

Knowledge Acquisition Session Report

Session Date(s): August 28, 2003

Session Time: 5:00-5:45pm EDT

Session Topic: User Requirements for a Clinical Trials Outcomes System – Medical Oncologist and CCR Manager Perspective

Knowledge Analyst: Bill McCurry, ScenPro, Inc.

Session Location: CCR Offices – Bethesda, Maryland

Type of Session:

Interview

Task Analysis

Scenario Analysis

Concept Analysis

Observation

Structured Interview

Requirements Generation and Analysis

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. The system produced by this project will facilitate the reporting and retrieval of outcomes data.

Due to the large overall 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 the Phase I effort is on gathering specific user data requirements and desired system functionality. Discussion areas of importance include the outcomes data elements to be accommodated by the system, the process by which this data is collected, stored and managed, current and future uses of outcomes data, and existing technologies facilitating reporting and retrieving outcomes data.

Session Goals

- Obtain an understanding of Center for Cancer Research's (CCR) roles and responsibilities in relation to Clinical Trials
- Identify and document CCR's user and data requirements for Clinical Trials Outcome data
- Identify the technologies CCR uses to record, manage and retrieve clinical trials outcome data

Report Summary

- The main data types supporting CCR clinical investigations are Patient Characteristics, Agents, Biomarkers, and Outcomes. These data are collected and stored in a variety of formats, including paper forms and small, internally developed databases. CCR plans to develop a central database to manage its data.
- The current process for gathering outcomes data across multiple clinical trials is manual and lengthy, requiring perhaps several weeks. Major benefits of an Outcomes System would be quicker and easier access to more outcomes data than are currently available.
- Patient-level data provides much more utility than data at an aggregate level (such as by treatment assignment). Many of these desirable data are potentially patient-identifying. It may be possible to strip some data and combine other data in patient records in order to meet HIPPA privacy requirements.
- Metadata standards such as Common Data Elements are designed to mitigate problems of data aggregation across clinical trials. However, Dr. Zujewski is concerned that widespread use of Common Data Elements may be years away.
- Several factors might encourage investigators to provide data for an Outcomes System. These include:
 - It takes little or no additional work on the investigator's part.
 - The investigator maintains control of the data .
 - The investigator retains access to their data.
 - The investigator does not risk having their data used without permission by other people.
- Publications are a source of outcomes data, but it is difficult to standardize and integrate published data because publication formats are inconsistent across the industry. Besides publications, non-standardized sources of useful outcomes might include billing forms and records, tumor registries, and adverse event reports.

CCR Overview

The Center for Cancer Research (CCR) forms the major intramural cancer research arm of the National Cancer Institute. CCR employs experts in both basic and clinical research, facilitating the translational research flow of basic science to the clinical research applications. More than 300 principal investigators drive basic and clinical cancer research, making CCR one of the largest cancer research organizations in the world.

CCR's has over 10 laboratories and branches conducting clinical research as well as the Medical Oncology Clinical Research Unit (MOCRU). The MOCRU supports initiatives in the areas of clinical research, clinical care, and clinical training. Within MOCRU, there are several sections. Dr. Zujewski is the head of the Breast Cancer Clinical Research Section (BCCRS). CCR clinical trials are conducted primarily in the areas of tissue acquisition, natural history, and early drug development (toxicity). Most BCCRS studies are small, no larger than 35 participants, with approximately 65% falling into the category of treatment trials. The remainder of the trials are categorized as prevention, natural history, or 'other.' Some CCR clinical trials will have hundreds of participants.

CCR Outcomes Data

The CCR currently collects and stores data related to clinical trials in a non-standardized manner across the organization. Some records are collected and stored on paper forms while some records are stored in spreadsheets and small, internally developed databases located throughout the CCR. The Center is implementing a centralized database in conjunction with NCI CB to support its clinical study operations and to use Oracle Clinical as the software platform for capturing clinical data. Implementation of this system has begun, but Dr. Zujewski believes it will be over a year until it is fully implemented. Dr. Zujewski believes that a complete modular data system supporting clinical research beyond primary data capture will take several years.

The types of data collected by CCR include patient characteristics, agents administered, toxicities/serious adverse events, biomarkers, and participant survival data. A more complete picture of the data collected by CCR investigators can be found by examining the CCR data elements in the Cancer Data Standards Repository and by reviewing the Theradex Case Report Forms (CRFs) from which many of the data elements were derived.

