Title
A Logical Representation for Capturing the Context of Observations and Quantitative Information in Clinical Trial Reports

Author
Tong, Maurine May-Lin

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Logical Representation for Capturing the Context of Observations and Quantitative Information in Clinical Trial Reports

A dissertation submitted in partial satisfaction of the requirements

for the degree Doctor of Philosophy in Biomedical Engineering

by

Maurine May-Lin Tong

2016

ABSTRACT OF THE DISSERTATION

A Logical Representation for Capturing the Context of Observations and Quantitative Information in Clinical Trial Reports

by

Maurine May-Lin Tong

Doctor of Philosophy in Biomedical Engineering

University of California, Los Angeles, 2016

Professor Ricky Kiyotaka Taira, Committee Co-Chair

Professor Alex Anh-Tuan Bui, Committee Co-Chair

Clinical trial experimental studies are the gold standard for obtaining evidence related to interventions for a given disease or chronic condition, and currently results are documented in free-text reports. Due to the current free-text representation, utilizing knowledge from these studies and interpreting results remains an ongoing challenge. This dissertation proposes a bridge representation that transforms information in clinical trial reports from a free-text format to a representation that is computer understandable and capable of assisting answering high level queries from bio-statisticians and clinicians. The objectives of this work are: (1) to specify a representation that will concisely synthesize fragments of information found in clinical trial reports, so users can readily understand the context of numerical data, follow the flow of the study, and assess the quality of the study; and (2) to support queries related to assessment of study quality and estimation of contextualized probabilities derived from various sections within the report (e.g., survival curve, p-values, etc.). The representation is based on a hybrid structure combining several modeling paradigms to create an intuitive and standardized way of describing the conditions of the experiments, the data generated, the analysis methods and the results. Query processing and navigation methods have been designed to operate on the representation to answer common questions related to clinical research, from the clinical and biostatistics side. Such queries include defining the conditions of the patient cohort and interventions, providing context to numerical frequency information, and providing a comprehensive summary of the methods used to compute statistical significance. The focus of the dissertation has been in the clinical research domain of oncology. The dissertation work offers a value-added and time-saving solution to standardizing and organizing information from clinical trial reports and synthesizing knowledge to advance clinical research.

This dissertation of Maurine May-Lin Tong is approved.

Denise R. Aberle

Thomas R. Belin

Gregory H. Leazer

Alex Anh-Tuan Bui, Committee Co-Chair

Ricky Kiyotaka Taira, Committee Co-Chair

University of California, Los Angeles

2016

DEDICATION

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge all who have guided me and supported me throughout my doctoral career. This research was continuously funded by fellowships, instructor positions, and teacher assistantships. I would especially like to thank the NIH for funding my research through the NLM Biomedical Informatics Research Training Fellowship (T15-LM007356) and UCLA.

I would like to thank my committee chair and mentor, Dr. Ricky K. Taira, who has spent countless hours around-the-clock regardless of the day and time helping me develop my research ideas, my presentation skills, and my writing skills. Thank you for making my passions and interest a reality and my career. I would also like to thank Dr. Alex Bui for his generosity with his time and willingness to provide thoughtful criticism at different stages of my research and to share his love of food. I would like to thank my committee members, Dr. Denise R. Aberle, Dr. Gregory Leazer, and Dr. Thomas R. Belin. The insight from our discussions have helped inspire unlimited opportunities for me in this exciting field. Without you, I may have never found my way. Outside my committee, I am thankful to Dr. William Hsu for driving me to expect more from myself and to Dr. Frank Meng for his constant and continual advice and encouragement.

I am extremely thankful to all the students of MII. Thank you, Anna, Mary, Jean, Kyle, Johnny, Edgar, Simon, Panayiotis, Nick, Nova, Eve, Karthik, Shiwen, Justin, Tianran, Jiayun, and the past students. Their willingness to help and genuine interest in all things has been such an inspiration. Thank you for always being there with enthusiasm, invaluable collaborations, and kind hearts. I would like to thank Isabel, Audrey, and Lew for their friendship and support as I went through graduate school at UCLA.

I would not be here now if it were not for my mentors at NIH and at UCLA from early on. Thank you Dr. Ludmila Prokunina-Olsson, Dr. Lori L. Bonnycastle, Dr. Michael R. Erdos, Dr. Francis S. Collins and the FUSION group. Last by not least, I would like to thank Dr. Irwin J. Kurland for taking me under his wing and believing in me during the infancy of my career in research.

Chapter 3, Section 3.4 contains materials published with RK Taira in AMIA Annual Symposium Proceedings. 2012; 2012: pages 1393-402. 2012 Nov 3.

Chapter 3, Section 3.5 contains materials published with RK Taira and W Hsu in Studies in Health Technology and Informatics, 2013;192: pages 856-60.

VITA

2015	University of California, Los Angeles MS in Bioengineering Specialization in Medical Imaging Informatics
2004	University of California, Los Angeles BS in Electrical Engineering Specialization in Biomedical Engineering

Awards

Nov 2014	Finalist, Student Design Challenge, American Medical Informatics Association
Aug 2013	Second Place Student Paper, Medical and Health Informatics World Congress

Professional Experience

2014 – 2016	Department of Life Sciences, UCLA Teaching Assistant
2011 – 2013	Methods International Journal Student Editorial Board Member
2009 – 2013	National Library of Medicine Pre-Doctoral Fellow Biomedical Engineering IDP
2008	Seagate Technologies, Engineer
2005-2007	National Human Genome Research Institute Pre-Doctoral Fellow

Publication and Presentations

Tong M, Hsu W and Taira RK. A Knowledge-based Representation of Clinical Trial Reports for Evidence-based Decision Support. RSNA Scientific Assembly and Annual Meeting, Bioinformatics Exhibit. Chicago, IL. Nov 2014

Tong M, McNamara M. A Patient Portal for Clinical Trials: Towards Increasing Patient Enrollment. 2014 American Medical Informatics Association Student Design Challenge: "Beyond Patient Portals: Engaging Patients with their Healthcare Providers." Washington, DC. Nov 2014.

Tong M, Hsu W and Taira RK. A Formal Representation for Numerical Data Presented in Published Clinical Trial Reports. Studies in Health Technology and Informatics. 2013;192:856-60.

Tong M, Hsu W and Taira RK. A representation for standardizing numerical data from clinical trial reports. Radiological Society of North America Scientific Assembly and Annual Meeting, Bioinformatics Exhibit. Chicago, IL. Nov 2012.

Hsu W, **Tong M**, Taira RK, Bui AAT. Visualizing evidence in biomedical literature: Integration and application of clinical, imaging, and genomic findings reported in research studies. RSNA Scientific Assembly and Annual Meeting, Bioinformatics Exhibit. Chicago, IL. Nov 2013.

Tong M, Taira RK. Improving the accuracy of treatment descriptions in clinical trials using a bottom-up approach. AMIA Annual Symposium Proceedings. 2012; 2012:1393-402. Epub 2012 Nov 3.

Tong M, Wu J, Chae S, Chern A, Speier W, Hsu W and Taira RK. A Tool to Formalize Information from Clinical Trials for Disease Modeling. RSNA Scientific Assembly and Annual Meeting, Bioinformatics Exhibit. Chicago, IL. Nov 2011.

Hsu W, **Tong M**, Wu J, Lin M, Bui AAT and Taira RK. Tools for Modeling Medical Imaging and Molecular Biology Correlates Using Published Literature. RSNA Scientific Assembly and Annual Meeting, Bioinformatics Exhibit. Chicago, IL. Nov 2011.

Tong M, Wu J, Speier W, Chae S, Chern A, Hsu W and Taira RK. A Knowledge-base to Represent and Visualize Information from Clinical Trial Literature. NLM Informatics Training Conference. Denver, CO. Jun 2011.

Tong M, Wu J, Chae S, Chern A, Speier W, and Hsu W, Bui AAT, and Taira RK. A Tool to Utilize Information from Clinical Trials: A Knowledge-Based Approach Incorporating Signaling Cascades. RSNA Scientific Assembly and Annual Meeting, Bioinformatics Exhibit. Chicago, IL. Nov 2010

Chapter 1 - Introduction

Advancements to evidence-based medicine have benefited from and are guided by rigorous scientific investigations such as randomized controlled trials (RCTs) [Wood 1999, Eisenberg 1999]. Clinical trials are regarded as the best approach to providing the most unbiased assessment regarding the efficacy of an experimental therapy or diagnostic procedure [Horowitz 1987]. Knowledge gained from clinical trials have the potential to improve our understanding of the causal nature of interventions and hence is a primary means of gathering scientific knowledge to drive developments related to disease characterization. Ultimately, these models will be used as an inferencing source for precision medicine applications [Chen 2013]. The translation of results from RCT experiments to patient care and/or disease models, however, is not straightforward. Some issues include: difficulties applying results from a population based study to an individual patient context; uncertainties associated with the assessment of the quality of a research study, especially in regards to conflicting studies; and ambiguities related to the interpretation of numerical data to correctly characterize, for example, observational frequencies.

The general problem addressed in this dissertation is a structured knowledge representation of clinical trial study results as reported in the primary literature. The main driving queries relate to investigating study quality and navigating context for numerical information. Table 2-1 of Chapter 2 provides a comprehensive list of intended queries to be answerable by the representation. The significance of the work includes: 1) Elucidation of relevant information contained in free-text publications toward improving patient care is a significant endeavor for the modern physician. Urick et al. point out that physicians and researchers must spend a significant amount of time and

have sufficient research training to appropriately integrate RCT study results into medical practice [Urick 2005]; 2) There are no consistent templates that allow reviewers of an RCT paper to quickly navigate to relevant information regarding study design, context in which data are collected, and the precise data and methods used to calculate statistical significance; and 3) Informaticians building models of diseases based on probabilities (e.g., Bayesian methods) require a precise understanding of what frequencies (and associated conditional probability estimates) are being reported. Without such context, the probabilities can be interpreted erroneously, resulting in models that mislead clinical decision-making tasks.

The problem of how to formalize information contained within a clinical trial study is not new. Most efforts have been motivated by patient recruitment applications (i.e., *which clinical trials does my patient qualify for?*) and/or information retrieval tasks (e.g., *which clinical trials have studied this disease with this drug?*). Representation issues have advanced along three themes: (1) Generating a checklist of required fields for characterizing a study [Schultz 2010], (2) standardization of terms and ontologic concepts [Sim 2000], and (3) management of study conclusions [e.g., Research Maps – Silva 2015]. Although these efforts have made strides towards standardizing certain informational aspects of a study, characterization has been mainly along the lines of concept indexing and/or high level propositional description that are inadequate for capturing a synopsis for a researcher and/or an evidence-based medicine practitioner.

1.1 Overview of the Dissertation

A large amount of effort and money is spent worldwide on conducting RCT studies. Research hypotheses are the heart of scientific endeavors; the accurate, unambiguous and operational representation of these hypotheses is vital for the formal assessment, synthesis and application of such investigations.

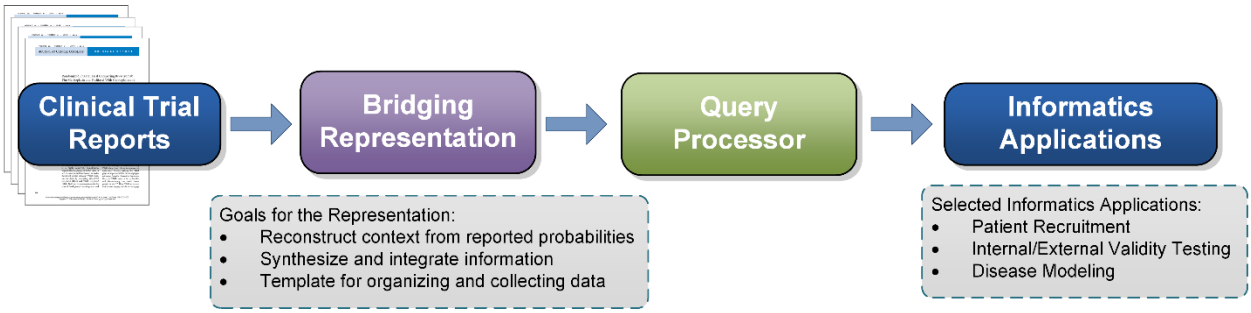


Figure 1-1. Overview of dissertation work

The main objective of this dissertation relates to the development of an improved representation for information presented in clinical trial studies. Two central issues are addressed (Figure 1-1): (1) How to formally represent the specific details relevant to "current best evidence" and study quality in a computer understandable format; and (2) Given this representation, what queries can be executed to support patient-specific evidence-based medicine and/or disease modeling inquiries. This work addresses the development of a more principled means to represent and assess quantitative evidence as presented in the clinical trials literature. The target users of the system are physicians and researchers, including clinical researchers, basic scientists, and informaticians.

Thus, the two specific aims researched in this dissertation are as follows:

Specific Aim 1
To specify a logical representation to concisely synthesize fragments of information found in clinical trial reports, such that users can readily understand the context of numerical data, follow the flow of the study, and assess the quality of the study.

Specific Aim 2
To provide a consistent template visualization and query processing engine to support inquiries drawn from the research paper related to concerns of clinicians who are interested in evidence-based medicine and/or biostatisticians who are assessing the quality and/or context of reported numerical information (e.g., observations, frequencies, probabilities, survival curves, and p-values).

The main hypotheses of this work are as follows:

H1.1 - The representation will be able to express in a logical form, a sufficiently detailed synopsis of the research paper for the purposes of clinical and research applications.

H1.2 – The representation will be intuitive to understand for the intended user base (i.e., clinicians, statisticians, and researchers).

H2.1 – Users will be able to answer specific targeted study questions on a paper more accurately using the proposed representation compared to the paper-based representation alone.

H2.2 – Users will be able to answer questions in a shorter amount of time compared to the status quo representation.

1.2 Specific Aim 1 – Representation

The traditional way of presenting knowledge in scientific papers has many limitations. The most important and obvious of these is the use of natural language, albeit augmented by various formalisms and mathematics. Natural language is notorious for its imprecision and ambiguity. Contextual information for proper interpretation of numerical information can be scattered in various locations within the text document, making it difficult to recall vitally connected pieces of descriptions. Many readers take away only crude summary information (e.g., p-values) and "throw away" essential related information such as sampling data, experimental methods, intervention details, and analysis descriptions. Documentation of clinical trial studies via natural language thus is a barrier in maximizing the use of computers to store, navigate, analyze, and integrate the results of disease-related clinical trial studies.

In this dissertation, I introduce a hybrid representation that utilizes components of process modeling and data spreadsheets. Hyperlinks between nodes in the process model and columns in the data spreadsheet allow information related to any variable described within the study to be linked to the experimental steps leading up to the collection and/or constraining of a variable. Rows in the spreadsheet map back to any relevant variable presented in the research paper. A cell in the spreadsheet of the representation describes a value, summarization, distribution, or data point collected for a given variable at a given node in the process model. Cell value types can be nominal, ordinal, descriptive or numerical. The representation standardizes the study properties using various publicly available ontologies, but is flexible enough to allow user supplemented entries in the cases of incompleteness. By hyperlinking the process model to the spreadsheet, data stored within the cell of a spreadsheet can reference back to the experimental step within the

process model to which this information was obtained. The expressivity of the representation was tested on a sample of RCT research papers from the domain of oncology, including studies from different RCT phases and mechanisms of intervention. A qualitative evaluation of the representation as an efficient medium of human expression [Davis 1993] was performed to assess how diverse users viewed the intuitiveness of the representation and how easily one could discuss aspects of RCT papers within the proposed representational framework. To narrow the scope of this dissertation, a mixture of research papers related to clinical trials within the domain of lung cancer were explored.

1.3 Specific Aim 2 – Query Answering

The second specific aim of this dissertation addresses methods for performing operations on the representation developed in SA-1 and providing an interface/visualization for users. Traditionally, researchers who have developed structured representations to summarize RCT studies utilized strictly relational data models, such as a standard SQL-like query processor. Current representations, however, do not capture experimental pathway information related to how data are collected, processed or constrained in the context of the overall experimental design. In this dissertation, I introduce a backtracking algorithm that can trace the context of variable assignments as defined by the study design flowchart component of the SA-1 representation. The nodes visited by the backtracking algorithm link to columns within the spreadsheet component of SA-1 and the combination of nodes and linked columns are used to infer the context of all other variable states. Details of this backtracking algorithm are presented in Chapter 4.

The query interface/visualization for the system closely resembles the underlying representation. Most users are comfortable interpreting flow charts and spreadsheets. One of the significant aspects of the work is the consistent layout of how information (i.e., design flow and specification of data and analysis methods) can be applied to RCT research papers in general. I sampled a number of published studies and found that almost all important aspects of the studies could be represented using the template layout of the system design. Importantly, this allows users to be conditioned to expect certain pieces of information spatially arranged on the layout and to have a standard format to assist in navigating to information of interest. An evaluation of how this aspect of the representation improved users' ability to answer questions related to a specific paper in a timely manner was conducted and reported in Chapter 5.

1.4 Summary of Contributions

The main contributions of this work are as follows:

1. A rich representation based on a hybrid combination of a process model and spreadsheet that systematically organizes descriptions of properties within the context of experimental design steps.
2. A query answering system that utilizes a backtracking algorithm within the process model to infer context for cells in the spreadsheet portion of the representation;
3. A consistent templated visualization for presenting and querying information allowing users to rapidly learn how to search for and navigate to desired information.

Taken together, these contributions attempt to: 1) structure heterogeneous information from clinical trial reports with the necessary context; and 2) locate and visualize answers for common

queries related to study quality and interpretation of quantitative information. The ability to quickly navigate papers is vital for helping researchers and physicians understand and act on available information.

1.5 Organization of the Dissertation

The remainder of the dissertation is organized as follows:

- Chapter 2 provides background on issues related to developing representations for clinical trial reports and summarizes the important literature related to this topic. Current trends in logical designs and descriptions of gaps between information needs and research efforts are identified.
- Chapter 3 documents the methodologies used to design the SA-1 representation. It includes an investigation of user needs, adaption of ontologies, and rationalizations and limitations of design decisions.
- Chapter 4 discusses the methodologies used to perform SA-2 query answering. It includes descriptions of the backtracking algorithm and the standardized visualization template.
- Chapter 5 discusses evaluation methods for the hypotheses of the dissertation work. The main evaluation is based on a modified crossover design to test the value-added by the dissertation work as compared to the status quo paper representation. Descriptions and results of qualitative evaluations of the interface are also presented.
- Chapter 6 provides a summary of the dissertation and discussion of the results and compares these results to capabilities of similar work in the field of medical informatics. The chapter

concludes the dissertation with a discussion of limitations of this research and potential areas of future expansion.

Chapter 2 - Background and Significance

In this chapter, I review the following: the motivation behind and user groups involved in evidence-based medicine (section 2.1), the difficulties in practicing evidence-based medicine (section 2.2), prior research on requirements for modeling clinical trial information (section 2.3), the current state of evidence (section 2.4), prior and related work in structuring clinical trial reports (section 2.5), and lastly, Bayesian modeling requirements (section 2.6).

2.1 Overview of Evidence-based Medicine

Evidence-based medicine (EBM) ideally requires that healthcare professionals make "conscientious, explicit, and judicious use of the current best evidence" in their everyday practice [Sackett 96]. This requires a comprehensive awareness of the relevant literature and the ability to retrieve, interpret and apply the results of the most appropriate scientific information within the context of the current patient case. Evidence-based medicine is an endearing concept for those working in the healthcare field and its practice should be guided by rigorous scientific investigations such as clinical trials [Wood 1999, Eisenberg 1999]. Elucidating the contents of these clinical trial reports can help inform clinical guidelines and provide healthcare procedures tailored to individual patients [Urick 2005]. However, utilizing information effectively to provide evidence accurately requires a significant amount of time, expertise and research training.

The current state of evidence-based medicine is challenging to comprehend and it is often difficult to directly apply evidence to individual patients in practice. One attempt to better define the issues needed to deploy evidence-based medicine practice is the five-step process proposed by Sackett et

al.: (1) define a clinically relevant question, (2) search for the best evidence, (3) appraise the quality of the evidence, (4) apply the evidence to clinical practice, and (5) evaluate the process [Sackett 1997]. While this list contains general guidelines for EBM, steps are vague and require considerable expertise to follow. The second step in the list, for example, requires a search for the best evidence; and these steps collectively pose a more fundamental question: how is the best evidence defined? As is common in clinical research, evidence is found in a number of sources (i.e., anecdotal experiences, case-studies, clinical experiments, systematic reviews, meta-analyses). The search for relevant papers is limited by search engines as physicians often prefer to review a handful of reliable sources of information rather than try to locate all the available medical evidence [Hoogendam 2012]. The third step in the list requires an appraisal of the quality of the evidence. However, the process does not give specific instructions on how to assess study quality. As a result, most clinicians have honed their own personal approach using their own critical-thinking skills to assess medical literature rather than to use an organized systematic approach [Steves 2004]. In fact, there are no precise metrics for appraising the quality of a study's findings. The fourth step is to apply evidence to clinical practice, but this aspect has not been well addressed in the clinical community. It lacks standardized yet personalized solutions in regards to the complexities associated with, for example, integrating evidence from conflicting clinical trial studies, or extrapolating results when a patient has different eligibility criteria. Because interpretation of conclusions and assessment of patient applicability for a given research study are clinician dependent, there is often an arbitrary application of methods that can be inferred from indications of the same research studies. The Cochrane Library provides a partial solution to these problems by supplying resources based on coordinated efforts to conduct and collect reviews on specific topics in medicine [Jadad 1998]. Despite the assistance with searching for and appraising

evidence, the context provided by the Cochrane Library needed by an experienced biostatistician to carefully assess each study is not typically provided with the appropriate level of detail and may be missing from the original source. Thus, while evidence-based medicine can enhance the scientific foundation upon which an upward improvement on healthcare can follow, steps to practicing evidence-based medicine are currently not straightforward and, in particular, the step requiring clinicians to appraise the quality of the report is unspecified.

To assist with practicing evidence-based medicine, this dissertation focuses on two groups of users: clinical practitioners and biostatisticians.

- Clinical practitioners must search through literature to identify relevant information for a patient of interest. After searching through literature, clinicians must read RCT papers then apply the evidence gathered and tailor it to their patient at hand.

- While biostatisticians are not directly involved in evidence-based medicine, they are instrumental in assessing the quality of a clinical trial study that are read by clinicians. Biostatisticians may read RCT papers for several reasons, including to determine if the correct statistical test is used, or if the analyses documented are replicable and were performed correctly.

2.2 Current Issues and Assessment of Needs

The randomized controlled trial (RCT) is a type of scientific experiment and is the most reliable method for ascertaining evidence in healthcare [Sackett 1996, Pearl 2000]. It is used to explore a causal hypothesis, carefully controlling for selection bias and spurious causal factors. The RCT

provides strong evidential basis for licensing and performing new procedures, and administering new medications and is required for regulatory authority approval. Furthermore, the evidence supports theories and best practices covering treatment, prevention, diagnosis, screening, and quality of life.

Its benefit prompts a great amount of effort and money spent worldwide on conducting RCTs studies, while ensuring patient safety and acquisition of high quality evidence. Despite efforts and money spent, there remains a disconnect between the acquisition of knowledge (i.e., the testing and validation of a clinical hypothesis) and the application of this knowledge (i.e., evidence-based medical practice). Several issues systematically undermine the ability to use research to either accelerate existing related research topics or apply the knowledge gained from such trials to individual patient cases. In this section, I elaborate more on these issues.

2.2.1 Issue 1 - Volume and Diversity

There exists a large base of literature in the biomedical sciences related to the testing and validation of new clinical theories. In all areas of medicine, this body of evidence needs to be assimilated, understood, and tracked by physicians. For example, in the area of lung cancer, there are over 6580 clinical trial studies registered in ClinicalTrials.gov from 2000 to 2016 covering diverse topics such as molecular agents, radio/chemotherapies, imaging, genetics, and psychiatry. The large number and variety of topics make it difficult for any single physician to stay up-to-date on all relevant findings in a given clinical area. Thus, most physicians struggle to keep up with their understanding of the benefits and harms of reported technological advances. The contributions of this dissertation towards addressing this issue is the development of a representation that could be

instantiated once for such papers, and subsequently be used as an informational surrogate for addressing important classes of commonly asked questions related to evidence-based medicine.

2.2.2 Issue 2 - Understanding the Statistics

The scientific community demands authors (i.e., writers of RCT papers) to utilize sound and formal mathematical models to characterize their data and ultimately to provide evidence for or against the given RCT hypothesis. Physicians (i.e., readers of RCT papers), however, often are uncomfortable with and do not have the background to fully interpret the mathematical descriptions of the study results. A study by McAlister et al. suggests physicians remain uncomfortable with the quantitative approach to medical practice [McAlister 1999]. [West 2007] reports that only 17% of surveyed clinicians believed their training in biostatistics was adequate for their needs to conduct their own statistical analysis. These statistics also extend to medical residents, as a 2007 study reported in *JAMA* from Yale University showed that 75% of medical residents in training do not understand the statistics used in the medical literature [Windish 2007]. In combination with the physician's lack of background knowledge, statistical tests published in clinical trial papers are increasing in complexity. [Horsfield 2005] conducted a study investigating statistical methods used in 311 research articles published by the *New England Journal of Medicine*. The authors discovered that physicians with a basic knowledge of statistics (i.e., *t*-tests, contingency tables, simple linear regression) would only be able to interpret 21% of the articles sampled due to an increasing use of advanced statistical methods in current studies. In a similar study, [Hellems 2007] reports that pediatric residents are not receiving sufficient statistical training. An analysis of 171 articles published in 2005 for *Pediatrics* reveals that pediatric residents with knowledge of the 10 most common statistical concepts (i.e., *t* test, χ^2 test) would be

unfamiliar with more complex statistical methods present in 53% of the journal articles. The contributions of this dissertation towards addressing this issue is the development of a framework that provides hyperlinks to data and trial execution steps to inform readers how particular statistical tests were selected and their calculations computed. When interacting with the framework, the data and trial execution steps, the appropriateness of statistical tests and/or from which execution paths within the study design data can then be judged by the user.

2.2.3 Issue 3 - Difficulty Assessing the Quality / Contribution of Research Paper

Even if a physician is relatively comfortable understanding the statistics stated within the conclusions of a given research study, there is a more general issue related to assessing the scientific quality of the study and its contribution to understanding the targeted disease. There are two sides related to objectively assessing the quality of the research paper: the writer and the reader.

