses using sophisticated statistical procedures. These analyses involve identifying relationships between patient baseline data, treatment assignments, and outcomes. The statisticians also develop statistical models that may be used for analysis.

User Interaction with the Outcomes Data Retrieval Processes

The Biometry Research Group needs to retrieve outcomes data and descriptive measures and details of that data in order to support the DCP with analytical and mathematical statistical analyses. Figures 3.0 and 4.0 below represent current and future data retrieval scenarios. By using the outcomes system to access stored data, statisticians will be able to retrieve more data and explanation of that data in less time than performing functions manually. These processes are described in detail below.

Current (As-Is) Outcomes Data Retrieval Process

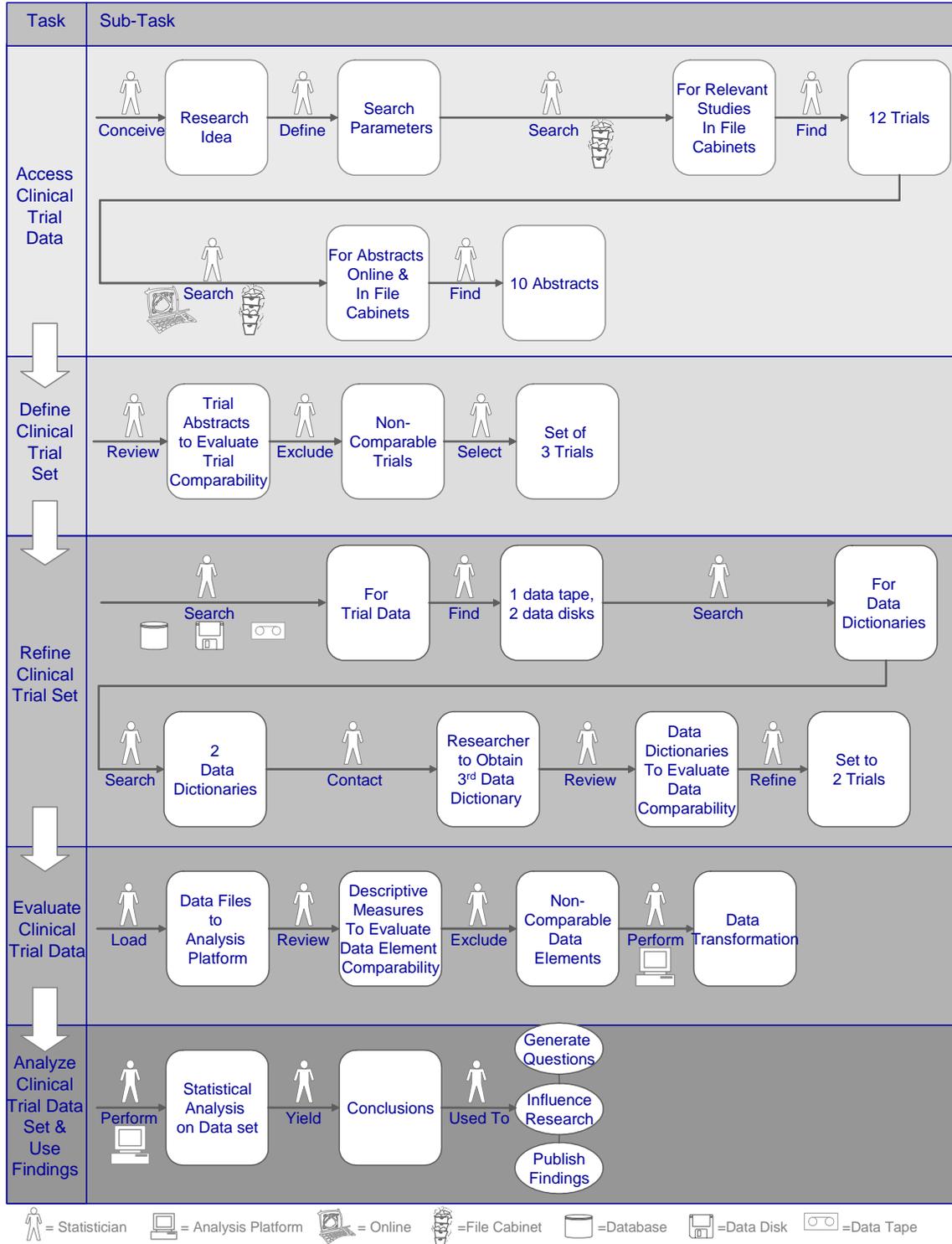


Figure 3. Outcomes Research Scenario: As-Is (Current State)

As-Is Scenario Text

While reviewing past analyses, a DCP statistician wonders if men with early prostate cancer suffer complications after their treatment with Proscar has ended. The statistician confers with fellow researchers and reviews his own records for applicable and accessible data sets. This preliminary research identifies 12 relevant studies. The statistician searches online and through filing cabinets and finds abstracts for 10 of the 12 studies. He reviews the abstracts to evaluate comparability of the studies. The statistician selects 3 of those studies that seem to have applicability to the research question.