BCCRS investigators (numbering around 60) currently lack the resources to access and review clinical trial outcome data across CCR studies. That said, it stands to reason that the ability to review outcomes

data across multiple studies and from a variety of organizations seems a long way off. It is believed that the majority of CCR investigators are primarily interested in information related to their own studies; however, Dr. Zujewski estimates that approximately a dozen of these researchers would be interested in exploring outcomes data across multiple studies and organizations.

The Exploration Process

Researchers are interested in utilizing four kinds of data, as shown in Figure 1. This data may be interrelated and used in any manner that is beneficial to the investigator. The exact data needed to support exploration is dependent on the question(s) being explored by the investigator.

Data Types Supporting CCR Clinical Investigations

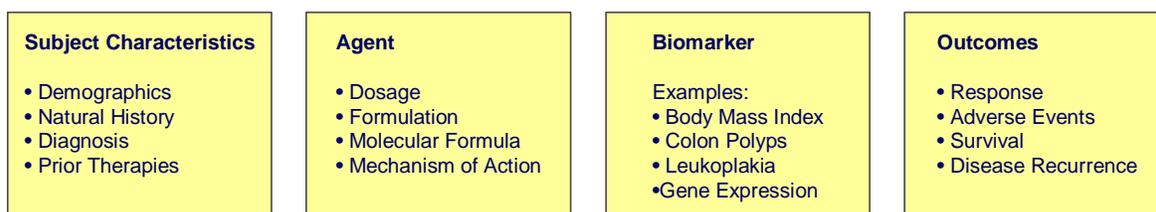


Figure 1. Data Types to Support CCR Clinical Investigations

In some cases, the data of interest are not overtly captured within the outcome records, but must be extracted from a thorough review of the data. For example, odd toxicities or unexpected outcomes are more commonly recorded because they seem significant, but other data may only seem meaningful when aggregated (such as less severe but persistent adverse events). Another example is that agent data will be readily available, but details of the therapies themselves will be more difficult to come by.

Outcomes Data Scenarios

Current (As-Is) Scenario

A CCR breast cancer investigator is talking to a colleague who is an expert in liver cancer. The liver cancer expert mentions that the increased obesity problem in America may be leading to an increase in a fatty liver syndrome, which may increase the incidence of liver cancer

Outcomes Research Scenario: As-Is (Current State)