Writers must accurately and unambiguously report research findings to prevent misrepresenting information. [Ioannidis 2005a] comments that most research findings reported in the literature are not entirely accurate in the conclusions they draw and readers can easily be deceived by the conclusions or the strength of conclusions for a given study. They cite methodological faults related to experimental frameworks (e.g., follow-up confirmation studies), bias (e.g., selective reporting, conflicts of interest, faulty randomization), lack of independent teams, and lack of statistical power (low number of samples and large state spaces). Additionally, the use of imprecise and ambiguous natural language as the representation for documenting scientific results has been shown to often mislead readers. [Hyland 1998, Hyland 1996, Light 2004, Roland 2007]

report on the frequency, type and effects of linguistic "hedging" that can influence a reader's perceived assessment of a scientific study. Hedging allows writers to express a perspective in their statements, and is an expression of tentativeness, possibility perspective or deference to the reader. It contrasts with factual language, but has been demonstrated as an effective means of gaining the reader's acceptance of a claim, possibly misleading readers.

Readers, on the other hand, must perform their own critical appraisal of the evidence and assess its scientific merit. A reader can be misled to assume that a finding is correct and could be used in practice simply because it was the subject of a research study published in a reputable journal. A more experienced reader can still be misled, despite careful reading if the article does not clearly depict the appropriate context. An important consideration, when appraising evidence, is the surrounding context under which a specific measure and/or observation is made [Mills 2012]. Context describes the conditions of the experiment and is important to rule out alternative explanations for observed effects as well as to guide appropriateness for certain generalizations. Reported findings, for example, may only be valid under certain conditions (e.g., a specific patient population), which may or not be obvious to the reader. Statistical quantities such as p-values alone can be misleading to clinicians unless readers know the characteristics of the populations that were tested, among other factors [Goodman 1999]. Finally, trying to reconcile and compare results from similar but heterogeneous experimental methods is a non-trivial task, even for highly proficient domain experts [Levin 2001]. For example, Simpson's paradox is a cited phenomenon regarding comparing statistics and confounding variables. Objectively, conclusions from even highly cited research studies have been seen to be contradicted or to demonstrate stronger effects than reported in the clinical research literature [Ioannidis 2005b]. The contributions of this

dissertation towards addressing this issue is the development of a framework that provides essential context for assessing statistical significance (i.e., type of hypothesis, sample size, test statistic, etc.).

2.2.4 Issue 4 – Difficulty Translating Scientific Findings to Clinical Practice

[Sackett 2000] defines EBM as, "the integration of best research evidence with clinical expertise and patient values for the task of clinical decision-making." Most physicians believe in and aim to practice evidence-based medicine, however, their abilities to determine the intervention and its circumstances that would provide the most benefit for a patient given his/her specific profile is highly variable. The fundamental disconnect lies in part in differing goals between the purpose of a research paper and what physicians seek in a paper. Research papers are designed to mainly explain the dynamics of a cause-effect relationship of a single research hypothesis for a pre-specified population, and is not necessarily designed to explain how to apply these findings to individual patients. The focus, the language, and motivation of the literature are science-oriented, rather than application-oriented. Thus, a large part of a scientific paper often describes hidden theoretical variables (e.g., biological parameters) that are often not routinely observable in clinical practice.

Additionally, a large part of the write-up of these RCTs is devoted to methodology and procedural setup related to a relatively controlled environment, which can be vastly different compared to operations of a routine clinical environment and the population to treat. Decisions as to what extent a clinician may generalize research findings for his/her patient is often unknown and applied in an ad hoc fashion. A complete understanding of how to diagnose, treat and/or manage a disease

may require a more comprehensive understanding of the complex causal chain and interaction dynamics of a disease process. RCTs often focus on a single proposition related to a disease, providing only a partial piece of the whole view. Physicians reading the article however, often then base decisions on this partial piece of the picture [Haynes 2007], this increases the difficulty in translating scientific evidence to practice. Thus, developing methods to integrate fragments of scientific knowledge into a more comprehensive mental model of a disease is a necessary, but complex challenge. Lastly, outcomes and results of RCTs are reported at different levels of detail and different pathways to effects (e.g., smoking causes cancer versus tar deposits on lung cause cancer). The more general the claim, the more straightforward its application appears. Physicians may become frustrated with understanding the scientific literature and eventually abandon the approach of EBM as an "ivory tower" concept. The contributions of this dissertation towards addressing this issue is the incorporation of a frame-based ontology for representing populations and intervention details (e.g., drug administration details) that can allow improved matching and improved assessment of expected outcomes.

2.2.5 Issue 5 – Access and Speed / Time Constraints

Finally, an EBM system must operate within the time constraints imposed by the workflow of a typical medical office. The original model of evidence-based medicine presented in 1992 in the *Journal of the American Medical Association* can be paraphrased as follows [Moher 1992]:

- A clinical question would arise at the point of care, and the physician would conduct a literature search yielding multiple (sometimes hundreds of) articles.
- The physician would then select the best articles from the results, evaluate the research,

determine its validity and decide what to do - all while the patient waited in the exam room.

In reality, this scenario does not happen due to time constraints and is impractical in a busy medical office. Currently, physicians are burdened with too many patients, and an aging geriatric population [Bodenheimer 2006]. Physicians are limited with the amount of time they can spend with each patient, the average being 10 minutes [Tai-Seale 2007, Uner 2013]. Furthermore, predictions show that physicians are expected to do more with less time. Colwill et al. predicts that population growth and aging will increase family physicians' and general internists' workloads by 29 percent between 2005 and 2025. Colwill et al. expects a 13 percent increased workload for care of children by pediatricians and family physicians [Colwill 2008]. Additionally, patients may feel uneasy about the confidence level of a physician who spends some time reading about their condition from an article or summary during the office visit. The contributions of this dissertation towards addressing this issue is the development of a consistent template visualization that should allow experienced users familiar with the representation to quickly find desired information.

2.3 The Need for Context in Clinical Trial Representations

Reported findings, are dependent on the surrounding context (e.g., specific patient population) under which a specific measure and/or observation is made. The accuracy of statistical quantities (e.g., p-values) are based on gathered data. These claims are illustrated with a running example from a typical clinical trial result:

> *"Bevacizumab in combination with carboplatin and paclitaxel improved overall response and time to progression in patients with advanced or recurrent non-small cell lung*

cancer.... Survival for the high dose bevacizumab was modestly longer than the control arm (17.7 vs 14.9 months; p=0.62)." [Johnson 2004]

To understand the benefit of this intervention, various types of context need to be considered before results can be trusted by the user. Several types of biases need to be investigated, involving questions such as: What were the details of the participant population? What data variables were used? How was the data collected? How were known confounders addressed? What was the test statistic used? What are the assumptions of the statistical test used to determine the p-value? Was the experiment adequately powered to test this hypothesis? What was the formal hypothesis used to test bevacizumab? What was the sample size? In general, the results of the clinical trial can change if the hypothesis or the conditions of the experimental procedures or analyses are different. In this section, I summarize the type of context needed in order to interpret clinical trial results and the importance of each type.

2.3.1 Study Participants

A description of the study participants is necessary to interpret clinical trial results as alterations in criteria may lead to differing conclusions. Significant contributors of variations in healthcare outcomes are due to racial/ethnic backgrounds and among biological, environmental, or social differences in causes of disease [Taylor 2005]. For example, [Adams-Campbell 2004] report that African Americans have the highest mortality rates and poorest survival from cancer compared with other ethnic groups. [Mosenifar 2007] urges the inclusion of the elderly in clinical trials as older age is an important issue for critical illnesses, especially respiratory diseases. Patient populations of different racial groups, age groups, or even a different ratio of males to females can influence results. Successful randomization, including a description of the controls, are also

necessary to interpret the results of an experiment [Holland 1986, Rubin 2011, Rubin 1975] by providing knowledge about the expected behavior in the absence of experimental procedures. If an error was to occur in the experiment, controls can help pinpoint this error. When control groups are used, context is needed to show that the study groups are initially equal and comparable at baseline, to ensure that one does not use partially or inherently heterogeneous data material [McCance 1995].

Details of participant flow are also necessary, such as when patients enter and leave the trial. With an intention-to-treat methodology, it is critical to report all subject dropouts carefully and truthfully. Failure to include all participants in the analysis may bias the trial results. For example, a study may have dropout rates that differ between treatment arms, so that fewer patients are followed up in one arm than the other [Bell 2013]. This is called "differential dropout." While this may not be alarming, context is needed to fully investigate this situation. Equal dropout rates between treatment arms do not imply that estimates of treatment effect are unbiased, and unequal dropout rates do not imply that estimates are biased. Instead, bias depends on the type of "*missingness*", the analysis method, and the effect that is being estimated. Thus, to associate different adverse effect profiles to a differential dropout, one needs to assess the cause of the different dropout by clearly documenting the patient profiles and potential biases. In summary, patient population profiles, the control population profile, and participant flow can influence interpretation of clinical trial study results and are needed to provide the appropriate context.

2.3.2 Experimental Procedures

Differences in experimental procedures can affect data and the study conclusions. The justification for the analysis lies strongly in the manner in which the data were collected. Therefore, it is important to clearly define and report the course of the experimental procedures along with change to these procedures [Smyth 2011]. Doing so would allow all methods to be replicated and the errors in measurements to be fully addressed. Even with accurate and defined methods, reporting is still not complete. Replicating treatments in practice depends on how well these procedures have been documented in research studies [Glasziou 2008]. In addition to replication, experimental procedures must be documented with appropriate context to assist with assessing inherent error. Certain experimental procedures can have unknown inherent error, affecting the results of a clinical trial. A 2001 article examined the effects of measurement error on therapeutic equivalence trials and reported that measurement errors inappropriately favor the goal of showing treatment equivalence [Kim 2000]. Such measurement errors can harm the evaluation of a new method of treatment and falsely prove it is better than the old method; or the opposite can be true where measurement errors can harm the evaluation of a new method of treatment and falsely prove it is equivalent to or as good as the old treatment. Thus, the context needed for experimental procedures includes the variables, when and how data were collected, and the error associated with each measurement.

2.3.3 Statistical Analyses

While it is universally accepted that context is needed when interpreting numerical outputs from statistical analyses, the question remains what is the necessary context. The selected test statistic

used in the clinical trial can have a significant effect on trial conclusions. It is directly tied to the hypothesis of the trial as the hypothesis is formulated in terms of the parameter space of the test statistics used [Berger 1987]. It is critical that an appropriate statistical methodology be selected and corresponding considerations in the trial design be implemented to objectively analyze the data. Because it is not uncommon that the data collection plan changes for unexpected reasons, it is important to adjust the statistical analysis accordingly. The context for statistical analyses requires a description of the analyses, the parameters of the statistical tests, and the assumptions made in the analyses.

2.3.4 P-value

The p-value provides a measure of the significance for the results of a clinical trial study. The p-value is defined as the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed during the trial. The less likely this is to occur, the lower the p-value, and the stronger the evidence is that the treatment actually did have some effect. While the p-value provides valuable information on scientific conclusions, there is the mistaken idea that a single number (e.g., $p < 0.05$) can capture scientific conclusions [Ioannidis 2005a]. Although the basic definition of the p-value in terms of a tail-area probability density is straightforward, its interpretation in terms of strength of evidence to support/refute a given scientific hypothesis (i.e., the decision rule) is subtle [Hubbard 2006, Hubbard 2008] and clouded by a number of confusing issues and implicit conditions. In 2007, a review of the literature was published that cataloged and described 47 specific statistical mistakes that are commonly made in the medical literature [Strasak 2007]. Results of medical research should not be reported as "significant" or "non-significant," but should be interpreted in the context of other evidence,

and along with possible biases or confounding factors [Sterne 2001]. Thus, interpretation of statistical significance requires a number of contextual details, including: (1) the type I error, or the level of significance, called the α-level, which is usually set to 0.05; (2) the exact statistical test methods; (3) the type II error, called the β-level, which is usually less than 0.2, or power of a study, which is usually greater than or equal to 80 percent; (4) sample size; and (5) the directionality of the test (one-tailed or two-tailed analysis). In addition to requiring the parameters of the statistical test as context, surrounding context is necessary as well. P-values do not give valuable information to making inferences or medical-decisions unless characteristics of the trial, such as the study population and collection methods, are thoroughly analyzed. In fact, understanding the context can aid in ruling out alternative explanations or sources of biases for observed results and allow for generalization [Kirk 2012].

2.3.5 Sample Size Calculation

Sample size calculations are necessary to justify any conclusions that may be made from an analysis. When testing if two treatments differ, studies with low power often find no significant differences between the treatment intervention and control groups. Most clinical trials that claim two treatments are equivalent are underpowered, lacking sufficient numbers of study participants [Clark 2011]. If the study was inadequately powered, then a type II error is more likely to occur. A type II error occurs if one fails to reject a null hypothesis that is false. Type II errors occur not only due to a limited number of subjects, but can also occur because there are too many measurements made on too few subjects. If one measures two groups of subjects twice, it is likely that some of the measurements taken on the second occasion will be different from the first set. Thus, a power calculation is critical in studies of equivalency to justify study claims.

A small sample size can add to the context because it is greatly influenced by bias, as compared with a larger sample size. A study reported in 2001 by Gluud et al. examined the influence of study size on study outcome [Gluud 2001]. Specifically, a meta-analysis involving 190 randomized trials over 8 different therapeutic interventions were divided into those with more than a thousand participants and those with less than a thousand participants. The results of this analysis demonstrated that the smaller sized studies had more positive therapeutic effects than those studies with the larger size. These researchers also reported that the larger studies were systematically less likely to report a positive effect, suggesting bias occurs more frequently and has a greater impact in smaller studies. Thus, the sample size contributes to the context in understanding clinical trial results and also requires context on its own. The sample size depends on four critical quantities: the type I and type II error rates α and β, the variability of the data σ, and the effect size d.

2.4 Current Representations of Clinical Trial Studies

Information without context can lead to ambiguities in evaluating the quality and strength of the study. Even when context is included in a paper write up, readers often have a difficult time creating a complete cognitive picture of how all such details fit together. One main reason is the use of free-text with minimal document semantic structure. In this section, I summarize the current representations and give an example of the difficulties of summarizing context.

2.4.1 Current Representations

The current representation for clinical trial evidence is a free-text report made public through academic journals. Typically, the clinical trial report contains several generic sections: abstract, introduction, methods, results, discussion, and conclusions. These sections contain information in

several formats and can appear in any page of the report. For instance, results are presented in narrative prose and often summarized in tables and figures. Figures and tables may not be spatially adjacent to their respective context and descriptions within the paper, and is demonstrated with a running example in Section 2.4.2. Ideally, the layout would clearly connect methodology with its corresponding results and appropriate figures and/or tables. Because this free structure does not connect the numerical data with its context, this can hinder the assessment of quality [Chalmers 1981].

Aside from the structure, the representation of statistics is not intuitive for applying statistics to individual patients. Clinical trials use orthodox statistics for the many types of trial designs, including but not limited to parallel group design. A parallel group design is used for confirmatory trials where subjects are randomized to different arms, with each assigned to a different intervention [ICH E9 1998]. It is suitable for assessing and comparing responses in patients with and without an intervention. Statistics such as confidence intervals and p-values are used to test the difference between the experimental and the control populations. These statistics are the key to identifying the strength and quality of the results collected [Pan 2013, Chootrakool 2011]. Moreover, the quality of conclusions reached by experimental studies are dependent on these statistics, sample sizes and significance levels [Davis 2006, Thornton 2000]. While these statistics provide essential information for assessing the study quality, there is a lack of methodology on applying statistics to individuals whose characteristics differ from a given eligibility criterion. This is partly due to the inherent modeling goals of classical hypothesis testing.

Classical hypothesis testing is used to test the null hypothesis on a sample population [Marden 2000]. Hypothesis testing requires two logical hypotheses: the null hypothesis and the alternative

hypothesis. A typical result rejects one hypothesis and accepts another as true. Within the framework of classical hypothesis testing, a test may lead to the rejection of the current theory, however, the rejection of the current theory does not imply that the alternative hypothesis is true [Senn 1991]. For example, a typical null hypothesis is that the intervention has no effect. A significantly small p-value indicates strong evidence against the null, but does not mean that the drug has an effect. Thus outcomes from clinical trial studies can only reject a null hypothesis and cannot be used to make predictions based on the alternative hypothesis. While clinical trials are well-designed and carefully conducted to test a hypothesis, it is not intuitive how to generalize significance from these frequentist statistics to a given patient. These limitations in determining applicability require a new way to represent knowledge that moves away from free-text trial designs. The representation described in this dissertation presents an important stepping stone towards providing researchers, particularly those interested in disease models, context for numerical data to support context-dependent inferencing on clinical trial knowledge.

2.4.2 Difficulty Summarizing Context: Running Example

To illustrate the difficulty in identifying context, the following discussion looks at the strength of evidence behind a statement written in the abstract and the information needed to support that statement. Within the clinical trial report of the running example, the abstract [Johnson 2004] states the following (Figure 2-1, Box):

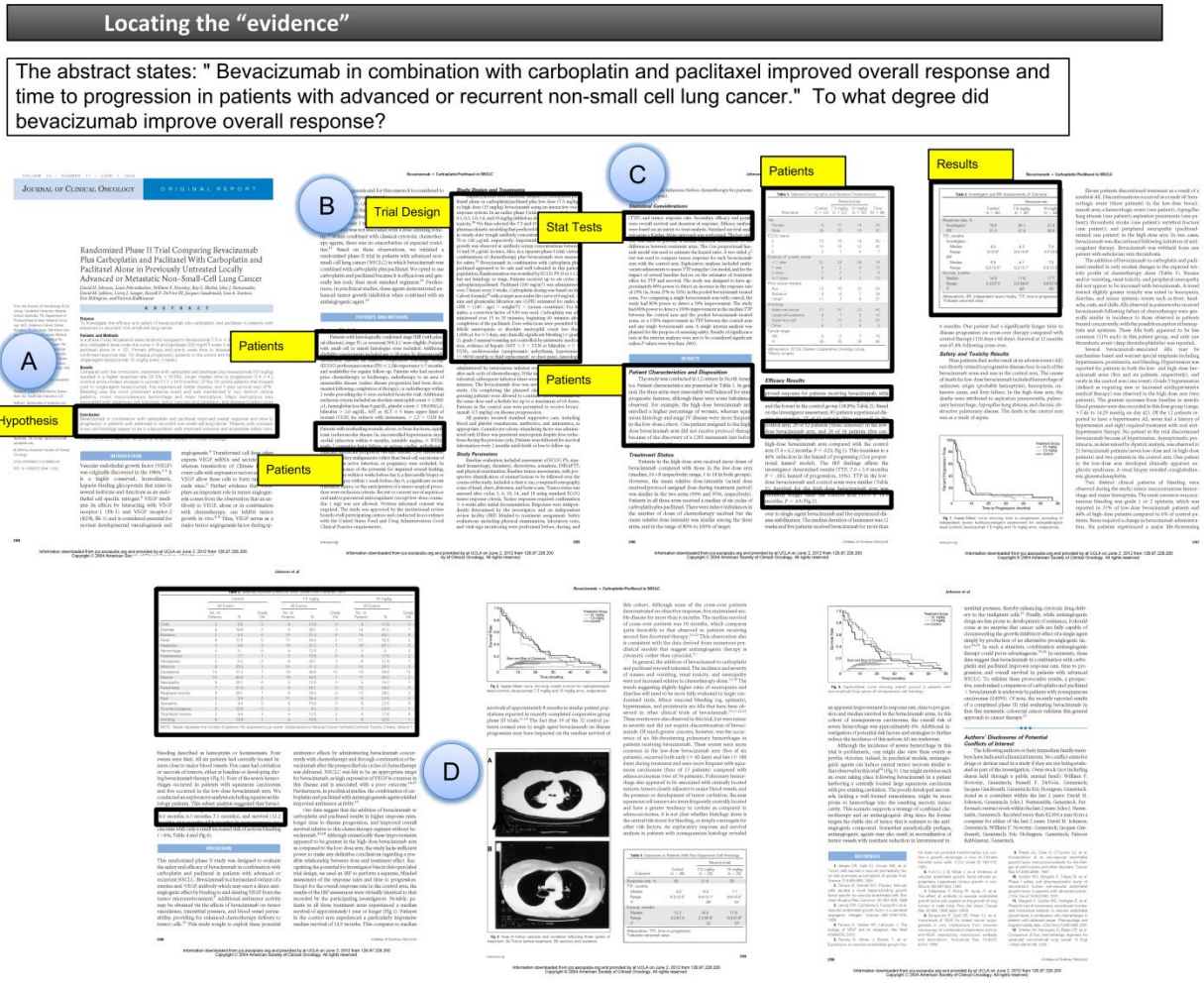


Figure 2-1. A screenshot of a clinical trial paper with relevant context on efficacy of bevacizumab highlighted

To fully understand the degree to which bevacizumab improves overall response, each piece of context should be carefully examined. However, the fragments of information scattered throughout the write-up is in no particular standardized and/or consistent layout, and makes searching and connecting such information difficult for a given reader. Examples of context for the running example include the following:

- Hypothesis. The hypothesis of the paper is, "to investigate the efficacy and safety of bevacizumab plus carboplatin and paclitaxel in patients with advanced or recurrent non-small-cell lung cancer." The hypothesis is seen within the abstract, and the patient recruitment characteristics are described in extended free-text and not in a computer understandable format (Figure 2-1, label A).

- Population Characteristics. The main patient population criteria, "with histologically confirmed stage IIIB (with pleural effusion), stage IV, or recurrent NSCLC were eligible," along with other criteria are described in detail in free-text and in a corresponding table (Figure 2-1, label B and C).

- Methods. The procedure for data collection related to the variable of response rate is embedded with a number of other variables in the study parameters section. The method to determine improved overall response and time to progression is embedded with other methodology being described within the text (Figure 2-1, label B).

- Data. Although the raw data is usually found within the results section, it is found to be captured in more than one format, including as a summary statistic, in free-text, and sometimes in a graph or table (Figure 2-1).

- Statistical Methods. Similar to the methods, the procedure for statistical methods related to the variable of response rate is embedded in a section with other statistical methods, requiring the reader to look for specific methods related to response rate. The power of the study is listed in the statistical methods section: "The study was designed to have approximately 80% power to detect an increase in the response rate of 25% (i.e., from 27%

to 52%) in the pooled bevacizumab treated arms." The statistical methods specific to the hypothesis of the study and the power calculations are listed with other statistical methods mentioned within the statistical considerations section of the write-up (Figure 2-1, label C).

- Subgroup Analysis. The report also contains subgroup analyses, which are not pertinent to the main hypothesis (Figure 2-1, label B and D).

In summary, context is necessary when interpreting numerical results; however, in the current representation, the inconsistency by which such context is authored in a free-text report leads to significant effort and time to manually gather.

2.5 Structuring Free-Text Clinical Trial Reports

To address the limitation of using a free-text report, structured full-text clinical trial reports are desirable. The difficulty in creating disease models occurs when knowledge is ambiguous or missing and/or not well linked. Information models and meta-data standards are approaches for improving the characterization of information. They use standardized vocabularies (ontologies), formal representation languages that promote semantic clarity, that support the free exchange of scientific data and knowledge; and vary widely in term of their functionality (syntactic interoperability, structural interoperability, semantic granularity).

The need for formalizing information contained within clinical trials research papers has been previously recognized and has been motivated by a number of on-going efforts in the informatics field. Driving needs include: 1) the need for editors, peer reviewers, and readers to understand

how the trial was performed and to judge whether the findings are likely to be reliable; 2) the need for decision support for evidence-based medicine; 3) the need to create comprehensive disease models; 4) the need for more sophisticated (accurate) retrieval systems. The specification for defining a good representation has evolved from many complementary efforts. In this section, a brief description of a sample of such efforts is given below:

2.5.1 CONSORT

Efforts and motivation for structuring and synthesizing treatment protocols have been researched by the Consolidated Standards of Reporting Trials (CONSORT). The CONSORT statement defines a 21-point checklist to aid RCT authors in deciding what to report [Altman 2001 and Hopewell 2008]. Table 2-1 summarizes the various items. Items related to methods, results, and analysis aim to improve the critical appraisal and completeness of clinical trial reports and has received powerful backing from journal editors including *JAMA*, *Annals of Internal Medicine*, the *British Medical Journal* and at least 70 other leading journals [Moher et al. 2010].

The CONSORT statement requires that interventions for each testing group be explained in sufficient detail to allow for reproducibility of results, including how the interventions were administered. Specific to treatments and interventions, there is a checklist of characteristics that consists of drug name, dose, method of administration, timing and duration of administration; conditions under which interventions are withheld, and titration regimen.

While CONSORT gives a detailed checklist for the necessary information a clinical trial needs to include, its representation is not standardized and there is no criteria for how clearly and completely this information is conveyed. It provides guidelines for what information should be

included in a scientific paper, but lacks structure with respect to: 1) semantic clarity – many fields are typically free-text descriptions with no constraints on how completely or detailed they should be; 2) connection of information fragments – the checklist does not model how the various items are connected including how data are collected and/or analyzed. Thus, the interpretation of summaries as represented in a CONSORT summary can still be ambiguous.

Section/Topic	Checklist item
Title and abstract	
	Identification as a randomised trial in the title
	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)
Introduction	
Background and objectives	Scientific background and explanation of rationale
	Specific objectives or hypotheses
Methods	
Trial design	Description of trial design (such as parallel, factorial) including allocation ratio
	Important changes to methods after trial commencement (such as eligibility criteria), with reasons
Participants	Eligibility criteria for participants
	Settings and locations where the data were collected
Interventions	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
Outcomes	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
	Any changes to trial outcomes after the trial commenced, with reasons
Sample size	How sample size was determined
	When applicable, explanation of any interim analyses and stopping guidelines
Randomisation:	
Sequence generation	Method used to generate the random allocation sequence
	Type of randomisation; details of any restriction (such as blocking and block size)
Allocation concealment mechanism	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned
Implementation	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions
Blinding	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how

	If relevant, description of the similarity of interventions
Statistical methods	Statistical methods used to compare groups for primary and secondary outcomes
	Methods for additional analyses, such as subgroup analyses and adjusted analyses
Results	
Participant flow (a diagram is strongly recommended)	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome
	For each group, losses and exclusions after randomisation, together with reasons
Recruitment	Dates defining the periods of recruitment and follow-up
	Why the trial ended or was stopped
Baseline data	A table showing baseline demographic and clinical characteristics for each group
Numbers analysed	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups
Outcomes and estimation	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)
	For binary outcomes, presentation of both absolute and relative effect sizes is recommended
Ancillary analyses	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory
Harms	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)
Discussion	
Limitations	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses
Generalizability	Generalizability (external validity, applicability) of the trial findings
Interpretation	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence
Other information	
Registration	Registration number and name of trial registry
Protocol	Where the full trial protocol can be accessed, if available
Funding	Sources of funding and other support (such as supply of drugs), role of funders

Table 2-1. Summary of CONSORT statement items [Moher 2009]

2.5.2 Ontology Development Efforts

Various informatics groups have worked on defining and standardizing information related to clinical trials. The Ontology of Clinical Research (OCRe) is a formal ontology for describing

human studies that attempts to consolidate multiple information standards (e.g., BRIDG, CDISC) and clinical terminologies (e.g., SNOMED CT) [Sim 2010]. OCRe is an extension of the RCT Schema, which captures concepts related to a trial's design, basic intervention description, execution, administration, and results. OCRe describes trial characteristics such as interventions, outcomes, and population descriptions, as well as funding sources and publication details.