The statistician manually obtains trial data from a data tape and two data diskettes for the 3 studies. This data is manually obtained by searching through completed DCP trial records and from other statisticians. One of the data sets has no data dictionary, so the statistician contacts the researcher to get the dictionary. The statistician reviews the 3 data dictionaries for the data sets to evaluate comparability of the data. He reviews detailed descriptions of the data such as:

- When were data collected?
- What was the meaningful identifier?
- What were the units of measurement?
- What are the values of the measures?
- What method was used to collect the data?

He determines that one trial is not comparable and refines his trial set to 2.

The statistician loads the data files, abstracts and data dictionaries to his analysis platform. He uses an analysis program to review the descriptive measures of the 2 data sets. The statistician identifies several data elements that need to be dropped from the data sets because they are not comparable between trials. The statistician performs the appropriate data transformations (recode, drop data elements, and exclude outliers) to consolidate the 2 data sets to one.

The statistician performs statistical analyses to answer the question of interest and draws conclusions from the analyses. The statistician implements his findings to generate new questions, publish findings, or to share the new lines of research with other researchers.

Future (To-Be) Outcomes Data Retrieval Process

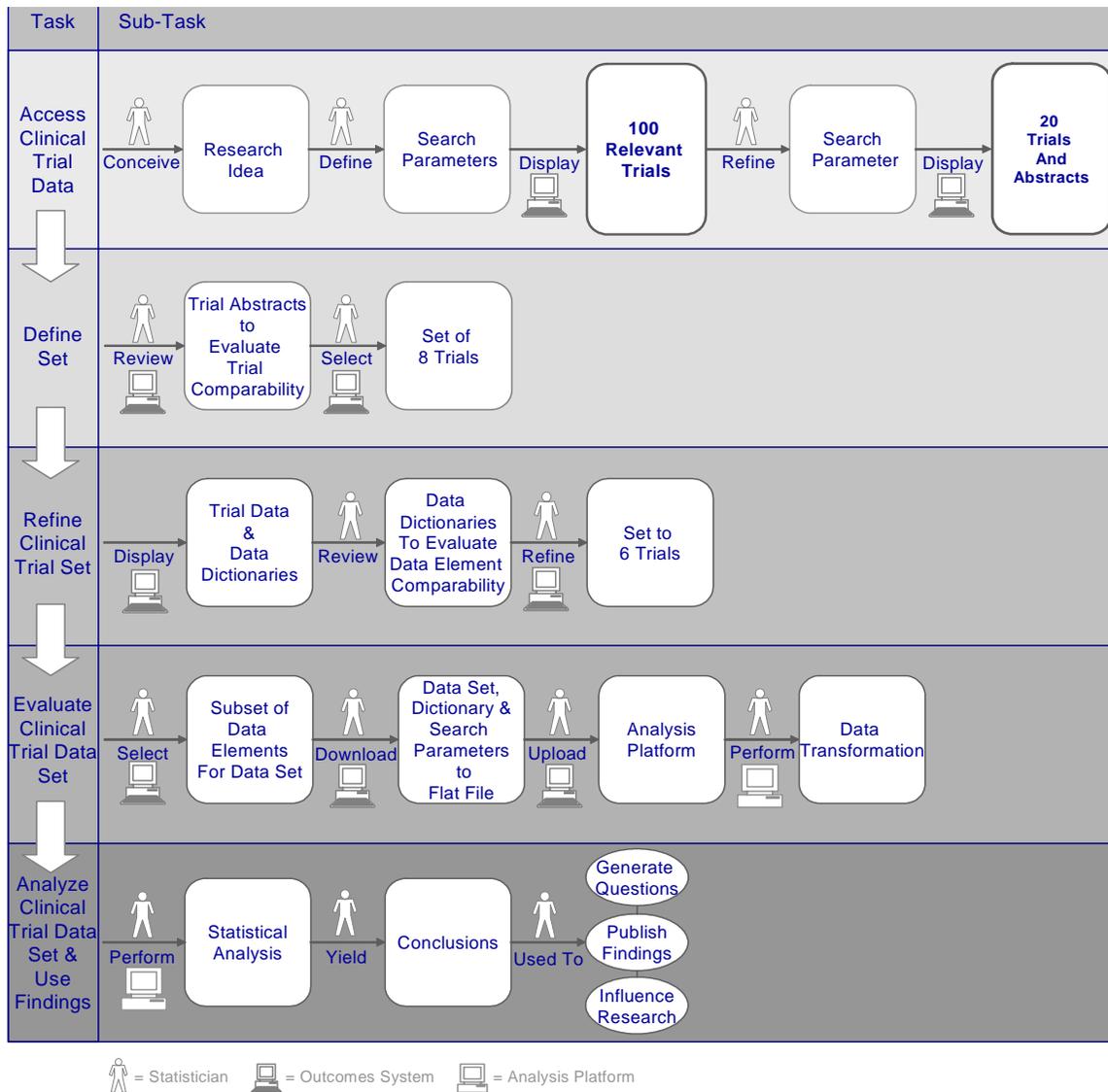


Figure 4. Outcomes Research Scenario: To-Be (Future State)