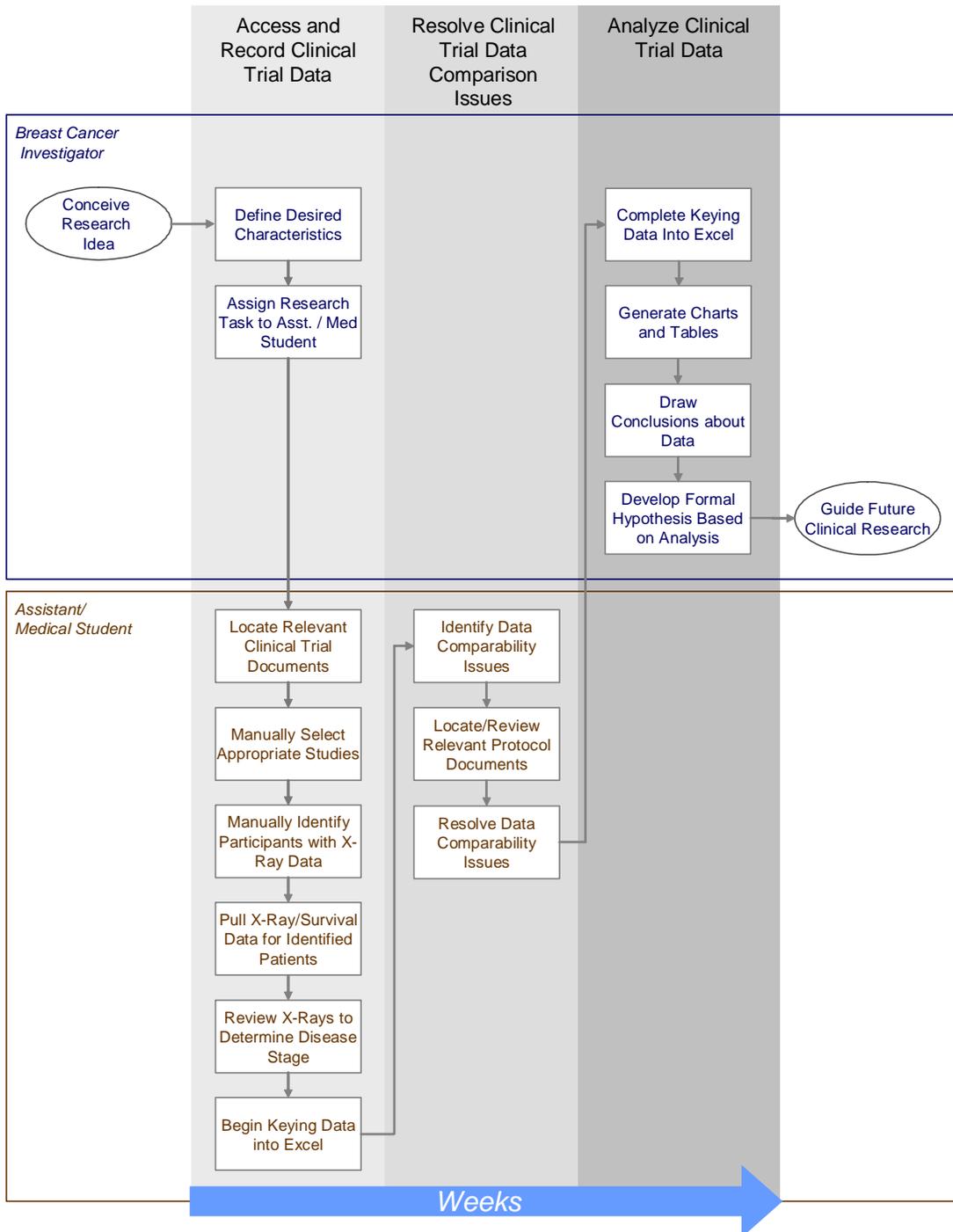


Figure 2. Outcomes Research Scenario: As-Is (Current State)

The investigator starts wondering whether fatty liver is also associated with lower survival among breast cancer patients. The investigator knows that CCR has conducted breast cancer studies and also knows that x-rays were taken as part of some of those studies.

The investigator then manually reviews the records of CCR trials to identify the breast cancer studies. The next step is to further evaluate the trial records and locate studies where X-rays may be available, access the survival data for those participants, and review each x-ray to determine the state of the liver. Because of the time consuming nature of work involved, the investigator assigns an assistant or medical student to support the part of the study process.

The assistant takes several weeks to access and review all the material. The results are then manually entered into an Excel spreadsheet. As the material is collated a number of questions arise regarding the comparability of the clinical trials. The clinical trial protocols are then reviewed manually to resolve any questions regarding comparability.

Once the data have been gathered into a spreadsheet, the investigator reviews the information by sorting in Excel it and creating charts and display tables. The investigator is able to draw positive conclusions about the correlation between fatty liver and breast cancer survival. For further analysis, the investigator saves the Excel file in a comma-delimited text file. The file is then forwarded to the investigator's statistician for in-depth analysis using SAS.

The investigator may use the conclusions as a basis for additional data exploration, or for future breast cancer clinical trials regarding diagnosis, treatment, or biomarker development.

Elapsed time of the entire analysis process: Several weeks.

Future State (To-Be) Scenario

A breast cancer investigator has determined that fatty liver is associated with lower survival among breast cancer patients. A possible underlying cause is obesity, but many breast cancer patients have been treated with Tamoxifen, which could possibly cause fatty liver also.

The investigator determines the need to examine breast cancer patient data from multiple studies involving patients with various relevant characteristics. Some patients should have fatty liver while others do not, and some should have been treated with Tamoxifen while others were not. The investigator also wants to capture the patients' height and weight so body mass index can be calculated. Finally, the investigator determines that survival and disease recurrence data are also needed for these patients.

Outcomes Research Scenario: To-Be (Future State)