Further efforts also include the Ontology for Biomedical Investigations (OBI) project, which developed an integrated ontology for the description of biological and medical experiments and investigations [Brinkman 2010]. This ontology aims to model the design of an investigation, including protocols, instrumentation, materials, and data.

Research has been done on requirements for structuring focused aspects of the clinical trial such as eligibility criteria. Weng et al., surveyed literature on current knowledge representations and identified five aspects of eligibility criteria [Weng 2010]. Their survey consisted of a total of 27 models or systems with computer-based eligibility criteria knowledge representations. Each representation was analyzed from 5 perspectives: (1) the use case of eligibility criteria knowledge representation; (2) the conceptual classification of eligibility criteria; (3) the choice of expression and query languages; (4) the encoding of medical concepts; and (5) the modeling of patient data. Their investigation determined that knowledge-bases for eligibility criteria were designed for 3 major use cases: determining eligibility, determining applicability, and classification.

Another effort is the Ontology of Scientific Experiments (EXPO), which standardizes experimental design, execution, and analysis of a scientific experiment [Soldatova 2006]. EXPO defines over 200 concepts for creating semantic markups about experiments.

These ontologies are built for various purposes and formalize information to varying degrees of granularity. The most common purpose is to match patients for trial recruitment or semantic markup. In contrast, the goal of the proposed representation is directed more specifically to the issue of providing the specific context (e.g., conditions, parameters, observational methods, etc.) to assist with understanding how various quantitative information reported in a trial is derived.

2.5.3 "Claims" Framework

Blake et al. introduced the Claim Framework that reflects how authors across the biomedical spectrum report findings in empirical studies [Blake 2009]. Information related to a claim is captured in four facets – two causal concepts (see description that follows), a change, and the basis of the claim. A causal concept reflects an abstract or concrete idea within a scientific domain and may play different roles in a claim. For example, the agent role reflects the concept that has initiated change, and the object role reflects a concept that has undergone a transformation. A change is defined as how the agent of interest influences the object. Although the number of change terms can be more than one, the results from this study suggest that authors typically use only one word to describe the nature of the change. The basis of a claim captures the author's rationale and evidence to demonstrate their claim. One key contribution is that the Claim Framework captures and informs classification of under-specified claims such as comparisons, observations, and correlations. These distinctions are important as the research moves from trial recruitment towards information synthesis. While the Claim Framework is useful, it lacks context when capturing trial results. Blake acknowledges the helpfulness of context, but the current version of the Claim Framework does not capture related work or experimental conditions. The framework designed in this dissertation addresses these weaknesses by capturing eligibility

criteria, methodology, and results and uses captured information to fully describe clinical trial conclusions.

2.6 Bayesian Considerations

Finally, a discussion of Bayesian considerations is provided as it is common for "big data" methods and disease modelers to extract probabilities and other statistical measures (priors, p-values, etc.) from the scientific literature for use in estimating disease model parameters [Druzdzel 2000].

Lehman et al. discuss the requirements for Bayesian Communication of clinical trials [Lehman 2000]. This work extends the progress made by evidence-based medicine clinicians and researchers, and contains a list of specifications for the creation of a Bayesian model. The list of specifications is divided into requirements for readers, authors, publishers and computers (see Table 2-2).

Lehman et al. developed a prototype web application that implements a subset of these specifications [Lehman 2000]. The example given focuses on the difference between two treatments using a *t*-test. Users specify a prior belief, and parameters include the experimental and control arms, sample size, mean results, standard deviations, units of the outcome, and an indication of which is better or worse. The application then presents the calculated posterior 95% Bayesian confidence interval based on elicited prior belief and the statistics communicated from a selected study. The user is able to specify a threshold for which he/she would not change his/her clinical decision. This threshold is determined by a trade-off of risk and benefit. For example, for

Specification	Bayesian Communication*	Evidence-Based Medicine
Express prior knowledge	Assess prior beliefs; sensitivity analysis for uncertainty in prior (F)	—
View effect size and variability	Mean of posterior beliefs; contaminated models for surprise (F)	Point estimate (F); confidence interval (H)
Express thresholds	Minimally clinically important difference (F if based on utilities)	Number needed to treat (H)
View inferences	Tail probability, credible set, Bayes factor, equivalence (F)	Post-hoc adjustments (H)
Receive explanations	Dynamic algorithms based on influence diagrams (F)	Static textbook explanations (H)
Evaluate study and statistical quality	Likelihood de-biasing (F)	Quality inventories (H)
Synthesize multiple studies	Confidence profile method, Bayesian meta-analysis (H)	Meta-analysis; Cochrane trial banks (F)
View beliefs of the community	Archived priors (F)	Post-publication peer review (H)
Protect authors' investment	Likelihood function (F)	Sufficient statistics (F)
Provide enough information	Information defined by decision problem (F)	Sufficient statistics, Outcomes research (F)
Make authoring easy	Applet libraries	Current program of education and tool-provision

Table 2-2. Bayesian specifications for readers and authors [Lehman 2000] Note: (F) indicates formal solution; (H) indicates heuristic solution.

premature infants in a Level III neonatal intensive care unit, a physician may decide that if administering adenosine did not raise the PO2 more than 10 mmHg higher than the control, the potential drug side-effects do not significantly outweigh the drug efficacy. Lehman et al.'s work envisions more sophisticated models for proportions, time series, and multivariate regression.

Their simple probabilistic model is not complete as much work still needs to be done on assessing priors. In the end, Lehman et al.'s work challenges the electronic publishing community on how the process of reporting data can or will change the language of discourse between investigators and research.

Bayesian considerations are also provided by the Food and Drug Administration (FDA). The FDA created a document for guidance in using Bayesian statistics in medical device clinical trials [CBER 2010] and is encouraging the use of Bayesian methods by providing the requirements for planning a Bayesian clinical trial. As always, good trial design is a necessary requirement for Bayesian statistics, including selection of relevant endpoints and the selection of appropriate controls. Next, similar to Lehman et al.'s list of specifications, the appropriate prior information is required to incorporate analyses correctly. Sources of prior information can include clinical trials conducted previously, patient registries, clinical data on similar interventions, and pilot studies. Prior distributions based on data are easiest to evaluate. Another requirement is the appropriate sample size and/or a criterion to stop the trial. The sample size is dependent on the variability of the sample, prior information, model used, distribution of the parameters, and the decision criteria. In general, the required sample size needs to be sufficiently large. In frequentist clinical trial design, sample size is determined in advance and the trial needs to go until completion. In the Bayesian approach, any particular criterion can stop the trial because at any point during the trial, the predictive distribution can be obtained and is not dependent on the sample size. Another requirement is a thorough evaluation of the operating characteristics in the planning stage, such as Type I and Type II errors. For example, Type I error inflation can occur when large amounts of valid prior information may be more acceptable than others that are irrelevant or statistically

inappropriate. This list, while written for clinical trials involving medical devices, can be adapted to clinical trials in the target domain.

2.7 Summary

This dissertation relates to the development of an improved representation for information presented in clinical trial studies. The representation described in this dissertation assists readers of RCTs, including clinicians and researchers, in understanding the context of numerical data to support quality assessment and clinical trial knowledge inferencing.

Chapter 3 - Specifications of the Representation (SA-1)

3.1 Overview of SA-1 Tasks

Specific Aim 1
To specify a logical representation to concisely synthesize fragments of information found in clinical trial reports, such that users can readily understand the context of numerical data, follow the flow of the study, and assess the quality of the study.

In this chapter, I explain the development and specification of my structured representation focusing on abstracting fragments of information, such as primary outcomes, statistical tests, and survival analyses from within clinical trial reports.

The general approach for developing the representation included the following: 1) identifying users and domain; 2) formalizing the functionality of the representation by sampling representative queries; 3) considering existing methods and incorporating existing knowledge sources such as ontologies for lung cancer; and 4) organizing and linking data elements, processing modelling steps, domain knowledge and analysis methods. The approach taken has involved working closely with domain experts, informaticians, and programmers; and many rounds of iterative design.

3.2 Identification of Users and Domain

The specification for a representation should be motivated by what targeted users are interested in modeling. In other words, the representation serves as a surrogate within a computer to model the essential information required by a given class of users and within the scope of some domain of reality [Sim 2010]. To gain insight on information required by a given class of users, I studied a specific clinical research group at UCLA consisting of clinical researchers, informaticians, and

biostatisticians in their attempt to synthesize the primary literature in the domain of lung cancer. I chose this group because members use medical research literature as the primary source of information, and individuals are required to make decisions based on synthesized evidence. To ensure that the representation supports user needs, a study for functional requirements was performed. The result of this functional requirements analysis was a table of evidence collected by a member of this group on clinical literature.

To narrow down the scope of the domain, I explored the needs of the lung cancer community, looking at both diagnostic and therapeutic trial studies. Lung cancer is a major health problem and is the leading cause of cancer-related deaths in the United States. In 2012, there were 86 740 deaths due to lung cancer in men, and 70 759 in women. An expected 158 040 Americans, in 2015, are predicted to die from lung cancer, making up 27% of all cancer deaths [CDC 2014]. Despite strategies for smoking cessation, the population at risk for lung cancer continues to grow. Diagnostics studies were chosen because most persons with a diagnosis of symptomatic lung cancer ultimately die of this disease [NLST 2011]. One diagnostic trial, the National Lung Screening Trial (NLST), was selected because it was a large multi-center study involving 33 US medical centers, enrolling 53 454 persons, and collecting computer understandable data on hundreds of variables. NLST compared two ways to screen for early signs of lung cancer: low-dose helical computed tomography (CT) vs. standard projectional chest x-ray [NLST 2011]. Because the NLST is a multicenter randomized controlled trial (RCT), it had more than adequate statistical power to detect a modest reduction in lung cancer mortality. Tumors characteristics were defined using variables, such as diameter, consistency, margins, etc.; and were unambiguously represented with numerical values or pre-defined categories. The contributions of

this trial are not only related to defining a measure of benefit for a test that can reduce lung cancer mortality, but also included subsequent publications including feasibility studies, psychosocial issues, study design and technical issues [Aberle 2008, Black 2007, Aberle 2011].

Therapeutic clinical trial reports were narrowed down from general lung cancer to non-small cell lung cancer (NSCLC). Non-small cell lung cancer was chosen for its abundance in the number of clinical trial studies and its complicated biological nature. Most NSCLC patients, if left untreated, have a median survival of 4-5 months after diagnosis and a less than 10% chance of one-year survival [Sharma 2007]. Because of its association with malignant proliferation during the development of lung adenocarcinoma cells, the epidermal growth factor receptor (EGFR) pathway is critical for therapeutic solutions. EGFR is part of the ErbB receptor tyrosine kinase family, which is often deregulated by cancer cells making it a validated target for anticancer therapies. Thus, small molecule reversible inhibitors specific for EGFR have great potential for clinical benefit [Price 2010]. Unfortunately, the clinical benefit of the EGFR-tyrosine kinase inhibitors (TKIs) has an added layer of complexity in that it is limited by primary and acquired resistance. Patients who initially respond to EGFR TKIs develop acquired resistance after a median of 12 months [Yap 2010]. With the number of biological parameters and surrogate observations to include, write-ups of therapeutic trials are often dense and study details can be complex and difficult for readers to follow.

3.3 Functional Requirements

Following identification of users and scope specifications, the next step was to perform a functional requirements analysis. The main goal of the functional requirements step was to

identify the information needed in the representation. This essentially involves understanding what types of queries the representation is intended to support, including necessary inferences, using a two phase process.

In the first phase, a set of "competency questions" were created to drive the design of the representation. An important first step in this phase of my research was to establish a steering committee consisting of three informatics professors, two biostatistician professors, and a clinical researcher in lung cancer (Figure 3-1). To gain familiarity with the general needs of a specific class of users, a comprehensive review of publications on utilizing evidence within clinical trial reports was conducted. Following the literature review, possible query items were collected and organized into categories. The steering committee provided overall guidance and also facilitated the identification of other individuals who routinely read clinical trial reports and analyze its evidence. These individuals helped to form small panels for several rounds of discussions related to the evolution of the required functional requirements. These discussions occurred over several months.

The meetings resulted in a set of compiled queries from free-form discussions. The intent was to create a "most-frequently asked" list of questions when looking at a clinical trial report from the perspective of the targeted users. Participants were asked at various stages of investigation to provide feedback of the relative importance of possible questions. Following these discussion, the list of queries was revised and circulated to the steering committee to ensure that it reflected relevant discussions. Study quality queries were broken down into categories relating to: 1) general information about the clinical trial study design; 2) clinical information necessary for specific patient case; and 3) statistical information on the strength of the trial. Query items related

to study design and analysis are listed in Table 3-1, and query items for disease modeling and EBM application concerns are listed in Table 3-2.

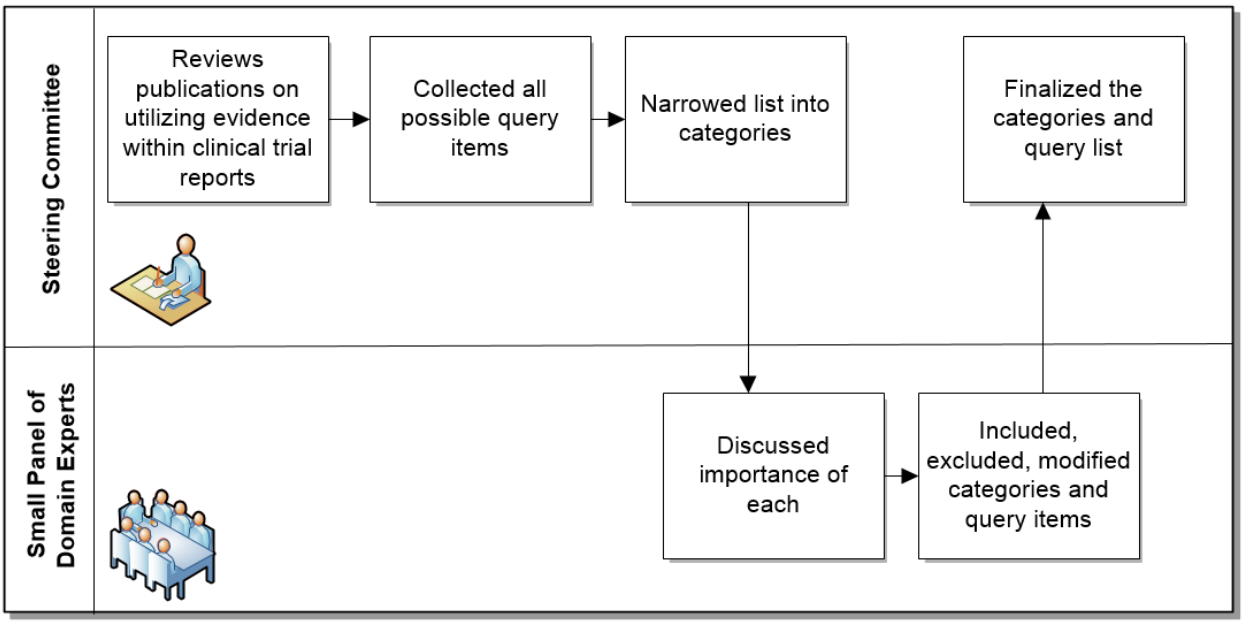


Figure 3-1. Flow chart of requirements analysis

The second phase of the functional requirements task was to observe how an existing research group in lung cancer attempt to summarize clinical trial studies. For this task, I reviewed the content, organization, and framework used by Professor Denise Aberle's research group. The primary tool was a spreadsheet used for recording hypotheses, statistical methods, results, confidence intervals, etc. Different worksheets organized the information contained in the background, methods, results, and discussion sections of a paper. Figure 3-2 shows an example from this research group and their attempts to capture the knowledge from a paper for the purpose of evidence based medicine.

Target queries for study design and analysis
What is the hypothesis?
Is the hypothesis one-sided or two sided?
What is the type I and type II error assumed in sample size calculations?
What is the trial design?
What are the specific conditions of the subgroups?
Have known confounders been appropriately controlled for?
How many patients were lost to follow-up, discontinued treatment, etc.?
What patient data is missing?
Does the sample size population match actual number of patients sampled?
Do post-hoc analyses address the sampling issue?
What is the data set associated with a particular statistical test?
What is the statistical test for a particular data set?
What statistical measure is to characterize the difference between two time-oriented curves? (example Bayesian statistic)
What is the p-value associated with the null hypothesis?
What is the threshold for p-value significance?
What is the estimated effect size for significant p-values?
What is the hazard ratio and confidence interval associated with the intervention?

Table 3-1. Sample of target queries for study design and analysis

Target queries for class disease modeling
What is the causal mechanistic hypothesis?
What are the various contextual factors that can affect the study hypothesis?
What is the context implicit in a stated frequency (probability) stated in the trial paper?
Can we estimate posterior probabilities from p-values or other reported information?
How do we estimate specific conditional probabilities required in the EBM model from the partial statistics reported in clinical trials?
How do we synthesize nodal relations indicated from multiple studies?
How can we estimate Bayesian parameters from orthodox statistics? (i.e., Bayes factor)
Do the clinical characteristics from this study's patient population apply to my patient's clinical characteristics?
Are results from the study generalizable to my patient?
Is this study too population-orientated for my patient?
What are the adverse effects associated with an intervention?

Table 3-2. Sample of target queries for class disease modeling

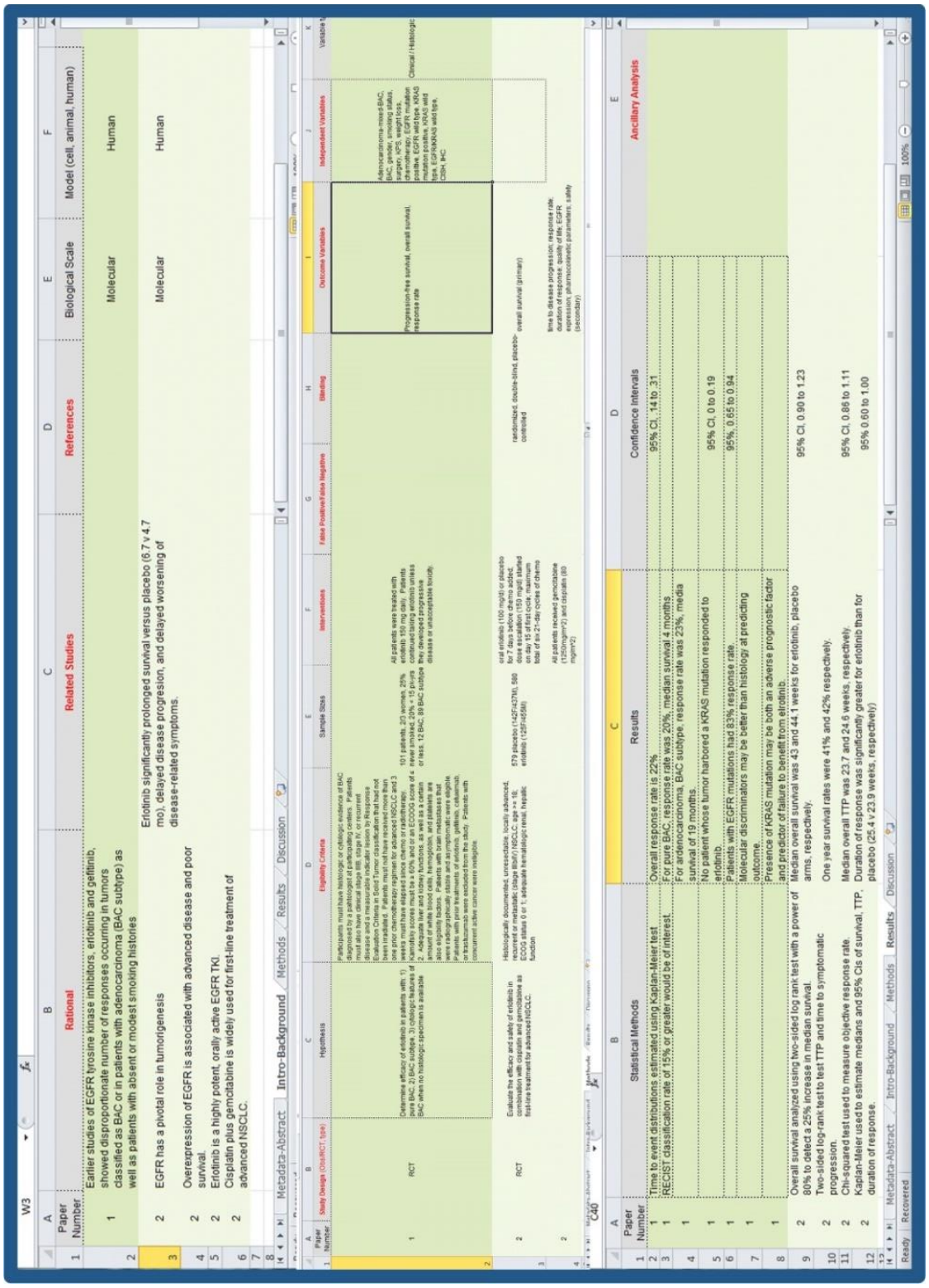


Figure 3-2. Early attempts by a UCLA research group to organize clinical trial literature in lung cancer (courtesy of Professor Denise Aberle)

3.4 A Situational Ontology for NSCLC Clinical Trial Reports

This section is drawn mainly from my work published in the paper below:

> **Tong M**, Taira RK. "Improving the accuracy of treatment descriptions in clinical trials using a bottom-up approach," Proc of the American Medical Informatics Association Fall Symposium, pp.1393-1402, 2012.

Clinical trial reports commonly have complicated therapy descriptions that are written in free-text. Not only are administration details important, details regarding protocol changes must be described clearly for reproducibility and quality assessment. Many times, this information can be imprecisely or incompletely described. An ontology can partially address the issue by making knowledge more explicit. As part of the dissertation research, an ontology of important specification topics related to clinical trials was developed for the domain of non-small cell lung cancer (NSCLC).

The purpose of this ontology is to standardize variables that appear in a clinical trial study for single study exploration and for across study comparisons. The situational ontology defines in a standardized way the concepts and vocabulary used in the domain of NSCLC clinical trials. To construct the situational ontology for NSCLC related clinical trial reports, I used both a top-down and bottom up approach.

3.4.1 Top-down Ontology Development: Adapting Existing Ontologies

A top-down approach starts with a general concept and translates that concept down to more detailed elements. For example, the methodology described in a clinical trial report includes drug interventions, which can then be described by drug type, frequency, and dosage. The top down

approach insures the generality and coverage to produce a robust ontology. The top-down approach usually starts with assessing and adapting, as necessary, the content of existing ontologies and knowledge sources. To develop the situation ontology for the proposed representation, I first extended pre-developed ontologies. In particular, I heavily borrowed ontological entries from the RCT schema [Sim 2004, Sim 2010], which is the most well-established knowledge source for specifying clinical trial summaries. RCT Schema consists of four top-level tasks and 62 subtasks that assist with standardizing the systematic reviewing task for clinical trials. These items relate to a trial's design, execution, administration, and results; and served as a base information model for my application. Other concepts specific to lung cancer were pulled from the Unified Medical Language System and the National Cancer Institute's Thesaurus (NCIT) [Ceusters 2005].

After the initial iteration of the ontology's development, the next iteration involved augmenting the ontology with elements gathered by human experts. To accomplish this task, I chose a test paper outside NSCLC, but within the oncology domain, to assess the ontology's adherence to the CONSORT RCT guidelines [Cloughesy 2008]. Three expert readers were asked to determine important elements of this report. Each expert reader identified elements alone without the influence of the other two readers. Afterwards, the expert readers gathered all elements and modified each element until a consensus was achieved. The master annotation list was compiled and the database schema was developed (Figure 3-3). The database schema was divided into several tables: hypotheses, recruitment, experimental procedures, raw data, statistical methods, and interpretations.