To-Be Scenario Text

A DCP statistician has been reviewing a data set and wonders if men with early prostate cancer experience complications after their treatment of Proscar have ended. The statistician defines a search parameter to find trials that includes males, prostate cancer, and the agent, Proscar. He inputs this search parameter into the outcomes system. The system returns 100 related studies. Due to the large number of potential trials, the statistician refines the search parameter to include males over the age of 50 with prostate cancer who received Proscar orally. Based on this refined search parameter, the system returns 20 relevant trials. The statistician saves the search parameter in the outcomes System.

The statistician uses the outcomes system to access the related trial abstracts. He reviews the abstracts to evaluate comparability of the trials. The statistician drops a few non-comparable trials and selects 8 trials to investigate further.

The statistician reviews the data elements, definitions, codes and descriptive measures from the 8 trials in the outcomes system to evaluate comparability of the data elements. On this basis, the statistician drops 2 trials to get a refined selection of 6 trials.

The statistician uses the outcomes system to select a subset of the possible data elements for his data set. He downloads his data set, the data dictionaries, and the refined search parameters to a flat file. The statistician then uploads the file to his analysis platform where he performs any desired data transformations (recode, drop data elements, exclude outliers, etc.) on the data set.

The statistician then performs statistical analyses using his analysis platform to answer the question of interest. He draws conclusions from the analyses. The statistician uses the conclusions to generate new questions, publish findings, or share with other researchers to influence new lines of research.

Challenges

Providing Data to the System

Even though statisticians do not see themselves providing data to the outcomes system, they did stress that an automated method of providing data to the system should make provision for manual overrides.

In the initial stages of system development, statisticians shared the following data limitation strategies:

- Store Phase III trial data first. These trials have more participants, larger data sets, and might be the easiest studies to obtain data.
- Begin populating the system with the most recent studies and work backward.
- Do not populate the system with studies that have incomplete data.

However, once data are in the system, they should remain available. Data should never be archived or removed from the system. This may bring up issues about managing data storage, but it was stated that some statisticians have been using the same data sets to answer many different research questions over a 20 year span.

Data Structuring

The statisticians see structuring the data storage as one of the major challenges to developing the system. An example is storing twenty different measurements for the same person on the same day. Are the outcomes stored by the date, or is there a reason to store the fact that they were all taken during the same visit to the physician?

High Level System Requirements

Threshold Requirements	Objective Requirements
Query the system to find clinical trials and patient records	Query the system by ANY variable to find clinical trials and patient records of interest
Select Data Elements for download	Understand and select a subset of detailed and well-identified Data Elements with associated information to clarify the meaning and source of the Data Elements
Perform simple analysis on screen	Download data in multiple formats (XML, comma-delimited, Excel) for further analysis in other systems

Summary of Desired Features

- Versatile Query Interface (List or Pop-Up Menu)
- Access Study Abstracts, Domain Dictionaries and other detailed information
- Universal, unique identifier for each study and piece of data
- Ability to refine a search
- Ability to save search queries
- Perform queries in seconds rather than minutes
- Obtain data in raw form
- Obtain metadata by clicking on data element
- Pre-processing functions (sorting, selecting, recoding)
- Indicate missing data
- Download data set to delimited flat, MX Excel or XML formats
- Incorporate a QA Pedigree
- Provision for manual overrides if automated data providing system

Entries for Domain Dictionary

Recoding: Collapsing data into more general categories

Sample Size: The number of units (persons, animals, patients, specified circumstances, etc.) in a population to be studied. The sample size should be big enough to have a high likelihood of detecting a true difference between two groups.

SAS: Statistical Analysis Software

Selection Bias: An error in choosing the individuals or groups to take part in a study. Ideally, the subjects in a study should be very similar to one another and to the larger population from which they are drawn (for example, all individuals with the same disease or condition). If there are important differences, the results of the study may not be valid.

Statistics: The science and art of collecting, summarizing, and analyzing data that are subject to random variation. The term is also applied to the data themselves and to the summarization of the data.

Statistical Methods: The use of statistics to analyze and summarize data.