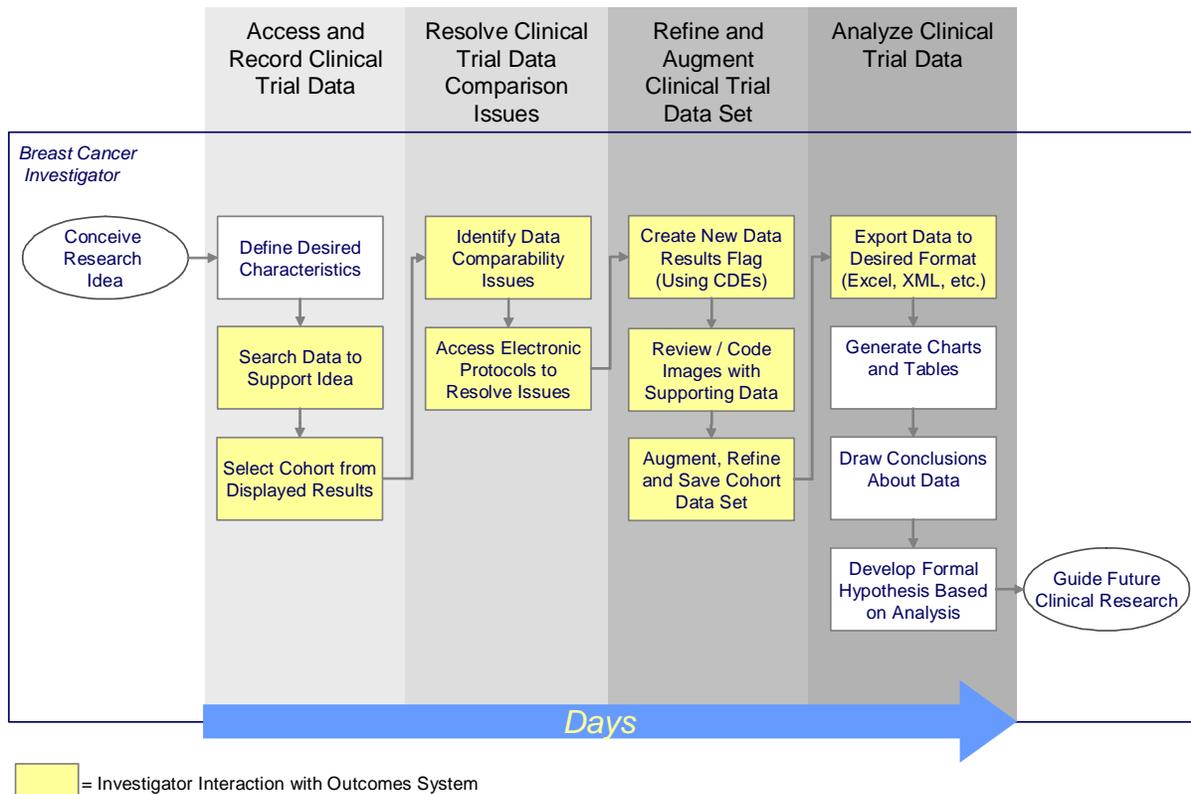


Figure 3. Outcomes Research Scenario – To-Be (Future State)

The investigator accesses the Clinical Trials Outcomes System website. The investigator conducts a search for all breast patient records containing the following data: clinical trial number and type, patient height and weight, agent type, x-ray availability, disease recurrence, and survival rate data. The system searches the available Outcomes data and displays patient records in a list which can be sorted based on the attributes in which the investigator is interested.

The investigator scans through the patient records for relevant data. These are then sorted by various attributes, including clinical trial. Reviewing the data raises a question about the comparability of one of the trials, so the system is used to electronically access specific protocol documents. After reviewing the protocols, one of the trials is determined to not be comparable, and the system is used to drop those records from the current list.

The researcher then uses the system to create a new variable related to the patient records. This variable will be used to flag the presence of fatty liver. The system will create the variable by accessing NCI's Common Data Elements.

The investigator then retrieves each relevant patient x-ray directly through the Outcomes System's imagery storage feature. As each x-ray is reviewed, the investigator codes the new Fatty Liver variable to indicate whether or not the patient had a fatty liver. During this review and coding process, search results are saved using the Outcome System. This allows the investigator to return to the data at a later time and complete the review process.

The investigator then uses the system's sort feature to again review the patient records. The investigator determines that disease stage data are also needed to complete the analysis. The system is accessed to add disease stage data to each available participant record. Patient records that lack disease stage data are stored for future analysis.

The investigator reviews the data online in a spreadsheet format. The system is then used to export the trial participant records to an Excel file on a local machine.

After reviewing the Excel file, creating some simple charts/tables, the investigator decides that Tamoxifen does relate to fatty liver, even in the presence of a high body mass index. At this point additional analysis is needed.

The investigator saves the Excel file in a comma-delimited text file. The file is then forwarded to the investigator's statistician for in-depth analysis using SAS.

The statistician's analysis confirms the relationship between Tamoxifen, fatty liver, and survival of breast cancer patients. The investigator may use these conclusions as a basis for additional data exploration, or for future breast cancer clinical trials regarding diagnosis, treatment, or biomarker development.

Elapsed time of the entire scenario: approx. 2 days

Recognized Challenges

Patient De-Identification

Dr. Zujewski sees much more utility in patient level data than in aggregate data. Countries that have near 100% registration of patients) can do a lot more with outcome data. It is a matter of societal good versus patient control of their information, and the USA definitely leans more towards the side of patient control.