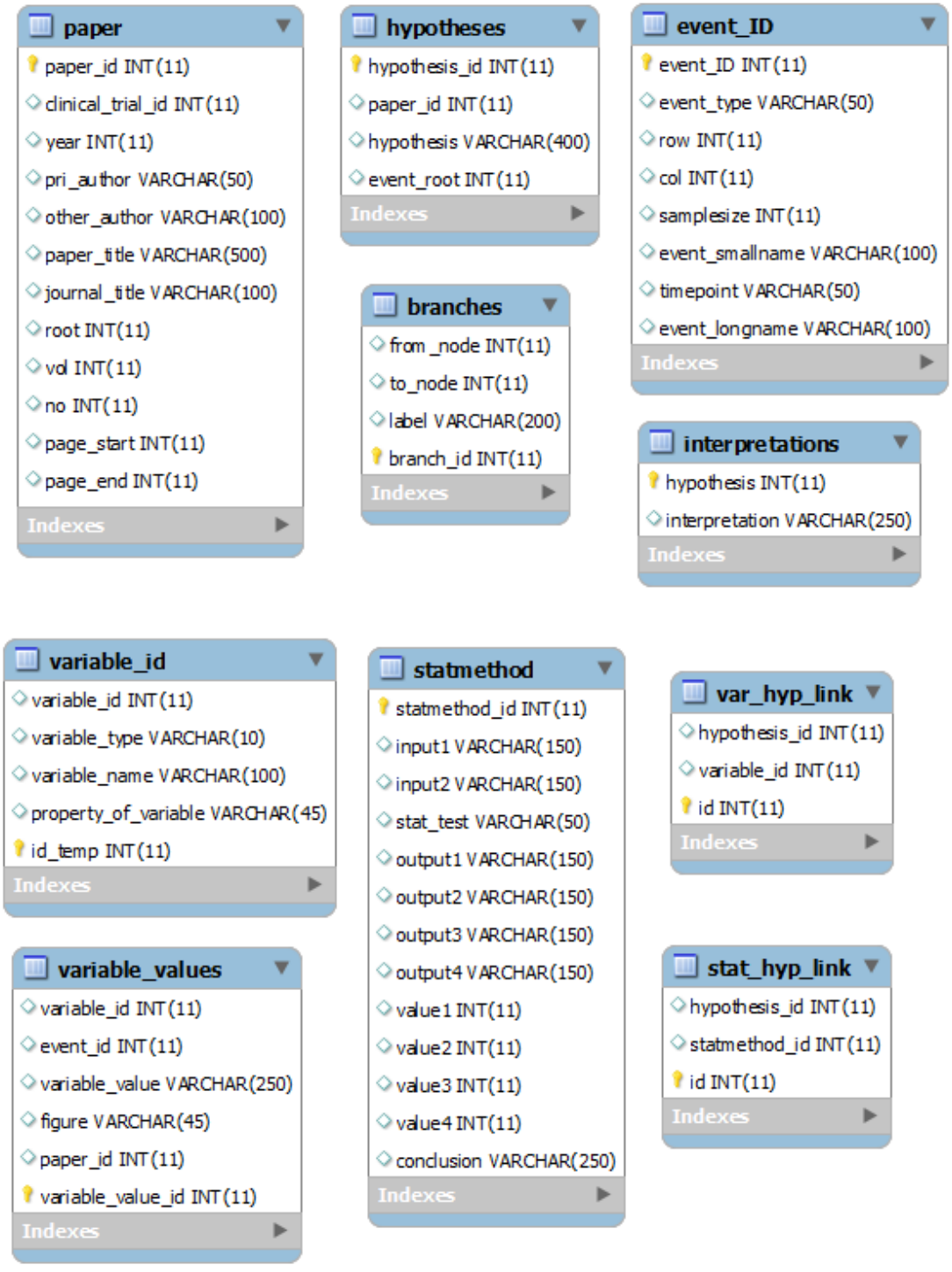


Figure 3-3. Base data schema for representation

While the top-down approach generated an initial base data schema, it contained many disadvantages. Despite being a solid base for data modeling, RCT Schema lacks standardization of some field values and the ability to add more granularity. Specific to treatment and intervention concepts, although there is a list of important intervention attributes (e.g., drug name, dose, method of administration, timing and duration of administration, conditions under which interventions are withheld, and titration regimen), the RCT model does not standardize their values and there is no scoring criteria for how clearly and completely this information is to conveyed. For example, the "*Interventions-Concept Class*" does not allow for detailing treatment descriptions as a sequence of steps and/or with decision points. These types of information are crucial for a knowledge source that is to guide evidence-based medicine practices. Thus, part of this dissertation work involved a more precise modeling of treatment descriptions.

3.4.2 Bottom-Up Ontology Development

In contrast to the top-down approach, the bottom up approach starts with a search of concepts from source documents and fills in missing entries and details with respect to the base ontology. The bottom up approach involves building up the ontology to include a more comprehensive compilations of concepts for the domain of NSCLC and is used to identify gaps in each aspect of the base ontologic model. For example, concepts such as "grade 3-4 dyslipidemia" and "6 cycles" mined from a sample of RCT papers (see below) have been grouped into toxicities and dose cycle classes, respectively. Similar to the top-down approach described previously (section 3.4.1), three expert reviewers assisted in the ontological curation of terms found in the various sections of the clinical trial reports using the same paper as before [Cloughesy 2008]. Reviewers identified and documented entries individually. The terms from the three reviewers were collected as the results,

and I used these entries to define concept classes as necessary and merged lexical variants into a master lexicon.

Following iterations involved further ontology development for areas where the representation was lacking. The goal of this structured representation was to capture the essential elements related to recruitment, steps of the experiment, the data collection process, the analyses, and the conclusions in a logical and consistent manner. A detailed description of one area which existing knowledge sources were deficient (interventions) is described in the next section.

3.4.3 Ontological Classes for Intervention: An example

Knowledge from RCT studies provide evidence related to the effectiveness of particular therapies. Therefore, it is important to clearly define the precise course of therapies, including change to the drug regimen. Despite efforts to control for regimen changes, many RCTs do not follow the initial therapies planned. Unexpected events can occur when conducting the study, resulting in differences in patients' treatment interventions. Discontinuities in treatment can occur, as well as individualized care from the clinical team. Another hindrance toward the precise specification of therapies occurs at the reporting level. The prevalence of incomplete protocol reporting is high, again often lacking details related to all outcome categories and/or protocol changes. These two reasons motivate the need for a more detailed therapy ontology. The utilization of a common standard ontology for treatments makes knowledge more explicit, helps detect missing data or errors, and promotes interchange and replication of treatments leading to better interpretation of patient results since treatment conditions can affect clinical outcome predictions such as survival.

Researchers and/or biostatistician can also then more easily assess the validity of the assumptions of the experimental design.

3.4.3.1 Document Corpus

The first step of the bottom-up process following the previous iteration was to collect a set of representative RCT papers for the selected domain. A PubMed search was conducted to identify clinical trials on NSCLC to serve as a set of representative reports for ontology development. The PubMed search constraints included the combined keywords related to RCTs (i.e., "*clinical trial*" AND "*Phase I*") and keywords related to the domain topic (i.e., "*lung cancer*" AND "*non-small cell*" AND "*EGFR*"). I then systematically reviewed each article that matched the search. Review and case-study papers were excluded and papers without access to full text in English were excluded. There were 28 remaining articles included in the ontology development. The remaining papers used in ontology development included 16 unique drug therapies. 13 (46%) of the trials used more than one drug. The most common drug for EGFR used to treat NSCLC patients was erlotinib, used by nine trials. 15 (54%) used a combination of two drugs. No trials used a combination of three or more drugs. I observed and noted several new stopping conditions. The most common stopping conditions was disease progression, and grade 3 or 4 toxicities. The most common protocol change action is dose reduction.

3.4.3.2 Term Identification

From the set of 28 documents from our development corpus, expert reviewers were asked to identify two categories of therapy-related terms: 1) treatments; 2) conditions that can affect the

course of treatment. An excerpt from the paper by Price et al., 2010 will be used as a running example to illustrate the process [Price 2010]:

> *"After obtaining informed consent, patients were treated with gefitinib 250 mg daily and everolimus 5 mg daily as determined in our earlier phase I study. Dose reduction of everolimus to 2.5 mg daily was allowed for toxicity not managed by optimal supportive care. Dose reduction of gefitinib to 250 mg every other day was allowed for side effects attributable to gefitinib. Dose interruption of both everolimus and gefitinib for grade 3 or 4 toxicities was allowed until resolution of the toxicity (≤ grade 1). For grade 3 or 4 skin toxicity, dose interruption of gefitinib only was allowed with continuation of everolimus unless the toxicity did not resolve within 1 week. For grade 3 or 4 dyslipidemia, dose interruption of everolimus only was permitted. Patients with grade 3 or 4 toxicities that did not resolve in 2 weeks were removed from the study."*

Examples of the drug treatments can be seen in the first sentence of the excerpt: everolimus and gefitinib, with dosages of 250 mg daily and 5mg daily, respectively. Examples of the conditions that can affect the course of treatment are seen in the sixth sentence describing the stopping condition "grade 3 or 4 dyslipidemia" along with the intervention that was stopped, the "everolimus drug regimen." The ontology also handled more complicated sentences such as the fourth sentence. This sentence contains two stopping conditions, "grade 3 or 4 skin toxicities" and "toxicity lasting more than one week." The resulting intervention process depends on both stopping conditions.

3.4.3.3 Class Definitions

I organized the classes related to therapy into four classes with a total of 26 attributes, building on top of the base ontology model (see Figure 3-4).

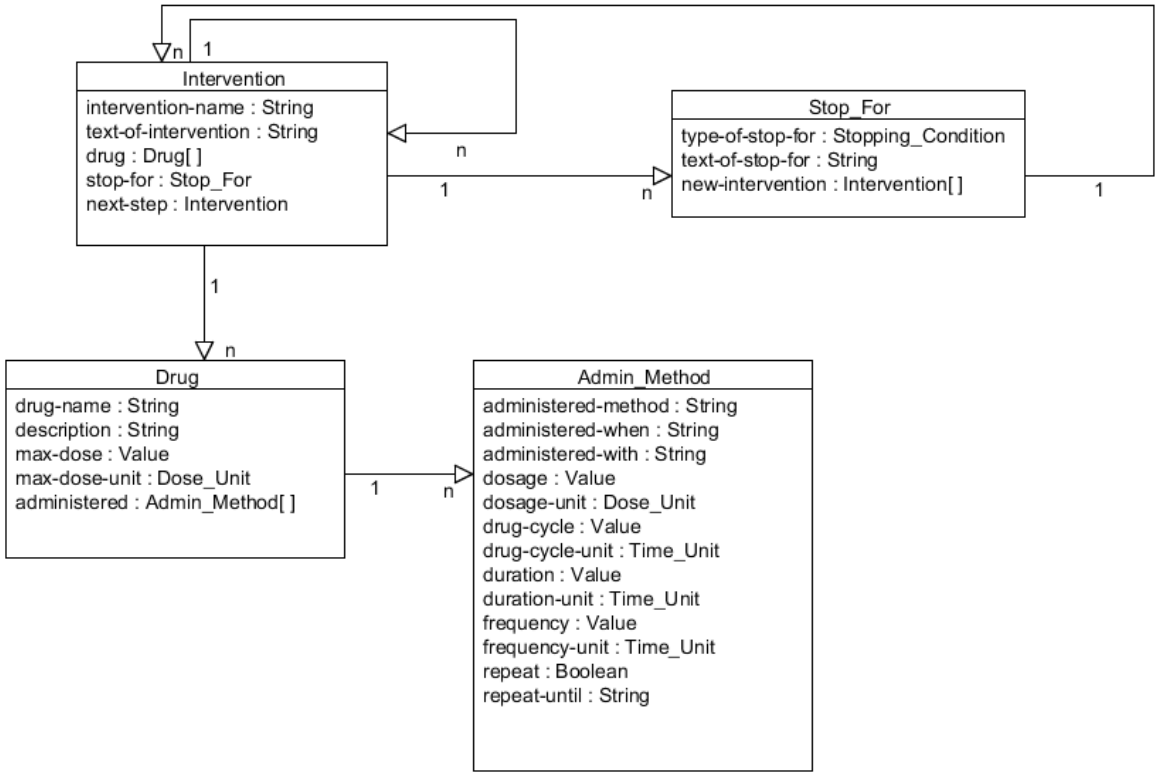


Figure 3-4. Data ontology for therapy intervention

- The Intervention class describes the events in a drug regimen. Because a drug regimen can have two or more drugs, the Intervention class is allowed to have multiple instances of drugs, each drug described by the Drug class. Because I was modeling interventions as a list of ordered events, each intervention event points to the next intervention event. Each intervention includes stopping conditions and subsequent changes in protocols. In addition, the Intervention class can be labeled with a name and description.
- The Drug class describes all information needed to replicate every administration detail of a particular drug, including administration method, dosage, drug cycles, duration, and frequency. The Drug class can contain multiple instances of administration methods,

capturing the various ways a drug can be administered. For instance, a drug can be administered daily, weekly, or monthly, as well as in different dosages.

- Each instance of the Admin_Method class describes one type of administration. Each drug can have various methods of administration. See Table 3-3 for attributes and example instances for this class.
- The Stop_For class describes the stopping conditions. Typically, drug regimens have complicated protocols to discontinue or change the use of a drug. Each Stop_For class is allowed multiple instances of the Stopping_Condition class (not shown in Figure 3-4). A standardized list of stopping conditions can be found within the Stopping_Condition class. When describing why the intervention changes, both the stopping condition and the new intervention needs to be captured. An example can be seen when protocols change due to participants showing grade 3 toxicities, such is the case for sentence 5 in our running example. The appropriate stopping condition is selected from a pre-defined list, and the protocol changes are described as a new intervention.

3.4.4 Example Data Representations for Intervention Class

In this section, I provide some example representations for the example excerpt from Price et al. (previously shown in section 3.4.3). In the excerpt, the text describes an intervention (here labeled "Interv_1") which can have five different stopping conditions, and subsequently, each stopping condition leading to a modified intervention as illustrated in Figure 3-5. Note in the figure that after "Interv_2," there is a stopping condition labeled "Stop_For_6," after which, the original intervention is reinstated. By portraying interventions as a process of events, the system is able to

Attribute Name	Description	Example Entry
administered-method	Method of drug delivery	"Orally", "IV"
Administered-when	Any addition information describing when the drug was administered	"Before breakfast"
administered-with	Co-delivery agents. This can include other drugs or non-active ingredients.	"250 mL saline"
Dosage	Dosage of the drug.	250
dosage-unit		Dose_Unit object containing information "mg"
drug-cycle	Length of a drug cycle, defined by trialists. This is different than the frequency property. For example, drug can be administered every day, however, the drug cycle can be defined for 2 weeks.	2
drug-cycle-unit		Time_Unit object containing information "week"
duration	Duration of drug infusion. This is useful to describe iv drugs, and is usually null for orally administered drugs.	90
duration-unit		Time_Unit object containing information "min"
frequency	Frequency the drug was administered. This usually takes the form of daily, weekly, etc.	1
frequency-unit		Time_Unit object containing information "day"
repeat	Answers the question: Was this drug pattern repeated? Allows for the entry for how long an event is repeated for	TRUE
repeat-until		6 cycles

Table 3-3. Attribute list of the Admin_Method class and example entries

better describe deviations from the initial protocol. For example, one can see an error if the intervention after a stopping condition matches the intervention before the stopping condition. In Figure 3-5, I notice that each modified intervention is different, as noted by a unique label identifier.

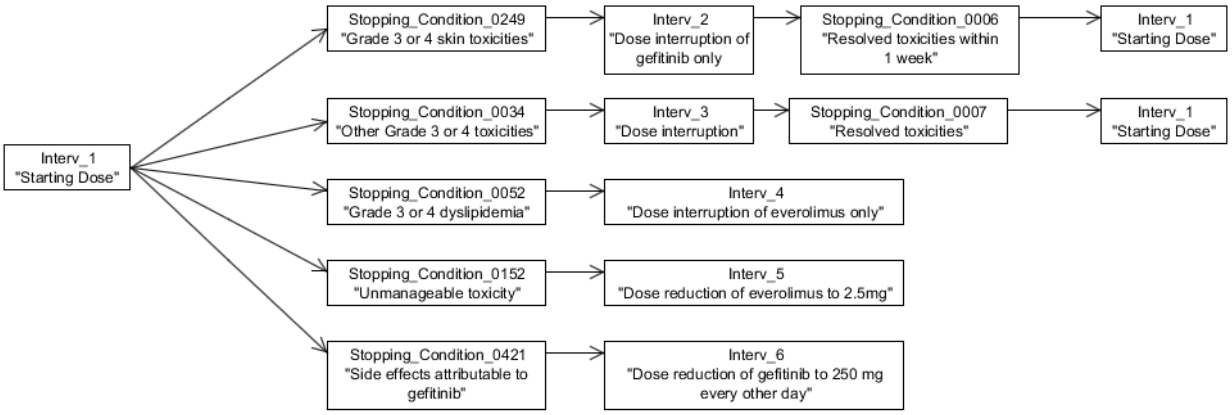


Figure 3-5. Modeling of an initial intervention with various treatment modifications

Details of the "Interv_1" instance representation for the excerpt are given in Figure 3-6. Recall sentence 1: "patients were treated with gefitinib 250 mg daily and everolimus 5 mg daily as determined in our earlier phase I study." This sentence mentions two drugs, gefitinib and everolimus. Hence, the representation instantiates two instances of the Drug class, "Drug_gefiti_1" and "Drug_everol_1." For each drug then, an instance of the Admin_Method class was created to account for administration details.

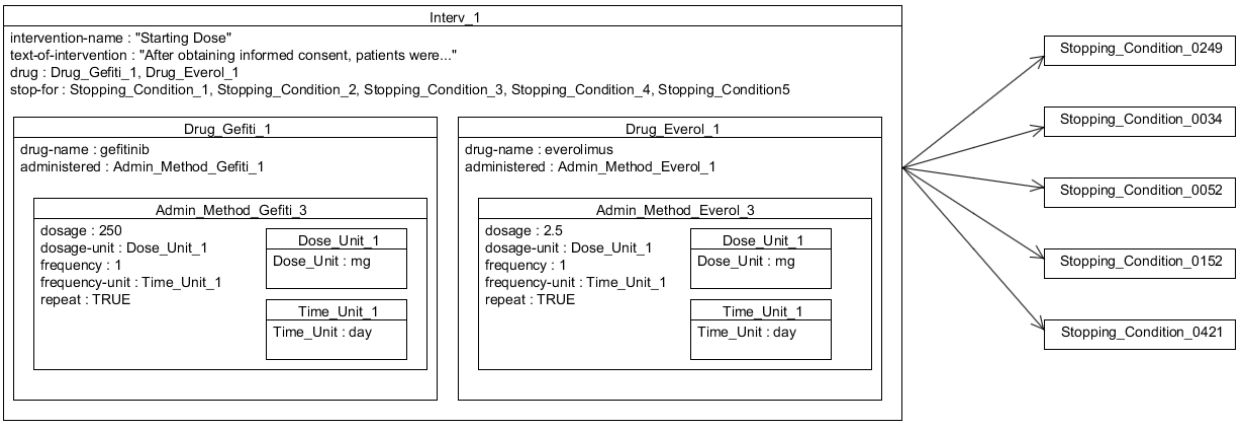


Figure 3-6. Modeling of drug and administration details of initial intervention

To illustrate how the stopping conditions work, consider the sentence: "Dose interruption of both everolimus and gefitinib for grade 3 or 4 toxicities was allowed until resolution of the toxicity (≤ grade 1)." This sentence contains a two-step stopping condition. In the first step, the stopping condition is the appearance of grade 3 or 4 toxicities. In the second step, the stopping conditions can be lifted if the toxicities resolve to grade 1 or better. Focusing on the first step, an instance of Stopping_Condition called “Stopping_Condition_0034” is instantiated. "Grade 3 or 4 toxicities" is assigned for the type property. In the new-intervention property, a new instance of Intervention was created called “Interv_3” (Figure 3-7), which was populated in a similar manner as “Interv_1.” In the second step, toxicities resolve, the dose interruption terminates and the original treatment continues. Note that the Stopping_Condition class defines not only stopping constraints, but can also be generalized to any changes in patient status, such as the resolution of toxicities. An instance of Stopping_Condition called “Stopping_Condition_0007” reflects this state. The type property is *"resolved toxicities"*, and the new-intervention property is “Interv_1,” which corresponds to the original intervention.

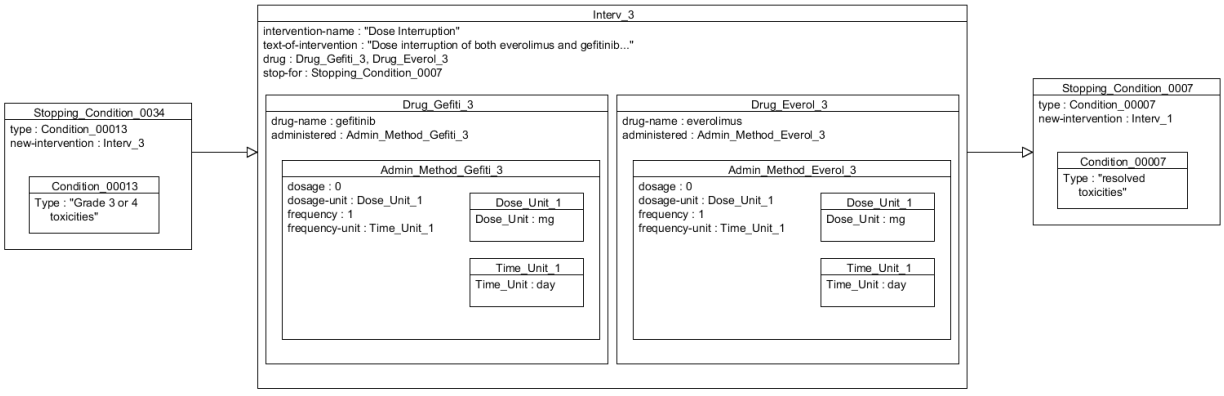


Figure 3-7. Modeling of stopping conditions

3.4.5 Process Model Representation of Clinical Trial Experiment Design

In designing a logical representation for clinical trial reports, a major issue in current efforts has been the lack of context for numerical information. Thus, understanding how a particular quantity is derived is of high methodological importance to arrive at the proper interpretation. To capture information on context, the overall experiment design flow is represented using process modeling techniques. Note that in section 3.4.3 and section 3.4.4, the use of process modeling was introduced in the context of capturing stopping conditions and their resolution. The process model documents the flow of events and populations associated with the clinical trial. Generic RCT event and decision template nodes were defined, examples include general population, sampling pool, decision boxes, recruitment criteria, control arm, intervention arm, randomization methods, etc., similar to [de Carvalho 2010]. Selected Unified Modeling Language (UML) activity diagrams served as the workflow specification language [Dumas 2001].

Process modeling methods were incorporated to characterize experimental design flow. The process model contains several types of building block elements (Table 3-4). The most common elements include: populations, eligibility criteria, and events. Ellipses are used to represent populations of individuals. Diamonds are used to represent decision nodes that affect the sample size number, such as eligibility criteria, discontinued treatment, etc. Rectangles indicate interventions and observational procedures related to hypothesis testing. Example interventional and observational procedures include genetic screenings, surgical interventions, drug cycles, imaging modalities, study end points, etc.

Element	Description	Example
Ellipses	Populations of Individuals	*Starting Population, Control, Low Dose Bev, High Dose Bev*
Diamonds	Decision nodes and stopping conditions	*No Prior Chemotherapy, Stage III or IV Cancer, Other Inclusion/Exclusion Criteria*
Rectangles	Interventions and Observational Procedures	*Imaging, Survival, Bevacizumab 15mg/kg*

Table 3-4. Description of symbols used in the process model

The steps performed in the clinical trial study relevant to understanding the context of recorded data are represented using the elements and linkages of a process model. Each step in the experimental procedure of a clinical trial study is labeled as an element and are linked to the following steps, which are represented as elements. The process model does not give a full specification of how to perform the experiment but instead gives a high level summary with enough detail to describe the full context for an assigned variable.

The linkages between process model elements allow node elements within the same pathway to be extracted, and has implications for recovering context. For a selected node element, a back-tracking algorithm transverses the process model following the semantics of the linkages heading towards the "Starting Population" node. The full path with respect to the starting population creates a subset of nodes, and information related to this subset is collected. The information maintained by the system for each node elements extracted is necessary to describe the context (e.g., population arm, the sample size of the population, randomization techniques, and ascertainment methods) for measurements performed at a given step in the process model. This is further described in section 4.4.

A specific use case of the process model for experimental procedures is demonstrated for Johnson et al [Johnson 2004]. The process model displays the recruitment period on the left and the inventions and observations on the right (Figure 3-8). The first node on the left is labeled as "Starting Population." The node is connected to three diamonds, each corresponding to a separate exclusion rule. Three exclusion rules determine patients eligible for the trial. The first filter is the presence of Stage III/IV cancer, the second is no history of prior chemotherapy, and the third is a combination of other exclusion criteria. After applying these three exclusion rules, the final set of participants is obtained. In the center, this set of participants is randomized into three study arms. Each row in the process model represents its own study arm. In our running example, the study arms are control, low dose, and high dose. Following randomization, the right side displays experimental procedures referred to as a sequence of events. The sequence of events branch from each population node displaying only the interventions specific to each study arm. For our running example, interventions were either 7.5 mg/kg Bevacizumab, 15 mg/kg Bevacizumab or no drug. After intervention, tumor status was measured, following that was survival, and lastly, adverse events.

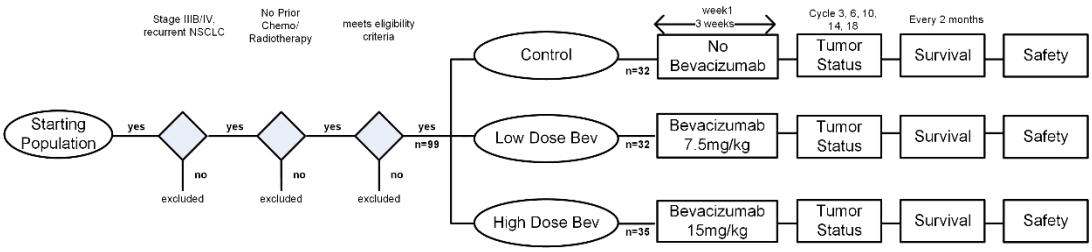


Figure 3-8. Example process model of experimental procedures

An example of the process model used for capturing drug intervention details is demonstrated and derived from Price et al. 2010. The process model illustrates how treatment interventions were

described based on the free-text (Figure 3-9). The starting intervention is given to all patients. The starting intervention is modified as patients experience complications such as grade 3 or 4 toxicities or side effects. The intervention is changed accordingly if other toxicities result or are resolved. One advantage of using the representation model is having the ability to characterize modifications of one protocol that were initially applied to all patients, with modifications typically occurring due to individual patient complications. Another advantage is the ability to pinpoint missing detail in the protocol. In this trial, a stopping condition is mentioned, but the resulting dose reductions were not specified. This research thus helped to address documentation issues that should be included in standardization of treatments written in a set of clinical trial reports for a target disease.

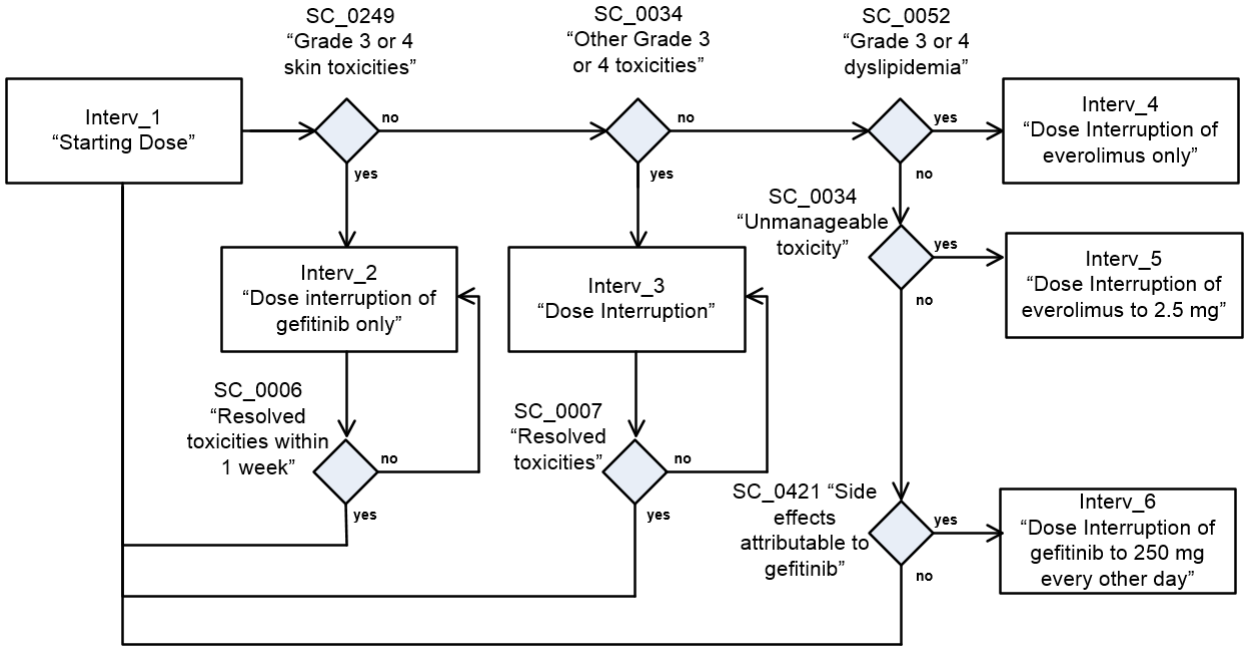


Figure 3-9. Fragment of the process model for the excerpt

3.5 Representation of Constraints, Observations, and Statistics

The assignment of categorical, logical, quantitative values and qualitative descriptions is often difficult to locate in current free-text reports and lacks standardization in terms of data type and level of granularity. In this section, I discuss the representation used for recording information within the context of the experimental design. I am concentrating on numerical information because it has been the least studied and because it is important for study quality assessment and evidence-based medicine; however, the same general methodology was applied to categorical and logical information.

This section is drawn mainly from my work published in the paper below which received a 2nd place best student paper award at the 14th World Congress on Medical and Health Informatics, Copenhagen, Denmark, 2013:

> **Tong M**, Hsu W, and Taira RK. "A formal representation for numerical data presented in published clinical trial reports," Proceedings of the International Medical Informatics Association 14th World Congress on Medical and Health Informatics, Stud Health Technol. Inform. pp. 856-860, 2013.

Additionally, a week-long computer exhibition of a prototype of the system was presented at the 98th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL. 2012.

> **Tong M**, Hsu W and Taira RK. "A representation for standardizing numerical data from clinical trial reports," RSNA Scientific Assembly and Annual Meeting Bioinformatics Exhibit, Chicago, IL. November 25-30, 2012.

3.5.1 Overview of Types of Quantitative Descriptions

Within clinical trial literature, numerical data is essential to understanding and providing the strength and quality of the clinical trial study. Analysis of numerical data can improve the interpretation of statistical analysis and allow for means of integrating evidence from different trials. However, numerical data is not stored with sufficient meta-information for interpretation. For example, to interpret survival curves, it requires an understanding of the processing steps leading to the collection of the data. A knowledge representation combining numerical data into the process of how the data point was generated has not previously been developed. Previous efforts formalize description of information within clinical trials but do not directly structure numerical data with sufficient context to describe the provenance of the data.

As a preliminary investigation to understand the types of quantitative descriptions contained in RCT reports, the types of numbers presented in a report were characterized using a bottom-up and top-down approach. Ten papers on NSCLC clinical trials were sampled from the primary literature. This was performed using a PubMed search containing the keywords "phase," "trial," "NSCLC," and "EGFR." Table 3-5 provides a number of sample text excerpts mainly from [Miller 2008] showing the diversity of situations in which quantitative information is used to describe various states, properties, trends and probabilities. Numerical values can summarize disease prevalence, population characteristics (i.e., distributions), estimated measurements, property constraints, potential errors, and statistical analyses (e.g., p-values, test statistics and confidence intervals). Trial design/recruitment constraints, for example, include information on eligibility criteria for participants, periods of recruitment, interventions with sufficient detail, and outcomes. Collected data includes baseline information on the starting population, as well as baseline data

about the populations and data from experimental procedures, which can be further divided into data about individual patients and data on each population.

Section	Natural language expression *mostly extracted from Miller et al, [Mil08]
Background	Bronchioloalveolar carcinoma (BAC), a subtype of adenocarcinoma, manifest as lepidic growth of tumor cells along the alveoli without stromal, vascular, lymphatic, or pleural invasion. Defined in that rigorous fashion, BAC is uncommon, comprising approximately 1% to 4% of NSCLC
Background	Mutations in KRAS and EGFR are critical to pathogenesis of a large number of lung adenocarcinomas, are mutually exclusive, and occur in approximately 40% of tumors from the US and 70% of tumors from East Asian countries
Background	Mutations in KRAS are found in approximately 30% of human lung adenocarcinomas
Background	More recently, mutations in EGFR have been identified in lung adenocarcinoma and have been associated with response to EGFR-TKI [refs 7, 13, 14]. Mutation in EGFR occur in 13% of unselected US populations, 33% of unselected East Asian populations, and overall in 30% of adenocarcinomas.
Background	Expression of the two most common EGFR mutations, exon 19 deletions and exon 21 L858R substitutions, lead to lung adenocarcinomas in mouse model systems.
Background	More recently, mutations in EGFR have been identified in lung adenocarcinoma and have been associated with response to EGFR-TKI [refs 7, 13, 14]. Mutation in EGFR occur in 13% of unselected US populations, 33% of unselected East Asian populations, and overall in 30% of adenocarcinomas.
Background	Expression of the two most common EGFR mutations, exon 19 deletions and exon 21 L858R substitutions, lead to lung adenocarcinomas in mouse model systems.
Results	Prior cytotoxic chemotherapy had no effect on RR, PFS, or OS
Results	Patients whose tumors had an EGFR exon 19 or 21 mutation had an RR of 83%, whereas in tumors with no demonstrable EGFR mutation, the RR was only 7% (p-value < .01).
Results	Patients with an EGFR mutation had a longer PFS (13 versus 2 months; P<.01) and a trend toward improved OS (23 v 17 months; P=.24)
Results	All patients with KRAS mutation failed to respond to erlotinib therapy
Results	EGFR IHC was of no predictive value.
Results	Presence of ≥ four copies of EGFR was identified in one of their patients and was associated with an RR of 43%. However, in patients who had increased EGFR copy

	number without EGFR mutation, the RR was 8%, similar to the R for all patients without EGFR mutation (9%).
Results	The poor RR and OS in patients with KRAS mutations are consistent with those of other studies. The poor outcome of lung adenocarcinoma and KRAS mutations has also been noted inpatients given adjuvant chemotherapy for early-stage NSCLC. Thus the presence of KRAS mutation may be both an adverse prognostic factor and a predictor of failure to benefit from erlotinib therapy in advanced disease.

Table 3-5. A sample set of text excerpts from RCT studies

A semantic label and format type were manually assigned to each mention of a numeric quantity. For the scope of this dissertation, I focused on numbers presented in the trial design/recruitment process and in the data collection process. Semantic labels describe the numerical data's type and where in the clinical trial report the number is presented (see Figure 3-10, column 1-2). The semantic label for numerical data is first divided into "Recruitment/Study Design" and "Collected Data." Within "Recruitment/Study Design," numerical data can fall into the categories of "Eligibility Criteria," "Intervention," and "Measurements." "Baseline Characteristics" and "Experimental Procedures" are the two divisions of the semantic label, "Collected Data." Within "Experimental Procedures," data can be further classified as "Individual Data Points" or "Population Data Points." In addition to the semantic labels, numerical data can take on a variety of formats which was also characterized (Figure 3-10, column 3), including: i) table data; ii) graph data, including axes, x-max, y-max, x-label, y-label, and x-y points for each series; and iii) free-text statements. The assignment demonstrated the variation and complexity of how numerical data is used to describe a measurement and how the information is conveyed.

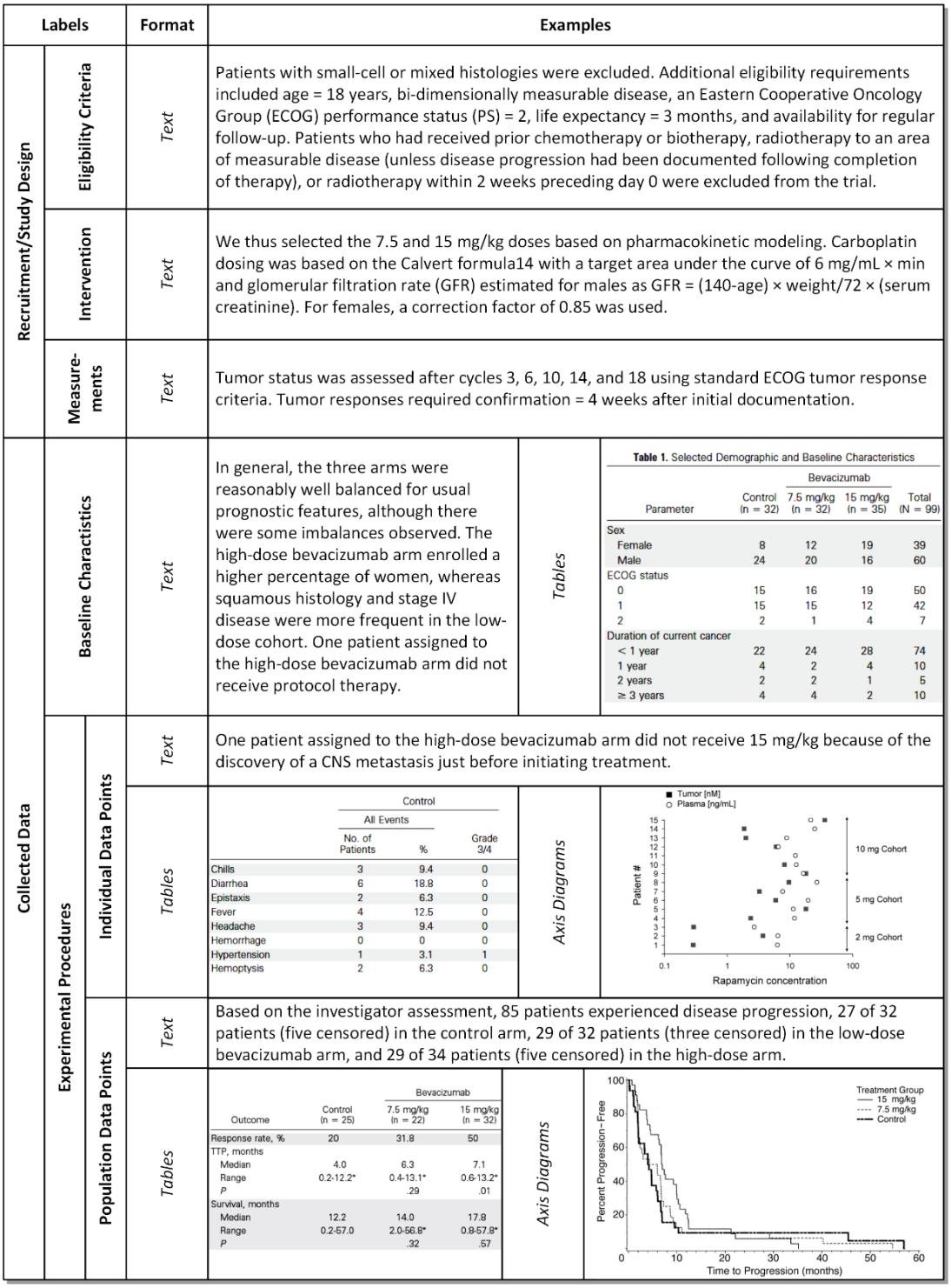


Figure 3-10. Typical examples of numerical data, organized by type.

3.5.2 Hybrid Data Spreadsheet – Process Model Representation Framework

To provide the connection between a data element and its context within the clinical trial design, a framework for organizing information from the clinical trial report was developed based on a hybrid consisting of a spreadsheet structure and the process model. The two-dimensional data grid structure of the representation is similar to a spreadsheet, which is used to specify property values, data constraints**,** and observational summary statistics**.** The rows and columns of the grid correspond to the following dimensions (Figure 3-11):

- Grid Columns (y_i): Each column of the grid corresponds to a different node within the process model. When the process model branches to specify a control arm and one or more intervention arms, separate columns are designated for each node within each arm. The maximum number of columns for the grid then correspond to the number of nodes in the process model. For visualization purposes, the position of a column within the grid (i.e., column number) contain a unique numerical identifier and the identifier matches with the node in the process model for which it is linked.

- Grid Rows (x_i): The rows of the grid correspond to a single variable presented in the paper. Thus, the number of rows in the spreadsheet correspond to the total inventory of variables that are mentioned as part of the study. As part of my ontology work, each row then is linked to an ontological definition within my NSCLC situational ontology. The situational ontology thus standardizes the row's label, its unique properties (most properties are defined by a template – e.g., size uses a template of dimension number, numeric value, unit

of measure, dimension name and value assessment), and also its presentation format (see Section 4.3).

At the intersection of the column and the row of the grid is a cell containing data corresponding to a procedural step and a standardized ontological variable. The characteristics of a cell within the spreadsheet area have the following properties:

- The address of the cell (x_i, y_i) corresponds to a variable (e.g., property), x_i, and node y_i of the process model.
- The value of the cell represents the specification or characterization of a variable, listed in row x_i, at a particular process, denoted by a node in the process diagram, corresponding to column y_i.

The instantiation of the value of a cell is tied to the ontological description of the property type defined by the row number. Thus, the format of the information contained in the cell is flexible and can in general be very different. A cell can be overloaded to hold 1) patient values for a given variable (these are often provided for small sample studies); 2) summary statistics (e.g., median, mean, standard deviation) or graphs; and/or 3) a constraint (e.g., > 18 years of age). Constraints are specific decision nodes to filter patients, and are commonly present in the recruitment process. An example of an overloaded cell can be seen in [Johnson 2004]. The cell of a "survival" node in the control arm and "overall survival" variable is associated with the following characterizations: distribution of a survival curve, and the mean and median months for survival (Figure 3-12). The overloaded cell is intended to handle the diversity and level of specificity for which information within a clinical trial paper are reported.

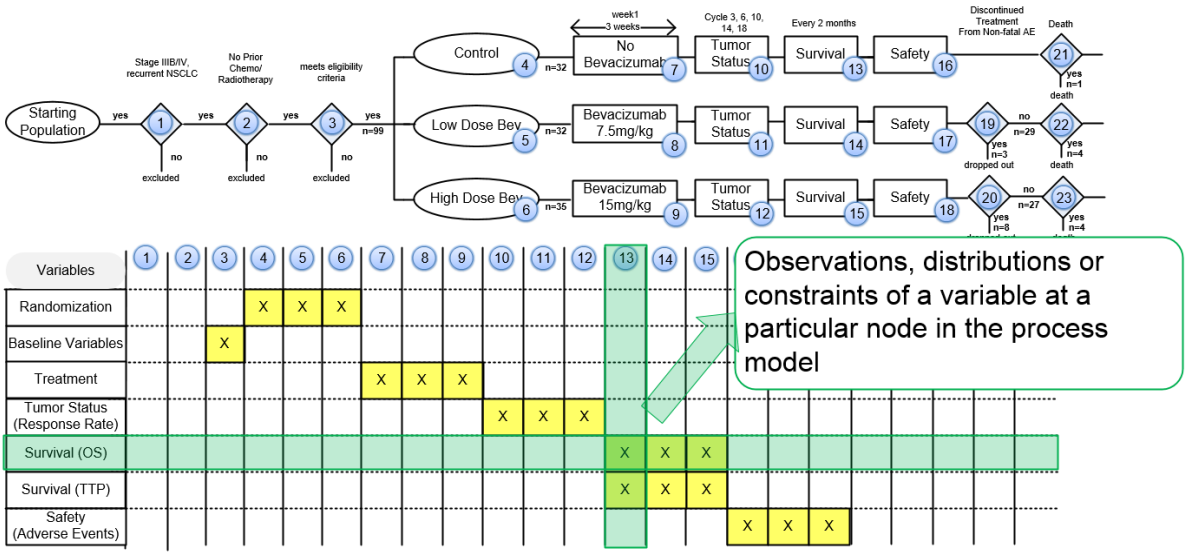


Figure 3-11. Hybrid process model – spreadsheet representation for capturing clinical trial specifics

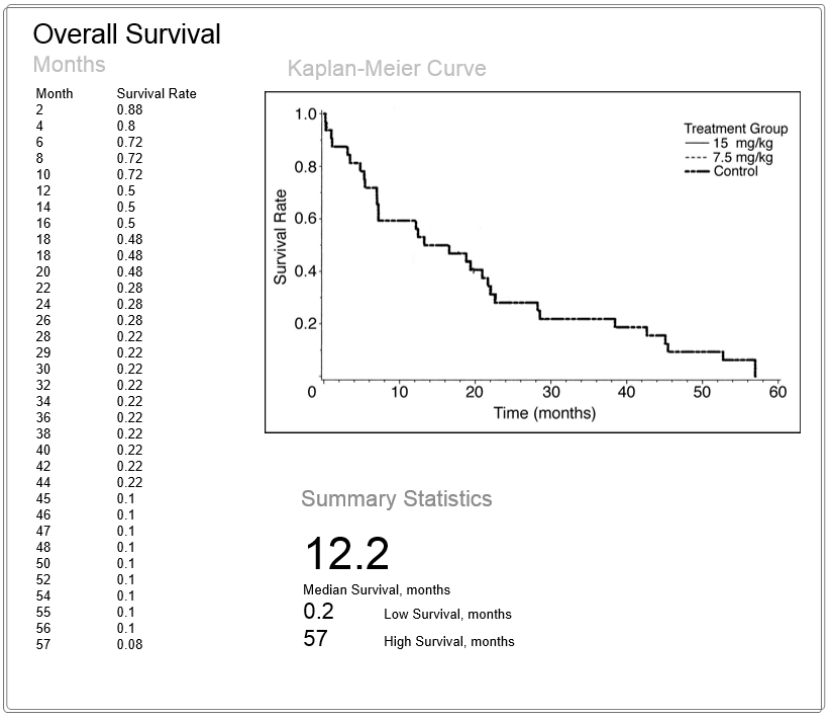


Figure 3-12. Contents of a cell for Survival

In addition to the contents within a cell, additional dimensions of the specification of the property can be present including: 1) time – within a single node of the process model, there may be a temporal component to its description (e.g., timeline of dosing schedule); the property of interest may also have a temporal component; 2) patient index – in recording of patient values, in some studies, the values on a per patient basis are provided. Thus, in general, a recursive frame-based representation (a slot value can be the instance of a frame) for each property is used to accommodate complex descriptions of properties of interest.

In the case of indexing each cell with respect to an individual patient, one can note that the representation can serve as a data collection template for the execution of an on-going clinical trial. This was noted by one of my biostatistician collaborators (Prof. Hyun J. Grace Kim) who has been involved in the design and analysis a number of clinical trial experiments.

3.5.3 Worksheet Area for Statistical Analysis Characterization

The purpose of the hypothesis testing worksheet area is to provide an inventory of all statistical methods performed in a study and to precisely specify the data, the test statistic, and other information required to assess how a hypothesis was tested and the quality of the test. One reason why the detailed context of a calculation is needed is because a statistical significance metric such as the p-value is not a calibrated value and depends on sample size [Marden 2000, Lin 2011], type of hypothesis [Goodman 1988], underlying distributions, effect size, and experimental design [Sestini 2009]. The goal is then to clearly document as much as these dependencies as could be deciphered from the written trial paper and format this information in a consistent manner.

The statistical worksheet area captures the essential information used to properly interpret a statistical analysis. In general, it contains fields that are specialized for each type of analysis method (e.g., test statistics). The lower portion of Figure 3-13 shows an example instantiation of the statistical analysis worksheet area in the context of the experimental flow (Figure 3-13, top) and the data recording spreadsheet area (Figure 3-13, middle). Common types of fields within this worksheet area include:

- Specification of the null hypothesis – Fields correspond to text form, causal agent, effect property, sidedness, size of the effect being tested, etc.
- Specification of the test statistic – Fields correspond to the inputs, outputs, and comments related to the particular test statistic (e.g., log rank test) used to summarize the data. As part of the ontology development, characteristics of common methods were compiled including the types of assumptions implicit in the statistical model.
- Data used to calculate the test statistic – Using the fields are captured in the data grid, the corresponding cells are identified and hyperlinked into the input fields of the test statistic specification.
- Data from the trial experiment from which the test statistics are derived – The input data for a test statistic (if reported) can be easily specified as a reference to a cell within the spreadsheet area of the representation. The cells in the spreadsheet component of the representation include the population context, as captured by its reference path within the process model. This querying of the context associated with a cell is described in detail in Chapter 4.

- Numerical measure of statistical significance – These fields include measures such as the *P*-value, Bayes factor, and hazard ratio.
- Statistical significance of the test – This field includes a statement of the statistical significant of the test (e.g., reject/fail to reject the null hypothesis)
- Clinical significance of the test – This field includes a statement of the practical significance of the conclusion.

This separation and clear indication of data, the test statistics, assumptions, and experiment context was designed to help identify interpretation errors and quality assessment of trial methods [Coultas 2007]. I now demonstrate the representation for statistical analyses with an example. Consider the excerpt from Johnson et al. 2004 [Johnson 2004]:

> *"Survival for the high-dose bevacizumab arm was modestly longer than the control arm (17.7 vs. 14.9; p=0.62...."*

This hypothesis test compared the survival outcome for a high-dose group and a control group. Using the log-ranked test, the test demonstrated longer survival time for the high-dose bevacizumab arm with a non-significant p-value (0.62). Selected details of the hypothesis test are as follows: 1) the test is the log-ranked test; 2) the input data to the test are survival measurements from the high-dose group and the survival measurements from the control group; 3) the *P*-value summarizes the statistical significance of the hypothesis, and in this example, is non-significant; and 4) the interpretation of the data was that survival is modestly longer with the high-dose bevacizumab as compared to the control group.

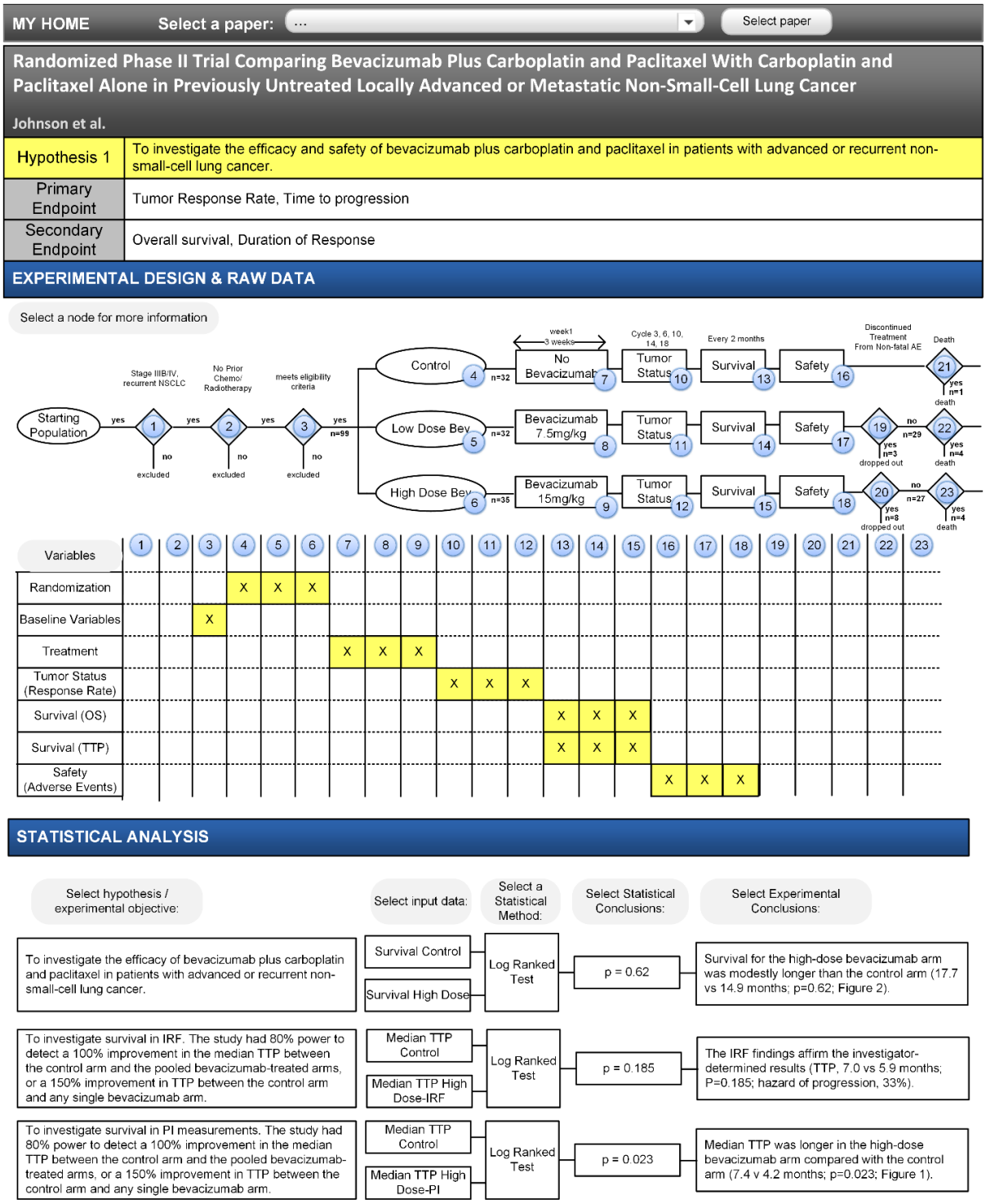


Figure 3-13. Instantiation of statistical analysis worksheet area

Chapter 4 - Query Processing and Visualization Design (SA-2)

4.1 Overview of SA-2 Tasks

Specific Aim 2
To provide a consistent template visualization and query processing engine to support inquiries drawn from the research paper related to concerns of clinicians who are interested in evidence-based medicine and/or biostatisticians who are assessing the quality and/or context of reported numerical information (e.g., observations, frequencies, probabilities, survival curves, and p-values).

In the previous chapter, I described the methods to specify a representation used to characterize the context of a clinical trial experiment. In this chapter, I describe how the representation is implemented into a pipeline incorporating two prototype applications. The first application, the Annotator, is used to instantiate research papers from a PDF file into the computer understandable format and guides the user to populate the fields of the representation. Users interact with the Annotator by answering pre-specified question, and answers are routed to the appropriate spot in the data model of the representation. The second application is a visualization tool driven by the data schema of the representation. The visualization displays the populated data elements in the representation and presents an integrated display of the sections of a clinical trial report. I explain the design of the visualization for the user and the methods used to perform query processing on the representation. As part of the query processing, a discussion of the types of inferences that can be made via the integration of knowledge sources are described.