Dr. Zujewski believes that with some effort patient-level outcomes data can be stripped of enough identifiers to comply with HIPPA regulations. An example of this is the way that the U.S. Census

protects individual identity by substituting the socio-economic status of a census tract for the individual's household income.

However, many raw, potentially patient-identifying characteristics remain important to investigators. These include gender, birth year, height, weight, general geographical location, and socio-economic status. There is a significant challenge to the clinical research and technical communities to find ways to include these data while protecting the rights of study participants.

Data Aggregation

Dr. Zujewski sees one of the biggest problems to be how to aggregate patient-level data across multiple trials in a meaningful way. Two factors exacerbate the problem:

1. Investigators and institutions tend to be individualistic in the way they operationalize and capture data.
2. The underlying concepts and the way they are captured change over time as the state of the science changes. (It becomes important to always know the date when the data were collected.)

Metadata standards (such as Common Data Elements in the Cancer Data Standards Repository) are designed to mitigate these problems. However, most trials in the USA do not yet use CDEs to represent data, and Dr. Zujewski feels that widespread use of CDEs is probably a number of years away.

Securing Project Participation

Factors that would encourage investigators to contribute data to a Clinical Trials Outcome System would include:

- It takes little or no additional work on the investigator's part
- The investigator maintains control of the data
- The investigator retains access to their data
- The investigator does not risk having their data used without permission by other people (they want to be the ones to publish from the data and get the academic credit)

Dr. Zujewski firmly believes that investigators will only provide data to the system if it requires little or no additional work for them. This means capturing data from whatever method they already use to provide the data to their sponsors.

Pharmaceutical companies typically require that investigators provide them with the paper case report forms, and the company keys their data into their system.

Electronic remote data capture systems are becoming more common. NCI groups DCP and CCR are all in the process of implementing Oracle Clinical remote data capture systems. NCI is currently capturing some outcomes data through the Clinical Data Update System (CDUS).

However, NCI frequently does not receive all the desired the study outcome data from the investigators. This is because NCI funds the trials but allows the cooperative group to “own” the trials. That is, the cooperative group owns the Investigational New Drug and is the official sponsor of the trial. The funding passes from NCI through the cooperative group to the investigators, and the investigators send the data to the cooperative group rather than NCI. In these cases, NCI must find a way to capture the outcomes data from somewhere in the data submission process.

Investigators will be more likely to provide their data upon or shortly after publication. Problems arise if the study is never published, and there is a bias against publishing studies that were not “successful”. Dr. Zujewski suggested that NCI or sponsors might require that results data be reported with a certain amount of time after publication, or within the end of the study if there was no publication.

Dr. Zujewski pointed out that Dr. Richard Peto, who has convinced breast cancer researchers all over the world to contribute their data to him for periodic meta-analysis that have been published in the British medical journal *Lancet*. (It would be worth researching how he went about convincing his colleagues to cooperate.)

Non-standardized Data Sources

Outcomes data are currently provided to project sponsors in whatever form is specified by the sender. Paper documents are submitted along with transmissions from numerous electronic systems (CDUS, Oracle Clinical and other remote data capture systems).

Publications are also a source of outcomes data. However, it is difficult to standardize and integrate published data because publication formats are inconsistent across the industry. If published outcomes data is included in the envisioned system, it must be flagged in some way to alert the user that they are accessing *published* rather than *original* data.

Dr. Zujewski identified several potential sources of outcomes data that might be integrated with data submitted directly from investigators or collected from relevant publications. These include:

- Billing forms and records
- Tumor registries
- Adverse event reports

Outcomes System Requirements

A number of Clinical Trial Outcomes System requirements were derived from this session. The requirements are categorized as Threshold Requirements (the minimum level of functionality needed to provide utility) and Objective Requirements (the desired functionality).

Threshold Requirement	Objective Requirement
Accessible to internal NCI researchers	Accessible to a widely distributed user base (Internal and External to NCI)
Easy to use for novice computer users	High level of participation/contribution by investigators
Resolve patient de-identification issues	Provide patient-level data
Strategy for ensuring proper data use	High level of participation/contribution by investigators
Provide access to critical data sets	Provide access to critical data sets including: <ul style="list-style-type: none"> • Diagnostic Data • Biomarker Data • Outcomes Data
Ability to aggregate data across studies and time	Aggregate data across studies and time without reliance on the widespread use of CDEs
Perform simple analysis on screen	Download data in multiple formats for casual and expert analysis
Provide access to multiple data sources	Provide access to multiple data sources including those considered to be non-traditional (billing records, etc.)