4.2 Instantiating the Representation

A library of instantiated research papers is needed to test the practicality of the system design. With the help of students and staff at the UCLA Medical Imaging Informatics Laboratory, I developed the Annotator software to accelerate the instantiation of individual research papers into the target representation. In this section, I explain the basic software framework that has been developed, the items that have been worked on during this dissertation effort, and the suggested functionality to be added.

The Annotator is the programming application for model instantiation. Software modules were either borrowed or built upon to develop this application including:

- Open source PDF library – After the user identifies the file location of the PDF file of the journal article, the file is displayed for the user in a Java window. The open source Java PDF library (PDFBox - http://incubator.apache.org/pdfbox/) is used to access and manipulate (i.e., highlight, cut, copy) the content of the PDF file.
- Java Panel for soliciting bibliographic and RCT information - The interface consists of templates soliciting basic reference data about the paper (i.e., title, author, journal, affiliated institutions, digital object identifier PubMed IDs, etc.). Additionally, the type of RCT study is specified including its purpose (e.g., prevention, diagnostic, therapeutic, quality of life) and study phase (e.g., II or IIIB, IV). Finally, reference labels for this report are also solicited from the user (e.g., "Miller 2008 Study", Tarceva NSCLC Trial). This information can be imported from an XML representation (e.g., *BibTex*), that is a common export feature for software such as *EndNote*; and is stored in a MySQL relational database.

- Process Model Visualizer – This java program receives inputs for specifications of a process model and generates a layout for the user. The program uses the Java Swing package.
- Grid Spreadsheet – The grid area is implemented in Java Swing using the JTable class.
- Information Forms – Various Java panels were created to reflect the specification of various classes as defined by the ontology developed for NSCLC and clinical trial reports (see Chapter 3).
- Digitization of x-y Graphs – The program Plot Digitizer (plotdigitizer.sourceforge.net) is used to digitize scanned plots of functional data. This allows data from survival curves to be digitally represented by the system.

The basic layout of the annotator application consisted of three main panels (Figure 4-1):

- The left panel consists of a navigation pane for browsing the sections of a clinical trial PDF report and a sectional map viewer.
- The middle panel displays the annotated contents of the published paper report.
- The right panel contains several forms to assist with populating the representation's data model, including: Paper Information, Hypothesis, Experimental Design, Observational Raw Data, Statistical Analysis, and Interpretation.

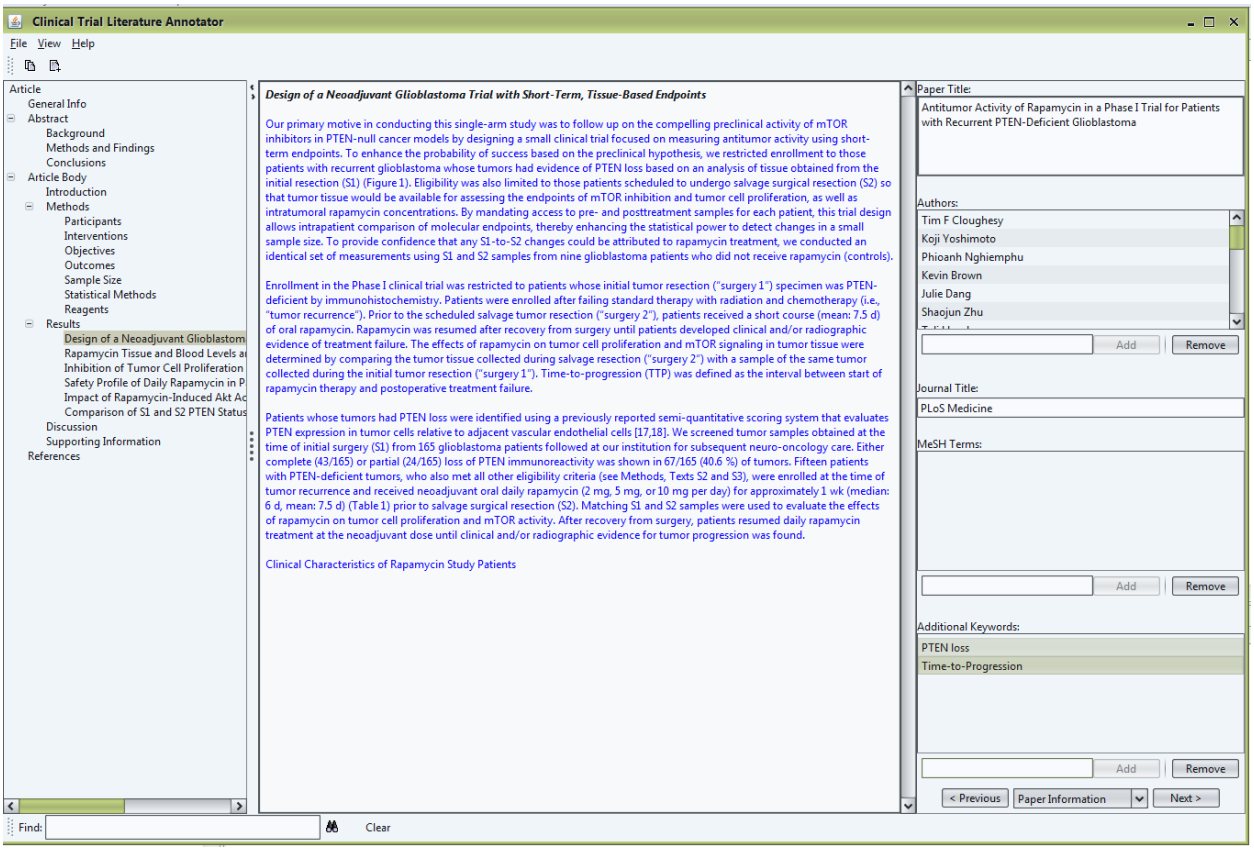


Figure 4-1. Screenshot of Annotator application

4.2.1 Annotation Forms

Paper Information Form: In the Paper Information section of the right panel, several text fields are available to copy, paste, and edit text for the title, author, journal title, and keywords. Upon entering and/or verifying the paper information, a button is available at the bottom of the screen to navigate through the remaining forms.

Hypothesis Form: In the Hypothesis section of the right panel, the hypothesis is entered from the clinical trial PDF viewer into a text field on the left panel. The hypothesis is currently stored as a text field, however, a text field may not be the best data type to store a hypothesis. Free-text

descriptions of hypothesis can introduce unnecessary ambiguity and semi-structured fashion prove to have an advantage over free-text descriptions. Hypotheses are generally broken down into a null and alternative hypothesis. The null hypothesis in practice is almost always stated as a hypothesis which is to be proven wrong [Gigerenzer 2004]. In a semi-formal free-text expression, the theoretical null and alternative hypotheses related to a population can be articulated as:

H_0: Drug has no effect (average mean life expectancy = 12.2 months, even with drug)

H_{alt}: Drug has an effect (normal life expectancy > 12.2 months) when the drug is given

Furthermore, semi-structured free-text can be further disambiguated into fully structured expressions. In the example above, it is uncertain what is meant by the phrase "no effect." The hypothesis needs to be formally stated in terms of a (or a possibly set of) relevant population parameter(s), $\boldsymbol{\theta}$. Assume the population parameter, $\boldsymbol{\theta}$, is the best estimate of the mean value of an observable outcome variable, X. Suppose that the outcome variable, X, represents how long the patient lives in months from the start of a trial; and θ_1 represents the mean months of survival for the drug intervention population, and θ_2 represents the mean months for the control population. Thus, the null and alternative hypotheses can be stated as:

$$H_0: \quad \theta_1 = \theta_2$$

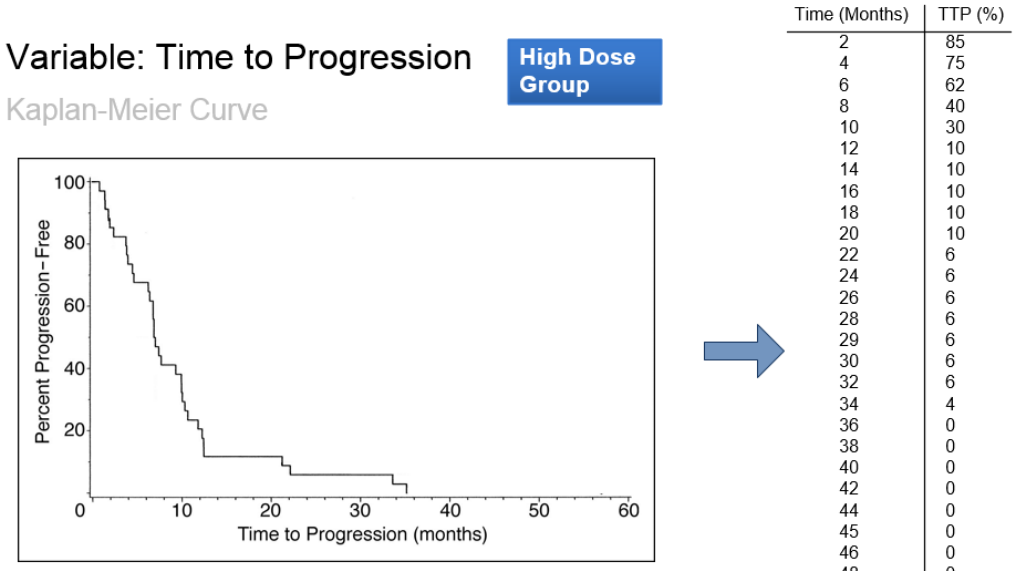
$$H_{alt}: \quad \theta_1 \neq \theta_2$$

Using this presentation, the free-text form of the hypothesis has been translated into precise mathematical terms. As part of future enhancements, semi-structured and structured entry fields defined from the ontology should be developed.

Experimental Design Form: In the Experimental Design section, users generate a process model by first establishing the nodes, and then adding the edges. To create a node, the user selects from a set of node types (e.g., starting population, selection criteria filter, randomization methods, control point, observational point, interventional arm, side effects, etc.), and provides information specific to that node class. As part of the node specification, a drawing grid allows users to indicate the positioning of nodes. After creating a node, the annotator assists with defining edges by providing a real-time drawing of the nodes and edges entered in the data model. Next, linkages between nodes can be specified to complete the creation of a process model. To decrease redundant efforts, future enhancement should include an inventory of common types of experimental design flows so that users can simply create a new instance of a process model by modifying an existing or generic experimental design configuration.

Observational Raw Data Form: The observational data, other types of information characterizing the state of a variable, or summarization of a variable is entered in this section of the interface. The property name and their possible values have been mined from within the ontology of Chapter 3 and/or are added to the ontology as needed. The key to the annotation process requires users to specify a node within the process model section of the representation to serve as a reference as the context for data entered. Data can be entered either as individual data points, or as a batch upload. Batch upload is ideal for cases where *x-y* graph data have been digitized using the open source software Plot Digitizer. To use the batch upload for digitized data, users would run the Plot Digitizer software on a selected axes diagram. The software would generate x-y coordinates while accounting for the scales of the x- and y-axis. Afterwards, the x-y coordinates can be uploaded to the database using the batch feature. For example, in [Johnson

2004], survival data for every two months in the high dose bevacizumab experimental arm is extracted from a Kaplan Meier curve into x-y coordinates (Figure 4-2).



Time (Months)	TTP (%)
2	85
4	75
6	62
8	40
10	30
12	10
14	10
16	10
18	10
20	10
22	6
24	6
26	6
28	6
29	6
30	6
32	6
34	4
36	0
38	0
40	0
42	0
44	0
45	0
46	0
48	0
50	0
52	0
54	0
55	0
57	0

Figure 4-2. Plot digitizer for survival data

As part of the functionality, observational data can also include image data files, such as those used in pathology and radiology studies. Future enhancements can include image mark-ups as raw data or an additional functionality to annotate images.

Statistical Analysis Form: The statistical analysis section allows users to annotate the details of statistical methods used to test a hypothesis. The interface is designed to allow users to reference a hypothesis, participants, interventions and the data inputs used for the test statistic reported. Figure 4-3 shows a few examples of how free-text extracted from a clinical trial paper can be represented. The population groups (i.e., 'Group A' and 'Group B') would be specified using the

combined modeling of the process model together with the property constraints within the data grid. Section 4.4 on query processing provides further details on how these statistical analyses are linked to the process model and data grid. Future enhancement can include an inventory of common types of statistical methods, with templates for their inputs and parameters displayed in the interface. Windish et al. tallied the statistical methods used in 239 original research articles, which can be a starting point for a situational ontology for statistical methods. Table 4-1 shows the most common types of statistical methods used in medical research as surveyed by [Windish 2007].

Text Excerpts	Structured elements of Statistical Hypothesis Testing

With the exception of ≤ 15 pack-year history of smoking, no clinical factors were associated with higher response rate RR (45% versus 5%, p<.01).	Statistical Hypothesis testing:
	• Group A: smoking status ≤ 15 pack-years;
	• Group B: smoking status > 15 pack-years;
	• Null hypothesis –
	  ○ RR (Group A) = RR (Group B)
	• Actual Observation:
	  ○ RR(group A) = 45%;
	  ○ RR(group B) = 5%;
	• p-value < .01

Patients with an EGFR mutation had a longer PFS (13 versus 2 months; P<.01) and a trend toward improved OS (23 v 17 months; P=.24)	Statistical Hypothesis testing:
	• Group A: EGFR mutation = true;
	• Group B: EGFR mutation = false;
	• Null hypothesis 1 – PFS (Group A) = PFS (Group B)
	  ○ Actual Observation:
	  ▪ RR(group A) = 13 months;
	  ▪ RR(group B) = 2 months;
	  ○ p-value < .01.
	• Null hypothesis 2 – OS (Group A) = OS (Group B)
	  ○ Actual Observation:
	  ▪ OS(group A) = 23 months;
	  ▪ OS(group B) = 17 months;
	  ○ p-value < .24

Figure 4-3. Examples of representations for statistical methods. PFS stands for progression free survival

Type of Test	Number (%)
Descriptive statistics	219 (91.6)
Simple statistics	120 (50.2)
Chi-Squared Analysis	70 (29.3)
t-Test	48 (20.1)
Kaplan-Meier Analysis	48 (20.1)
Wilcoxon Rank Sum Test	38 (15.9)
Fisher Exact Test	33 (13.8)
Analysis of Variance	21 (8.8)
Correlation	16 (6.7)
Multivariate Statistics	164 (68.6)
Cox Proportional Hazards	64 (26.8)
Multiple Logistic Regression	54 (22.6)
Multiple Linear Regression	7 (2.9)
Other Regression Analysis*	5 (2.1)

Table 4-1. Common statistical methods used in medical research [Windish 2007]

Interpretation Form: For the Interpretation section, a text field is available to copy, paste, and edit text from the clinical trial PDF viewer on the left panel, similar to the Hypothesis form. The interpretation fields include a free-text field for statistical interpretation and a separate free-text field for practical/clinical significance. Unique to this section, the Interpretation section can be used to help clarify unclear language in reporting interpretations. The language used to report interpretations can be a source of confusion to readers, if readers are bordering on whether interpretations are statistically sound and are deciding what are judgements and non-conclusive trends made by the writers toward the effect. An example of confusing language involves the p-value. In some rare cases, the p-value described is reported inconsistently throughout the paper. In the RCT paper by [Miller 2008], text excerpts were encountered with contradictory statements regarding the p-value stated in the text and value presented in a table. In this case, the connection between whether the effect is or is not significant was ambiguous to a reader. More commonly, the source of confusion in the literature and textbooks is the use of the symbol α with regard to

hypothesis testing. The measure for statistical significance is the controversial decision criterion, represented as:

$$p - value < \alpha$$

This criterion, or α-level, is the central product that has led to the widespread misunderstanding of the p-value and classical testing [Gigerenzer 2004]. In the literature, the phrase "level of significance" is used liberally in a number of different contexts and actually refers to three different philosophically distinct definitions, which to most non-statistician researchers are not entirely obvious. The ambiguity of the symbol α stems from the competing philosophical Frequentist views of Fisher and Neyman-Pearson for statistical testing [Goodman 1999]. The hybrid approach in today's use inconsistently combines Fisher's calculation of a p-value and Neyman-Pearson's rule-based. The three different references of this phrase include:

1. The standard level of significance, a conventional standard for all researchers, is simply a p-value threshold, typically 0.001, 0.01 or 0.05 (Early Fisher). These are common preset thresholds that are used for all experiments that calculate p-values. For the scope of this dissertation, I refer to this threshold value as α_{Thresh}.
2. The exact level of significance is determined after the experiment and is represented by the exact value of the p-value (late Fisher). The exact level of significance reflects a relation between the experimental data and theory. In this context, α is a property of the data.
3. The level of statistical significance, as characterized by the α level, is the relative frequency of Type I errors in the long run, decided on by using cost-benefit considerations before the experiment (Neyman and Pearson). Like the p-value, it is determined assuming that the

null hypothesis is true. In this context, α is a property of the test, not the data. For the scope of this dissertation, I refer to this definition of level of significance as $\alpha_{Type\ I}$.

The symbol α is ambiguously used generally to refer to both uses 1 and 3 above. For this reason, further enhancements could implement standard requirements to characterize each α-level specified. I distinguish its use in regard to the first definition as α_{Thresh} and $\alpha_{Type\ \mathrm{I}}$ for the third definition. As part of future plans for annotating interpretations, the distinction between the currently ambiguous use of the term "level of significance" is proposed here and recommended.

4.2.2 Annotation Guidelines

While annotation guidelines are currently being researched, there is little or no consensus on the type of information that should be collected from a clinical trial report and the format of this information. Moreover, clinical trial reports are extremely rich with information, requiring extraction of knowledge to be almost entirely conducted by manual efforts.

I experimented with the implementation using a variety of clinical trial study reports to demonstrate the robustness of the system to extract a diversity of information within my domain of NSCLC and the ability to put all information in the appropriate fields in the representation. The result was a set of annotation guidelines for keeping track of how information is generated in a clinical trial study; and an interface with a specific line of questioning to collect and populate user entries into the appropriate fields in the representation. While creating the annotation guidelines, I tried to balance the trade-off between scalability (allowing for the representation to be populated quickly) and performance (allowing a variety of information to be fully captured).

As part of the learning process to define how users should annotate clinical trial reports, an early version of the prototype annotator application was demonstrated at a week-long computer informatics exhibit at the Radiological Society of North America in 2010 [Tong 2010]. In this exhibit, I demonstrated the concept of how to review, annotate and structure trial reports in the domain of brain tumors. The system demonstrated to a user how to download (e.g., via PubMed) the text of a journal article, and took the user through the analysis of each part of the study, vetting the results manually. This extracted information was then stored in the database together with past structuring results involving similar hypotheses or interventions. Information from the database can then be retrieved, culminating in a graph of interrelated study variables. During the week long exhibit, several hundred people visited the booth, and their feedback and recommendations were valuable in re-sculpting the system's requirement specifications. A few of the most common comments are paraphrased below:

> *"I appreciate the organization and am especially interested in the results of a fully annotated paper."*
>
> *"The line of questioning helps me figure out what I need to understanding from a clinical trial paper"*
>
> *"I would like to see this extended to clinical reports and physician notes."*

In another experiment testing the validity of annotations guidelines, I informally tested how a number of students could instantiate paper versions of the representation. In spring of 2011 and 2012, students from the graduate course BE226 (Bioengineering 226 – Medical Knowledge Representation) were each assigned to select a clinical trial study and create a poster (4' x 8') of

the papers representation according to the annotation guidelines. In total, each of the students (n=15) were able to create poster representations of the trials, which was put on display and reviewed by various UCLA faculty and students. The posters were viewed positively by a number of researchers and biostatisticians for their clarity and structure. A few positive comments are paraphrased below:

> *"I wish I could teach clinical trial design using this representation"*
>
> Professor James Sayre (Biostatistics and Radiology)

> *"This representation also seems very nice for organizing and collecting data during the execution of the trial as well."*
>
> Professor Hyun J. Grace Kim (Biostatistics and Radiology)

By instantiating a number of paper-based forms of the representation, I discovered ways to improve the representation and hence improve its specification. Asking users to instantiate paper versions of selected clinical trial research papers confirmed that at least potentially, the representation can be instantiated using the guidelines specified.

The diversity of natural language, the clarity of writing, the complexity of experimental designs, etc. all contribute to difficulties in developing a mature annotation system for even a focused field, such as NSCLC. The creation of a general line of questioning to be inclusive enough to annotate all variations of experiment types within clinical trial reports required more years of development and is outside the scope of this dissertation. My goal in this dissertation is, thus, to concentrate on the specification of a representation for clinical trial paper to support quality assessment and/or

evidence based medicine. A basic framework for the representation has been developed and extensions to the representation can be built to allow additional operations.

4.3 Visualization Design

The visualization of large complex information spaces can greatly affect a user's acceptance and ability to optimally benefit from the knowledge stored within the system. A visualization was developed to assist with viewing and interacting with the contents of the representation from each clinical trial report (Figure 4-4). Visualization of the clinical trial study was designed to be as consistent as possible across a general sample of clinical trial reports, and intuitive in its organization, navigation, and query formulation. It is intended to provide a template spatial-organization layout of information, such that navigation for information is made easy. Specific features include:

- The visualization allows for summary views of a single clinical trial report. On one view, the purpose of the trial, a visual display of the recruitment and experimental procedures, a list of statistical methods and values, and a list of interpretations is provided.
- The visualization helps with understanding a hypothesis by connected fragmented information related to a particular hypothesis together. Hyperlinks between process model nodes, data grid cells, and worksheet inputs allow users to easily track related data items.
- The visualization shows an overview of all variables being described in the paper. Various semantic links between variables, inherent and inferred from ontologic relationships, were defined in the previously described ontology for NSCLC (Section 3.4).

- The base visualization includes two main linked components: a process model diagram and a synchronized "spread-sheet" like interface.

Figure 4-4 shows the basic layout of the visualization including areas for stating the purpose of the study, an area designated to display the experimental design and associated reported data / states of variables collected/reported at each stage of the experiment, a statistical analysis area and an interpretation area. Figure 4-5 shows a conceptual illustration of the instantiation for a single research paper.

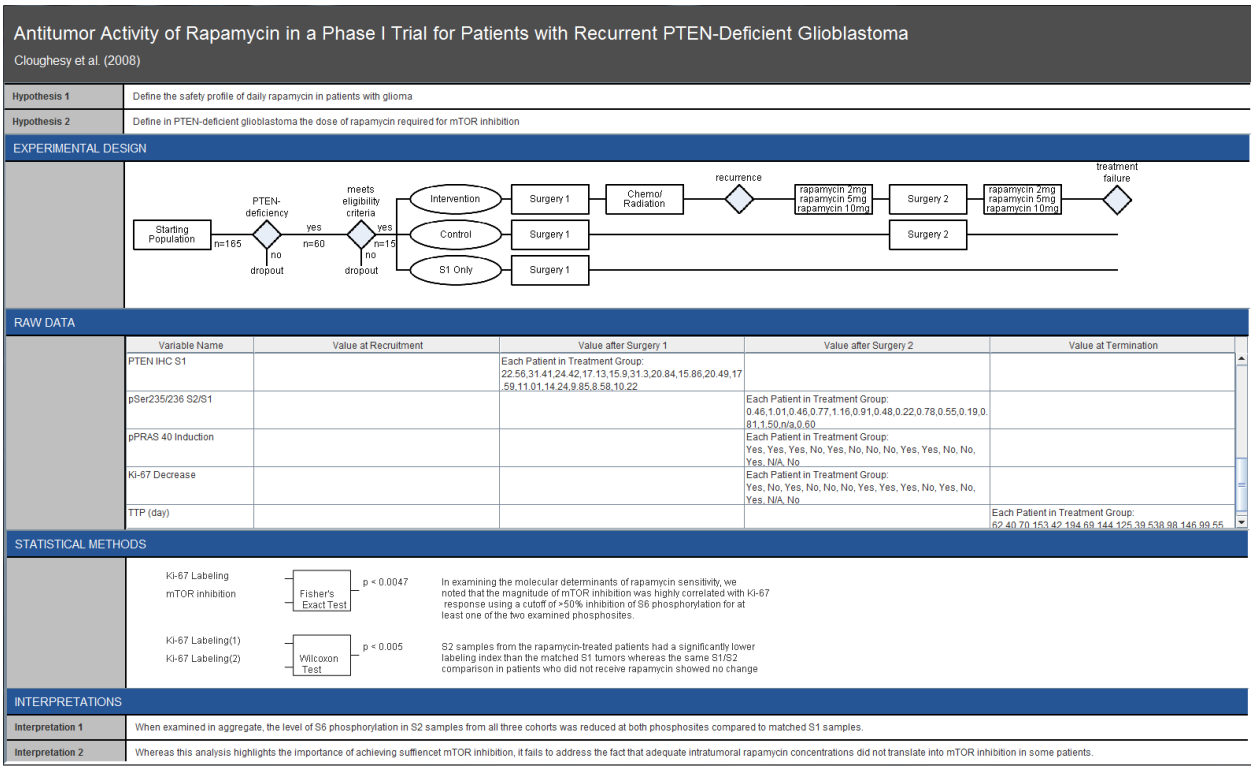


Figure 4-4. Screenshot of basic visualization layout

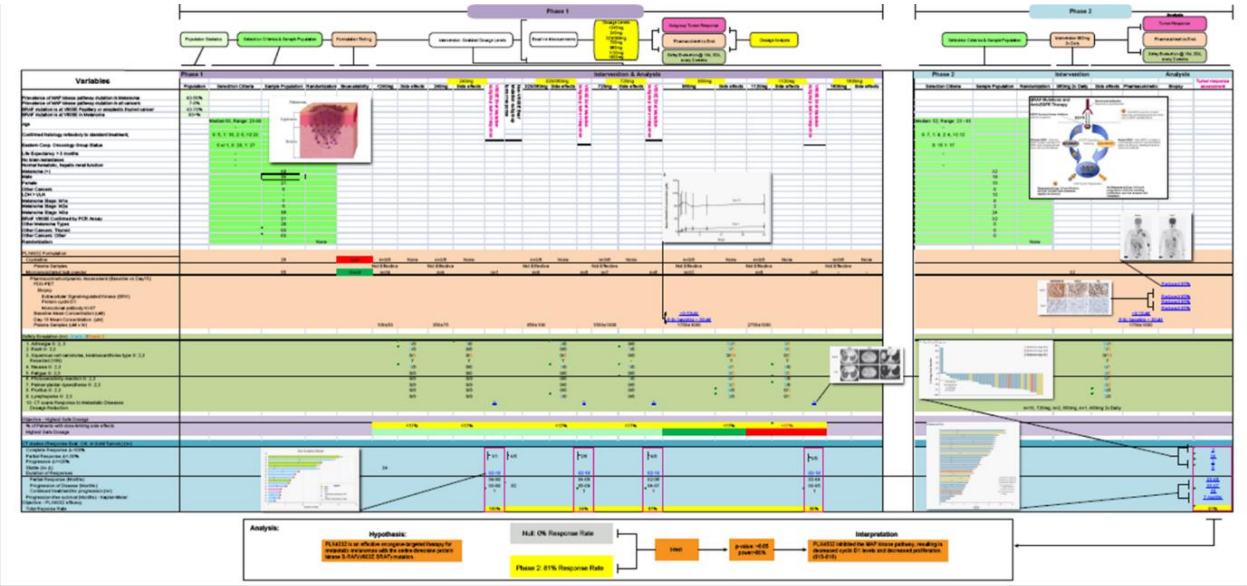


Figure 4-5. Example illustration of the visualization layout for RCT paper

Hypothesis: The top portion of Figure 4-4 is an area to show the objective of the study. This is presented to the user in free-text as well as fields defined within the ontology (e.g., primary outcomes, secondary outcomes, statistical parameter involved in null hypothesis, etc.)

Process Model: Below the hypothesis area shows the designated placement of the study flowchart. As previously noted, the process model documents the flow of events and populations associated with the clinical trial. The complexity of the models was designed to not be unnecessarily complicated when depicting the experimental context for how variables were obtained or stated. A time component can also be specified with the process model using annotated links, where links between connected elements in the flow diagram can be annotated to specify a time period.

Spreadsheet Data Grid Area: The spreadsheet contains the variable constraints and the recorded values and/or associated summary statistic related to a variable. This structuring framework is a

key innovation of this proposal by allowing knowledge/informational fragments to not only be structured, but also placed in the context of the entire experiment.

Variable List: The left column of the spreadsheet area contains a list of all unique variables described in the research paper (see Figure 4-6). Importantly, each variable is mapped to on ontologic concept within the NSCLC RCT situational ontology described in Chapter 3. The ontological relationships defined in this knowledge base allows the visualization to intelligently group similar variables together via similarity (i.e., belonging to the same class or superclass) and/or via association with a common frame head (i.e., belonging to a semantic frame category). The ontologic normalization of variables also allows the list of variables to be linked to external knowledge sources.

One possible use case for linking variables to external knowledge sources is the possibility of identifying confounders. Causal graphical models of a disease (e.g., NSCLC) could be used to identify possible confounding relationships between two variables. This insight could help in the assessment of quality, by checking if the investigators had controlled for such known confounders.

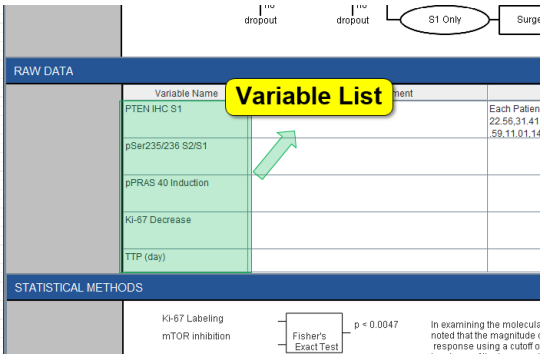


Figure 4-6. Left-hand side of the data grid area

Variable Characterization Area: The variable characterization area of the data grid corresponds to values and/or constraints assigned to each row property at a given node in the process model (Figure 4-7). Thus, each cell in the data grid is associated with an event node from the flow diagram for a particular variable, and the cell itself corresponds to the specifications or characterization of a variable for an experimental procedure of a group of patients for which the node refers to in the flow chart. A cell's value can be semantically overloaded – in an object-oriented sense – depending upon the semantics of the property in question. For example, a cell's value can show: 1) a categorical semantic state; 2) the individual values for each patient for a given variable, if available; 3) the entire distribution for a given variable over the sampled population; 4) summary statistics of the distribution. To visualize the contents of a cell, a limited number of customized visualizations have been developed (e.g., survival curves and pie charts – Figures 4-8a and 4-8b) and each variable type may require a custom module. In addition, functions can be used to derive useful information from the cells assigned data. For example, in Kaplan-Meier curves, these functions can provide a quantitative measure of the difference between two curves, in addition to showing the curve. These functions are left for future work.

Statistical Methods Worksheet Area: The panel for statistical methods provides a visual inventory of all the tests performed. Each test is listed with its corresponding inputs from the data grid, the test statistic, output statistics such as a p-value, and a statement of significance. This portion of the representation uses seven fields to specify the results of an analysis (Figure 4-9): 1) text of the hypothesis; 2) the property of the outcome data being tested; 3) input data - from the data grid, the annotation process involves identification of the comparison cells that are part of the hypothesis testing modules; 4) the test statistics deployed (e.g., Fisher F-test, mean, hazard ratio,

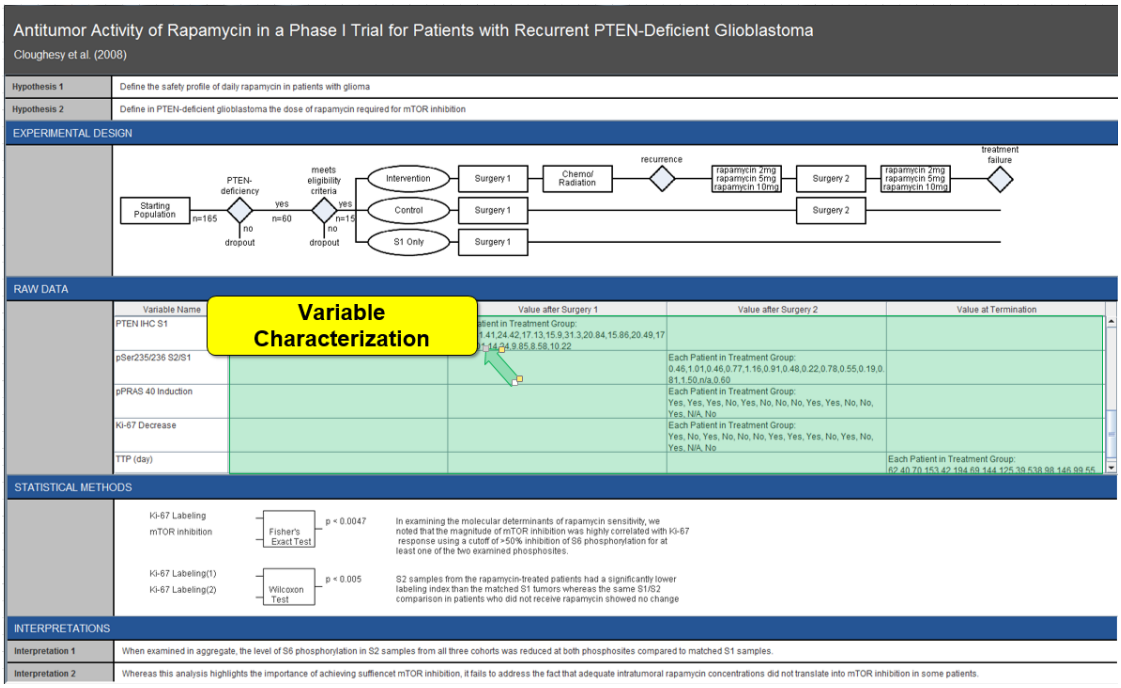
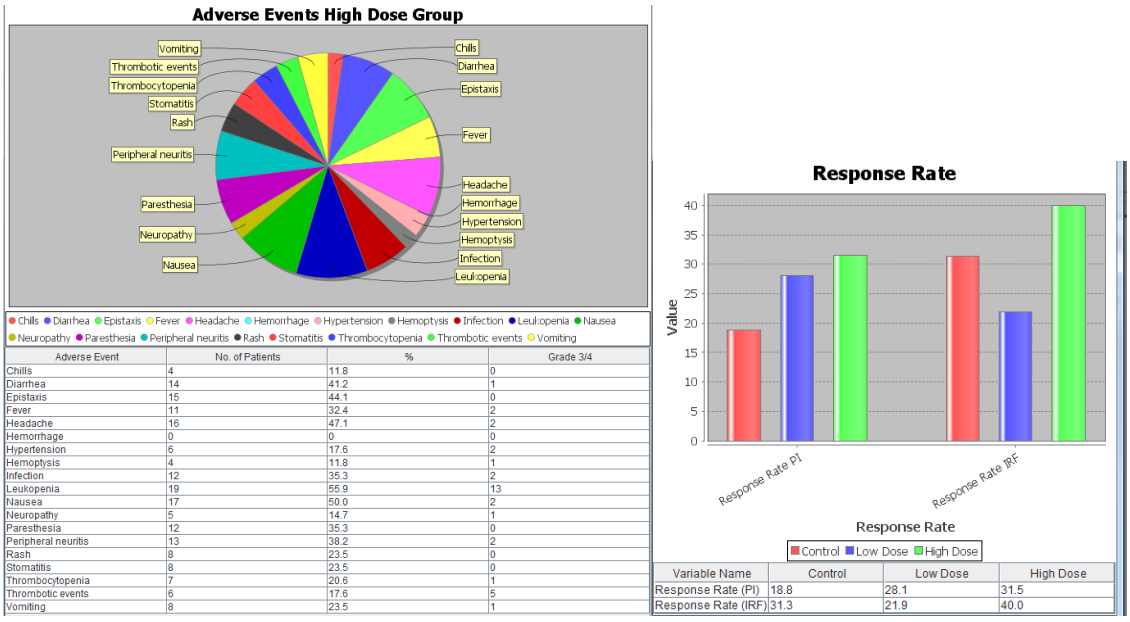


Figure 4-7. Variable characterization area of the data grid



Adverse Event	No. of Patients	%	Grade 3/4
Chills	4	11.8	0
Diarrhea	14	41.2	1
Epistaxis	15	44.1	0
Fever	11	32.4	2
Headache	16	47.1	2
Hemorrhage	0	0	0
Hypertension	6	17.6	2
Hemoptysis	4	11.8	1
Infection	12	35.3	2
Leukopenia	19	55.9	13
Nausea	17	50.0	2
Neuropathy	5	14.7	1
Paresthesia	12	35.3	0
Peripheral neuritis	13	38.2	2
Rash	8	23.5	0
Stomatitis	8	23.5	0
Thrombocytopenia	7	20.6	1
Thrombotic events	6	17.6	5
Vomiting	8	23.5	1

Variable Name	Control	Low Dose	High Dose
Response Rate (PI)	18.8	28.1	31.5
Response Rate (IRF)	31.3	21.9	40.0

Figure 4-8. Summarization of adverse effects for the high dose experimental arm (A), and comparison of response rate in control and experimental arm (B).

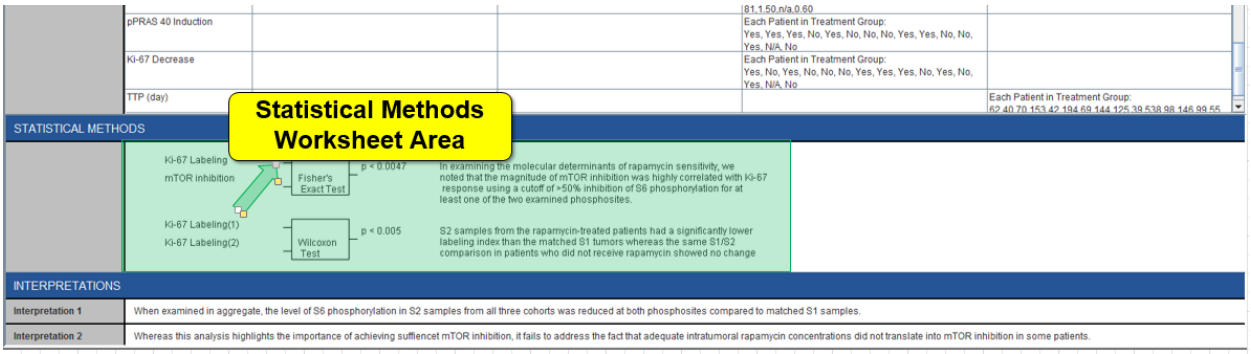


Figure 4-9. Statistical evaluation worksheet area of visualization

etc.); 5) numerical output of statistical test; 6) statistical interpretation (statistical significance level of the test / reject null hypothesis); and 7) clinical significance of test.

4.4 Query Processing and Inferencing

Ultimately, the representation and visualization methods developed must support the intended queries stated in Section 3.3.

4.4.1 Queries for Specific Papers

The prototype database to store the representation was developed in this dissertation using MySQL. The query formation to search for specific papers is facilitated by the NSCLC ontology. The NSCLC ontology has been compiled to include entries for chemo- or radiotherapies, properties, and statistical methods, and all the sanctioned values, or "states," that entries can assume. For example, the recruitment class in the ontology includes all typical attributes for study inclusion and exclusion in the domain of lung cancer. When instantiated with a user's search constraints, the recruitment class can be used to generate a SQL relational query for a given paper.

Thus, queries can be fashioned by simply providing a structured form for users to fill out. Similar types of queries for causal agents and study design type can be generated in the same way.

4.4.2 Queries Related to a Node in the Process Model

For a selected paper, users may wish to view the data and/or constraints related to a certain node in the process model. Within the framework of the representation, each process model node is hyperlinked to a column in the data grid, and each row in the column contains relevant concepts. For example, if users are interested in the study inclusion criteria, they can click on the corresponding graphical node in the process model area of the visualization. The processing of the query can quickly locate the cells in the data grid relevant to the user's query, given that the inclusion criteria has an ontologic definition that specifies the types of possible attributes (e.g., age, sex, therapy history, smoking status, etc.), and return the data within these cells.

In another example, suppose a user wants to view the interventions of the clinical trial study by [Johnson 2004]. This query involves selecting the node "High Dose Bev" in Figure 4-10. After this node is selected, properties associated with this node can be found from the designated column in the spreadsheet area of the representation, summarized in Figure 4-11. The drug administered is bevacizumab at a dose of 15 mg/kg using an intravenous infusion over 90 minutes. The drug cycle was 3 weeks with a maximum dose of 18 doses. If the drug was not well tolerated, this can be denoted in the exceptions field under new action, such as reducing the dose from 90 minutes to 30-60 minutes. The information is abstracted into the fields: "Cell Name," "Drug," "Dose," "How it was administered." Under field "How it was administered," the ontology includes the following

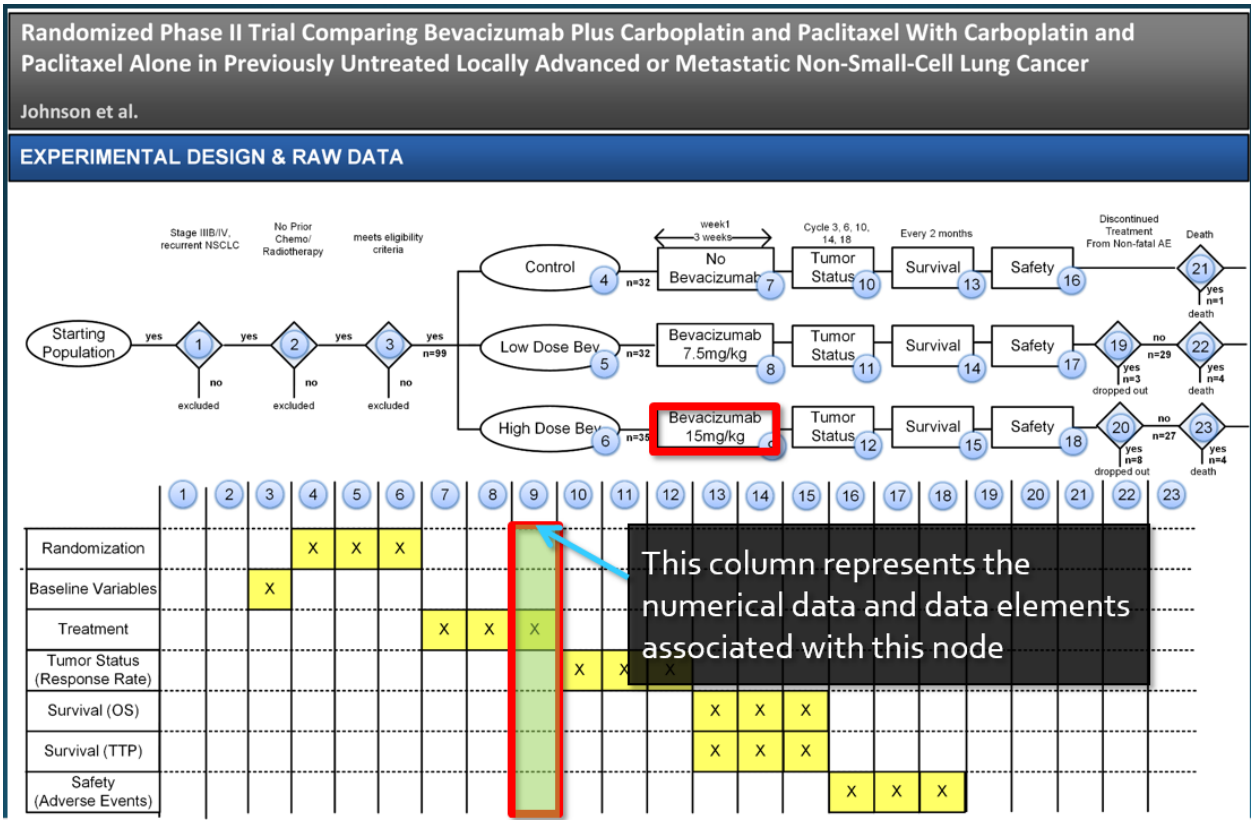


Figure 4-10. Queries related to the intervention node, "Bevacizumab 15mg/kg"

```
Cell name: Bevacizumab
Drug: Bevacizumab
Dose: 15 mg/kg
How was it administered:
   Vehicle:  Intravenous infusion
   Duration: Over 90 minutes
   Cycle: 3 weeks
   Maximum dose: 18 doses
   Exception: Well tolerated
      Resulting Action: New duration
      Duration: 30-60 minutes
```

Figure 4-11. Drug administration details recovered from the node "Bevacizumab 15mg/kg" in process model

fields: "Vehicle," "Duration," "Cycle," "Maximum dose", and "Exception." Under the field "Exception," the ontology includes the fields: "Resulting Action," and "Duration."

4.4.3 Queries Related to a Cell in the Data Grid Area

An important class of queries the system is designed to support is related to providing context for a given observational value. An example user query can be: What is the context associated with a frequency that is reported for a property in the data grid area? The context consists of the sample population, observational method, and/or interventional details. Without this context, the appropriate interpretation of such observational properties remains difficult to realize and in the worst case, interpretations can be misapplied.

The query processing steps to reconstruct the context for data reported in a cell is as follows. The example references the paper shown in Figure 4-12. In this query, the node of interest refers to the “Safety” node of the high dose intervention arm. The user first selects the node of interest in the process model, afterwards, the representation performs the following steps:

1. Find the corresponding column. Each node in the process model is hyperlinked to a corresponding unique column in the data grid area.
2. Find the variable of interest. Each row in the data column corresponds to a collection of variables. Within the data column, the variable of interest is identified. In this case, it is the row containing the property “hemoptysis.” Note that in Figure 4-12, the group of variables for adverse events is collapsed in the row labeled “Safety (Adverse Events).”

3. Backtrack through the process model to obtain context for observations and get associated data to each backtracked node. Thus, given the path for a target node in the process model, one can map the nodes of the path to columns in the data grid area, which indicate context. In the current example, the backtracking through the process model identifies the nodes: “Imaging," “Survival,” “Tumor status,” “Intervention,” “Baseline,” etc. (see Figure 4-13). The highlighted process model path leads from the start node, “Starting Population,” to the node of interest “Adverse Events” for the high dose intervention arm. The path can then be used to recover the context of observations made at any point in the experimental procedure.
4. Construct logical representation of context. Each process node within the backtracking path contains a different set of variables and values. The information for each node and its set of variable values is compiled to detail the context for the target variable.
5. Repeat steps 4-5 until the start node. After the query is complete, the flow identifies the eligibility criteria and the experimental procedures. Nodes in the flow for eligibility criteria include "Stage NSCLC cancer," "Prior Chemo Radiotherapy," and "Other Eligibility Criteria." Nodes in the flow for experimental procedures include “Baseline,” “Intervention,” “Tumor Status,” “Survival,” “Imaging,” and “Adverse Events.”

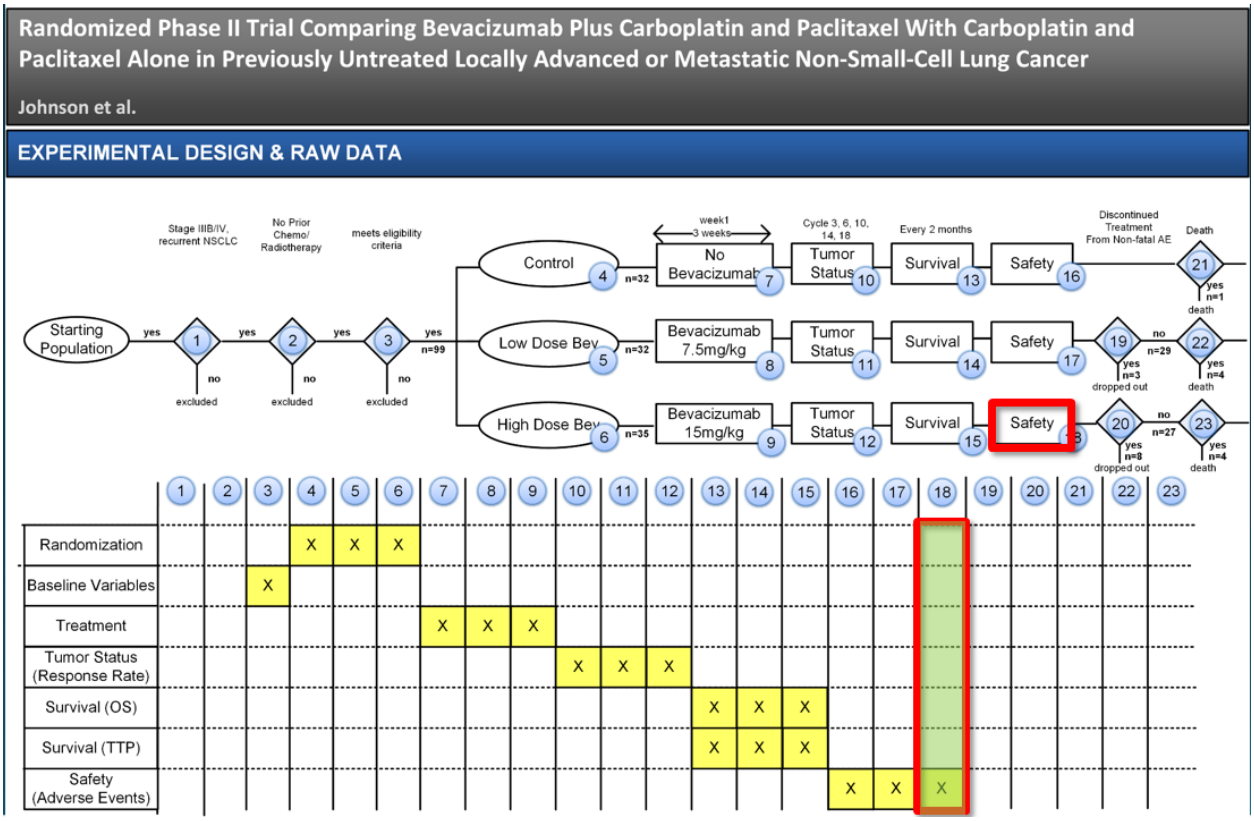


Figure 4-12. Example for identifying context for a reported frequency of observation

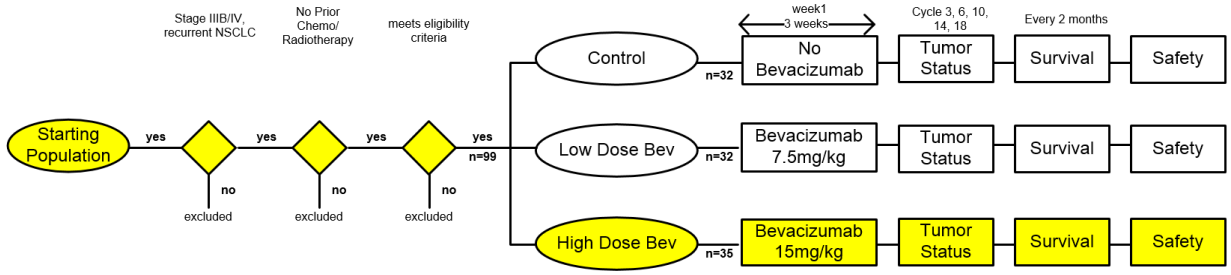


Figure 4-13. Highlighted process model path leading to the node of interest “Adverse Events” for the high dose intervention arm

To summarize, after identifying the node of choice, “Safety,” denoted in a red box, a pathway can be constructed demonstrating the result itself and how the numerical value was generated. For this observation, one can trace the flow through the pathway starting from the "Starting Population"

node on the very left of the figure in Figure 4-13. Thus, the result of this query allows the user to follow the semantics of the process model, and backtrack through the process model. Each process node within the path provides a different part of the context. After the variables and values within each node are aggregated and compiled, one can use relational information from the ontology (e.g., frame definitions) to visually display the conditions (i.e., context) associated with a particular observation. Because information is structured and explicitly linked to the process model and variable list, the representation can provide information for interpreting a particular probability and identify other factors that may have contributed to the result.

Following identification of context, the value of the probability can be estimated from the query results. The numerical information for "Adverse Events" under the high dose population is displayed in the data grid. The numerical data itself is captured as a table of frequencies per adverse event (Figure 4-14). The table contains the types of adverse events, the number of patients having that adverse event, the percentage, and the subset of patients with adverse events of grade 3 or 4.

The notion of survival outcomes for this clinical trial is one key piece of evidence and the outcomes are typically expressed as a Kaplan-Meier survival curve. The probability of survival for a time point in the sample population can be estimated from a Kaplan-Meier survival curve, where each time point displays the proportion of patients surviving in the high dose group. As a second example of cell-level observational context, specific for the clinical trial in our running example, a user query can be posed to investigate survival in the high dose group (Figure 4-15). Note that context is extracted in the same way as in the previous section by using the backtracking path for

	Control			7.5 mg/kg			15 mg/kg		
	All Events			All Events			All Events		
	No. of Patients	%	Grade 3/4	No. of Patients	%	Grade 3/4	No. of Patients	%	Grade 3/4
Chills	3	9.4	0	4	12.5	0	4	11.8	0
Diarrhea	6	18.8	0	9	28.1	3	14	41.2	1
Epistaxis	2	6.3	0	10	31.3	0	15	44.1	0
Fever	4	12.5	0	11	34.4	2	11	32.4	2
Headache	3	9.4	0	10	31.3	1	16	47.1	2
Hemorrhage	0	0	0	4	12.5	2	0	0	0
Hypertension	1	3.1	1	5	15.6	0	6	17.6	2
Hemoptysis	2	6.3	0	9	28.1	3	4	11.8	1
Infection	8	25	1	10	31.3	0	12	35.3	2
Leukopenia	10	31.3	7	15	46.9	10	19	55.9	13
Nausea	15	46.9	1	16	50	1	17	50	2
Neuropathy	9	28.1	0	4	12.5	0	5	14.7	1
Paresthesia	7	21.9	0	9	28.1	0	12	35.3	0
Peripheral neuritis	9	28.1	1	8	25	0	13	38.2	2
Rash	3	9.4	0	11	34.4	0	8	23.5	0
Stomatitis	3	9.4	0	5	15.6	0	8	23.5	0
Thrombocytopenia	5	15.6	0	2	6.3	0	7	20.6	1
Thrombotic events	3	9.4	3	4	12.5	2	6	17.6	5
Vomiting	6	18.8	1	6	18.8	1	8	23.5	1

Figure 4-14. Data embedded within the node of interest “Adverse Events” for the high dose intervention arm

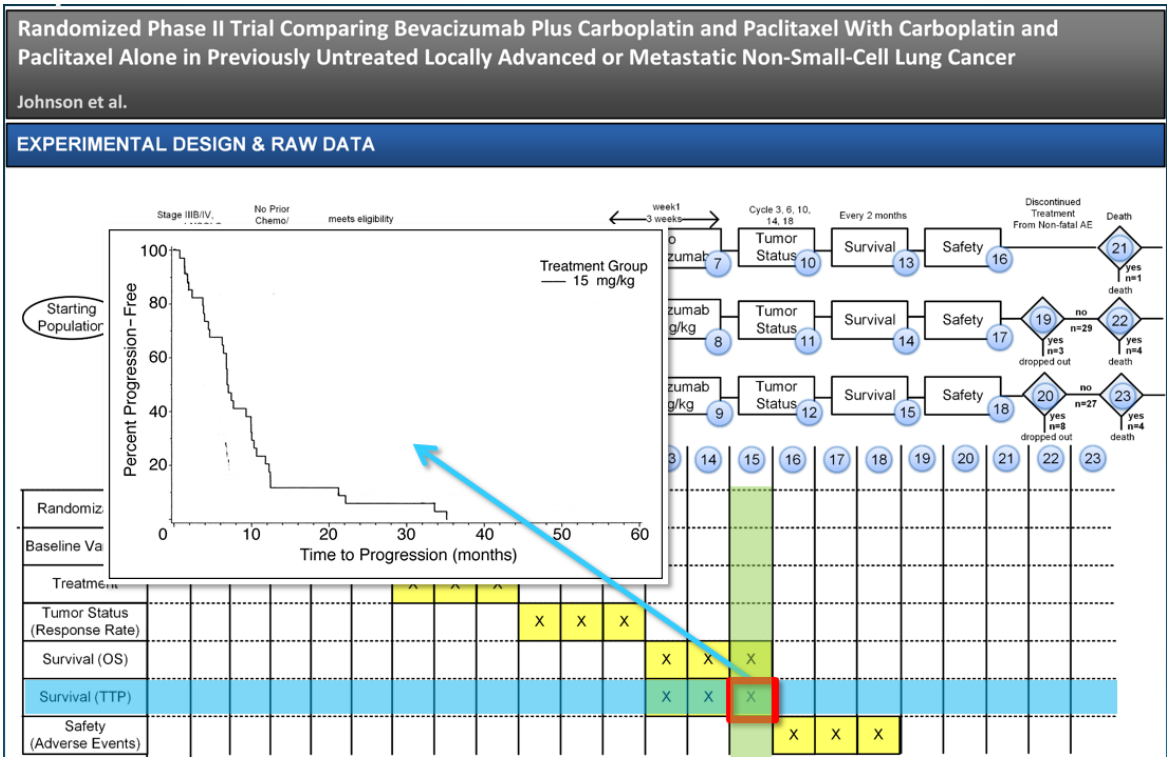


Figure 4-15. Querying for survival data

context. In addition, context for survival probabilities can also include comparisons of treatment arms used for hypothesis testing.

4.4.4 Context Related to Statistical Methods

Statistical testing typically involves comparing distributional parameters associated with comparison arms of the trial. For example, a log rank test could be used to infer a difference in median survival between the treatment arm and control arm. Because the statistical test involves observational data from different nodes in the experimental process model for a particular outcome variable, this uniquely refers to a cell in the data grid of the representation. After the user selects a statistical test, an observational cell and the context for that cell (e.g., population profile, *N* per arm, *etc*.) can be recovered by the backtracking algorithm (see Section 4.4.3). Thus, the representation provides a direct link from the details of statistical methodology to the raw data involved.

.

Chapter 5 - Evaluation

5.1 Description of the representation

To evaluate the dissertation framework, two experiments were performed to assess the following outcome measures: 1) ability to improve speed and accuracy to answer modified CONSORT questions; 2) ability to improve speed and accuracy to answers targeted questions posed by a biostatistician and clinical researcher when the representation is used as a supplemental resource in addition to the status quo paper printout of the trial report.

In Experiment 1 of this evaluation, I evaluate the effectiveness of the representation for clinical trial literature for the purpose of answering standardized and general CONSORT-type questions. I determine: 1) how well the representation can assist users with understanding the published report's content, and 2) whether its presentation is intuitive to navigate and comprehend. Results of the usability study are based on a comparison of interpreting information using the status quo versus using the dissertation representation. In Experiment 2, I investigate the system from these aspects: 1) ability of users to answer commonly asked questions generated by biostatisticians and domain experts; 2) quantitative results from a Likert scale survey on preferences; and 3) qualitative results based on user's comments about the evaluation design and the representation. My results suggest an instrumental role of computer understandable representations to not only reduce manual effort and save time, but also to assist with synthesis of information and knowledge discovery.

5.2 Experiment 1: Alternative Systems Evaluation

The hypothesis tested in this experiment is as follows: given the representation, participants can answer paper-specific query questions with higher accuracy (and faster time) as compared with the status quo. Paper-specific query questions targeted comprehension and information retrieval type questions and were modified from accepted standards.

5.2.1 Study Design

A two-arm randomized trial design (Figure 5-1) was used to compare user task performance using the alternative methods of the status quo (i.e., paper version) versus the dissertation intervention (i.e., visualization). Eleven participants were recruited drawn from graduate students in medical informatics, bioengineering, participants with medical school education, and medical researchers.

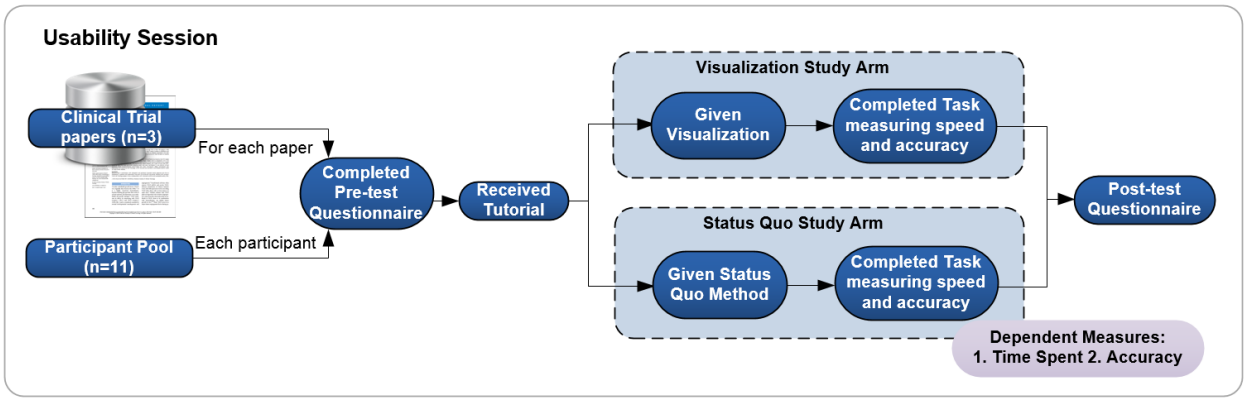


Figure 5-1. Study design consisting of a 2-arm randomized design

5.2.2 Paper Test Cohort

Clinical trials in the domain of non-small cell lung cancer (NSCLC) were chosen to narrow the variability of clinical trials used. A PubMed search was conducted using the keywords "EGFR",

"lung cancer", "non-small cell lung cancer", "clinical trial", and "phase II". The search yielded 261 published reports. For the initial scope of this initial pilot study, three papers were randomly selected that met the criteria of being a clinical trial about NSCLC involving EGFR mutations to assess time spent and accuracy while answering the questionnaire. The paper selection is diagramed in Figure 5-2 and paper characteristics are summarized in Table 5-2 below.

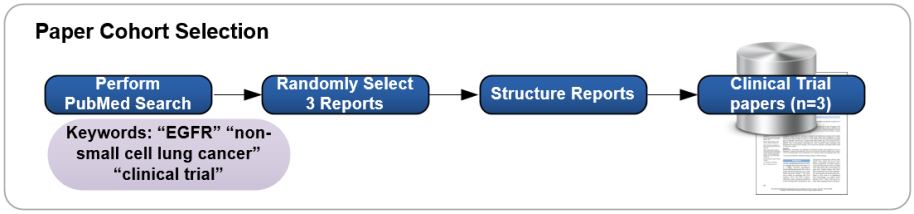


Figure 5-2. Methods for Paper Cohort Selection

Paper	Title of Report	Outcome variable	Sample size	Total Events	Date
1	Randomized Phase II Trial Comparing Bevacizumab Plus Carboplatin and Paclitaxel With Carboplatin and Paclitaxel Alone in Previously Untreated Locally Advanced or Metastatic Non-Small-Cell Lung Cancer[13]	Response Rate (RR)	99	33	2005
2	First-Line Gefitinib in Patients with Advanced Non-Small-Cell Lung Cancer Harboring Somatic EGFR Mutations[14]	Objective Response Rate (ORR) = sum of patients with confirmed complete and partial responses / number of patients treated	98	15	2008
3	EGFR expression as a predictor of survival for first-line chemotherapy plus cetuximab in patients with advanced non-small-cell lung cancer: analysis of data from the phase 3 FLEX study[15]	Overall Survival (OS)	1125	24	2011

Table 5-1. Summary of clinical trial papers used in Experiment 1

5.2.3 Study Execution

Each participant reviewed Clinical Trial Papers 1, 2 and 3 (Table 5-2). For each clinical trial report, participants were randomized into the representation study arm or the status quo study arm. In either study arm, the flow of the study is as follows (Figure 5-1):

(1) Participants filled out paperwork (consent form, pre-test questionnaire) and received a tutorial on how to interpret the representation based on two example questions. Each participant was asked to sign a study participation consent form. A pre-test questionnaire was administered to each participant to characterize their familiarity with cancer, biology, and statistical methods. The pre-questionnaire relates to their level of understanding of cancer, biology, statistical methods, and clinical trial designs (Appendix C).

(2) Participants completed the usability sessions either with the status quo or representation (Appendix A).

(3) Participants answered a post-questionnaire about the visualization of the representation. The post-questionnaire consists of Likert scale survey asking participants to rate the effectiveness of the visualization and the preferences of the user towards the representation and the status quo (Appendix C).

5.2.4 Generation of Test Questions

User tasks for Experiment 1 were divided into two types:

(1) Comprehension task to assess whether the individual is able to synthesize evidence from the published report,

(2) Information retrieval (IR) task to focus on locating specific pieces of evidence in the report.

Comprehension questions were developed based on the CONSORT reporting guideline requirements, and specifically focusing on the test subject's ability to interpret the objective and claims made in the published report. For example, one comprehension question asked: The trial states, 'This large prospective biomarker study found that patients with activating EGFR mutations derive the greatest PFS benefit from erlotinib maintenance therapy.' Describe the method, numerical data, and analyses for this statement." IR tasks focused on the ability of a test subject to locate key information as again adapted from applicable CONSORT requirements. IR questions include reporting the eligibility criteria, locating the experimental arms, summarizing the methodology, and identifying the results of statistical tests.

Questions of both types were presented using multiple choice, fill-in-the-blank, and short answer. All questions were reviewed by a biostatistician who was not involved in the development of the system to reduce bias in word-choice and to ensure conformance to standard guidelines and terminology. The gold standard was created by a domain expert who was given an open amount of time. Tasks were timed and graded for accuracy by determining the percentage of questions answered correctly.

5.2.5 Dependent Measures and Statistical Analyses

The participants used the status quo or the representation to answer questions demonstrating their comprehension of the clinical trial and recorded the time required to answer the questions. The dependent measures of this usability study included time spent, measured in minutes; and accuracy, calculated as the percentage of questions answered correctly. The accuracy was calculated by tallying the number of questions answered corrected, and dividing by the total number of questions.

Overall time spent and accuracy was determined by averaging over all values in each condition. Groups were conditioned on having either the status quo paper or the representation. A pilot study was used to estimate the amount of time and accuracy for each task that was considered reasonable. A power calculation was performed to determine the appropriate sample size for the combination of participants and clinical trials needed. With an estimated time difference of 10 minutes (30 minutes vs. 40 minutes) and standard deviation of 8 minutes, a sample size of 12 per group would yield an 83% power with 5% significance level. With an estimated accuracy difference of 15% (70% vs. 85%) and standard deviation 17%, a sample size of 12 per group would yield an 80% power with 5% significance level. Hence, a sample size of at least 24 is needed, meaning at least 8 participants each reading 3 clinical trial reports. This is satisfied by the number of participants enrolled. A 2-sided student's *t*-test was used to compare accuracy and time spent using the representation versus using the status quo method.

5.2.6 Participants Characteristics

Eleven participants were involved in the study. All participants have read a clinical trial report before and took on average 80 minutes to read it completely. Participants expressed confidence in their understanding of the knowledge presented within clinical trial reports (average 6.0 ± 1.9). While most participants were confident in their understanding of statistical methods (average 6.3 ± 1.5), participants were less confidence their assessment of the quality of statistical tests (average 4.4 ± 1.9). General participant characteristics are presented in Table 5-2. For confidence measures, scale values are 1 = not confident to 10 = very confident. For courses, values indicate number of college-level undergraduate or graduate level courses.

Characteristic	Mean	SD
Confidence with understanding of cancer mechanisms	5.0	1.8
Confidence with knowledge on NSCLC	4.3	1.9
Confidence in understanding knowledge within reports	6.0	1.9
Confidence in understanding statistical methods	6.3	1.5
Confidence in assessing the quality of statistical tests	4.4	1.9
Courses on biology-undergraduate	3.1	2.6
Courses on biology-graduate	3.0	5.6
Courses on statistics-undergraduate	1.2	0.7
Courses on statistics-graduate	2.2	1.1

Table 5-2. Characteristics of participants in Experiment 1.

5.2.7 Results

Overall accuracy was similar between the representation and status quo, however, participants with the representation had on average a quicker overall time than participants with the status quo (representation 26 ± 10 minutes vs. status quo 36 ± 10 minutes; p=0.008) (Table 5-3). This suggests that information is easier to locate in a visualization of the representation than in the status

quo. In an exploratory analysis, the decrease in time taken to answer comprehension questions contributed more to the significant difference than the time taken to answer IR questions (comprehension questions p=0.012 vs IR questions p=0.047) (Table 5-4). Accuracy was maintained in both the representation and status quo despite stratifying by question type. While the representation provided similar accuracy, the tradeoff is a significant times savings when compared to the status quo alone.

System	Accuracy (%)	SD	Time (min)	SD
Representation	73.70%	13.30%	26	10
Status Quo	67.00%	15.90%	36	10
P-value	0.207		0.008	

Table 5-3. Measures of performance as a function of overall accuracy and overall time

Task Type	System	Accuracy (%)	SD	Time (min)	SD
Comprehension	Representation	68.20%	16.20%	18	7
	Status Quo	60.10%	19.70%	24	8
	P-value	0.209		0.012	
Information Search	Representation	80.00%	13.90%	8	4
	Status Quo	75.40%	15.40%	12	5
	P-value	0.462		0.047	

Table 5-4. Measures of performance as a function of overall accuracy and overall time stratified by question type

When stratifying by clinical trial study, non-significant differences were found in both time and accuracy between the representation arm and the status quo arm for each clinical trial study (Table 5-5). The point estimate of Report #2 was shown to have decreased accuracy as compared with Report #1 and #2. The accuracy can be affected due to an increase in complexity of the study

design and greater amount of content for both the representation and status quo method. This trend was explored in Experiment 2. The accuracy for comprehension questions and for IR questions were separated for exploratory analyses (Table 5-6 and Table 5-7). The mean accuracy for comprehension questions within one report suggests a difference between the representation condition and the status quo condition, favoring the representation (83.3%, 69.6%, 51.3% vs. 76.4%, 58.8%, 44.8%). This suggests that using the visualization can increase comprehension. This trend within reports is currently being studied in an attempt to significantly increase accuracy in the visualization of the representation and in the representation itself.

In summary, the results of the usability study were consistent with my intuition. Having the representation required on average 27.8% less time than having the status quo (representation 26 min vs. status quo 36 min; p=0.008) while maintaining similar accuracy. These findings did not appear to be affected by participants' varying levels of familiarity with the statistics, clinical domain (i.e., non-small cell lung cancer) and clinical trial procedures. This suggests that having essential information placed in context of the entire experiment helps users cognitively critique and apply contributions of clinical trials on a deeper level in a timelier fashion. This enables informatics tools to query information to be used for meta-analysis and probabilistic disease modeling and assist with the difficult task of assessing the quality and usefulness of each trial.

While all participants favored the representation over the current method, questionnaires revealed that much work is needed to improve the satisfaction and usability of the representation. One solution to avoid bias of a less completely documented clinical trial study is to use the representation to supplement an individual's understanding gained from reading the status quo published report (see Experiment 2). While the study design proposed in Experiment 1 assigns

participants to either the status quo or the representation condition, in actuality, the two conditions are not mutually exclusive. This suggests that the combination of having the representation to reference while reading the status quo published report can further help to save time and increase accuracy. In an unstructured interview with potential users, one biostatistics professor anecdotally

Trial	Representation				Status Quo			
	Accuracy (%)	SD	Time (min)	SD	Accuracy (%)	SD	Time (min)	SD
1	77.3%	5.34%	27.3	10.73	68.9%	10.01%	34.5	10.63
2	58.5%	11.67%	25.0	8.86	53.8%	14.80%	34.3	9.63
3	84.5%	6.10%	24.6	12.42	78.8%	10.23%	38.2	11.20

Table 5-5. Measures of performance as a function of overall accuracy and overall time stratified by trial number

Trial	Representation				Status Quo			
	Accuracy (%)	SD	Time (min)	SD	Accuracy (%)	SD	Time (min)	SD
1	69.6%	8.7%	19.3	6.4	58.8%	11.8%	23.0	4.2
2	51.3%	10.3%	17.8	6.5	44.8%	19.5%	23.5	8.8
3	83.3%	11.8%	15.2	8.1	76.4%	12.3%	26.5	10.2
Combined	68.2%	16.2%	17.6	6.7	60.1%	19.7%	24.4	8.0

Table 5-6. Measures of performance as a function of accuracy and time for comprehension questions stratified by trial number

Trial	Representation				Status Quo			
	Accuracy (%)	SD	Time (min)	SD	Accuracy (%)	SD	Time (min)	SD
1	85.4%	10.2%	8.0	4.4	76.6%	14.8%	12.3	6.3
2	70.0%	18.7%	7.2	4.3	68.3%	9.8%	10.8	5.8
3	86.0%	5.5%	9.4	4.6	81.7%	19.4%	11.7	3.2
Combined	80.8%	13.9%	8.2	4.2	75.4%	15.4%	11.5	4.8

Table 5-7. Measures of performance as a function of accuracy and time for IR questions stratified by trial number

noted that she liked the hybrid process model-spreadsheet for contextualizing observations and statistics.

5.3 Experiment 2: Added Value Evaluation

The hypothesis tested in this experiment is as follows: given the representation and status quo, participants can answer paper-specific query questions with higher accuracy (and faster time) compared with the status quo, alone. Query questions are generated by biostatisticians and clinicians.

5.3.1 Study Design

A two-arm, modified cross-over randomized design was used to test the "value-added" effect of the intervention (Figure 5-3). Thus, a one-sided hypothesis was tested. The intuition is that in practice (i.e., in a mature real-world implementation) the system would be used as follows: Users would read a research paper; however, questions would come up as needed sometime later during clinical practice (for the EBM clinician) or during a research endeavor (for the disease modeler). In this case, the user would revisit the previously read paper to search for a specific answer to a question. Experiment 2 was designed to test whether answering such questions upon re-visiting the paper is more accurate and timely using the status quo (i.e., the paper copy) as compared to the dissertation work intervention.

Twelve research participants were recruited from a graduate level bioengineering class at UCLA (Bioengineering 226 – Medical Knowledge Representation). This population of students served as proxy subjects for the ultimate users of the system, which is envisioned to be biostatisticians,

clinical practitioners, and clinical researchers. Three papers from the pool of 21 test papers were randomly assigned to each participant, and stratified with each participant receiving one paper from each complexity level (i.e., one low complexity, one medium complexity, and one high complexity paper).

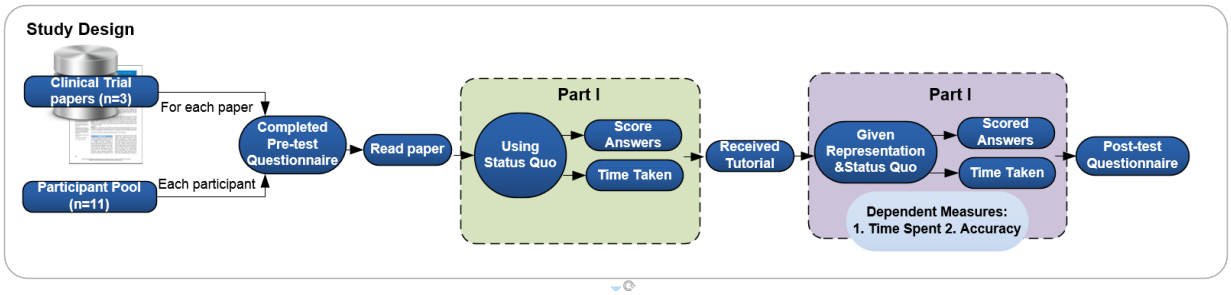


Figure 5-3. Study design consisting of a 2-arm randomized modified cross-over design

5.3.2 Paper Test Cohort

Clinical trials in the domain of non-small cell lung cancer (NSCLC) were chosen following the same procedure as Section 5.2.2.

It was noted that the "comprehensibility" of research papers can vary widely. Comprehensibility entails aspects related to language, organization, level of detail, and experiment complexity which can affect a reader's ability to recall details of a study for question answering. Thus, from the pool of retrieve papers, a study coordinator (MT) randomly selected a paper and assessed its level of difficulty, roughly categorizing a sampled paper as either: a) low complexity; b) medium complexity; or c) high complexity. The level of complexity was assigned based on a number of factors including the number information elements in tables and figures, the page length of the report, and the time taken to read the paper as determined by domain expert annotators. The intent

was to have a stratified sample of papers with equal amounts in each category (Figure 5-4). This stratification was performed to test the intuition that the system would be most beneficial for papers that were deemed "difficult" with high complexity. Within each participant, even distributions were maintained with respect to level of complexity. In other words, each participant received one paper from low complexity papers, one of medium complexity and one of high complexity. Due to time constraints, seven papers from each category were identified for a total of 21 unique papers to be used for testing.

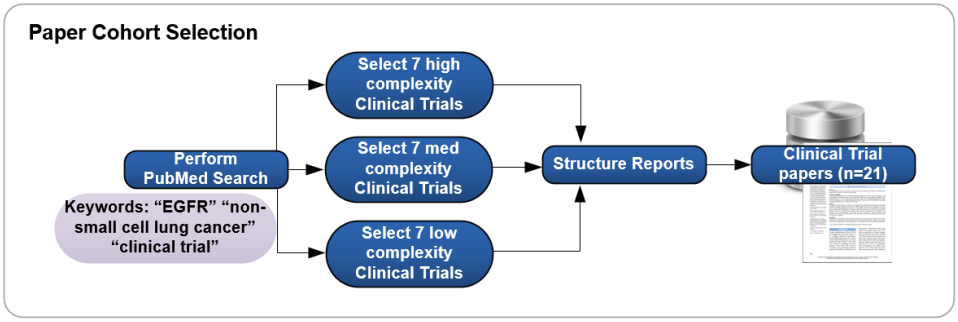


Figure 5-4. Categorizing sampled trial reports according to level of complexity

5.3.3 Study Execution

The flow of the study is summarized as follows (Figure 5-3):

(1) Each participant filled out paperwork including a study consent form and a pre-test questionnaire. The questionnaire is shown in Appendix C and relates to their level of understanding of cancer, biology, statistical methods and clinical trial designs.

(2) Each participant received a 2-hour tutorial on how to interpret the questions created by the experts (see Section 5.4.2). The tutorial reviewed an example paper and example questions with expert provided answers.

(3) Participants were given the status quo paper copy of their assigned clinical trial papers. Each participate was instructed to read all 3 assigned papers at their leisure but within 24 hours prior to their scheduled usability sessions. The participants were asked to read these papers as if they were normally inquiring about a particular line of research. No restrictions related to note-taking, highlighting, etc. were imposed.

(4) Participants completed Part I of the usability session, which involved using the previously distributed paper copy to answer the questions generated by the biostatistician and domain experts for their given papers. There answers were recorded on a standard form. Time to complete each question was self-reported. See Appendix B for samples of the forms used and details sample questions for a given paper.

(5) A washout period was imposed (at least one week). The assumption is that during the washout period, users would forget most of the questions and their answers provided in the status quo arm.

(6) After the washout period, each subject, for each assigned paper (same set as in Part I), were subsequently placed in the intervention arm. After a second tutorial on how to interpret the representation based on an example representation, lasting 1 hour; participants completed Part

II of the usability session with the representation for all three papers. Answers were recorded on a standard form. Time to complete each question was self-reported.

(7) Participants answered a post-questionnaire to gather impressions on the adequacy of its content and to provide feedback on design, interface, and suggestions for additional functionalities. The post-questionnaire is shown in Appendix C and included responses related to the effectiveness of the representation for characterizing various aspects of a trial study (e.g., purpose, interventions, study design, observational data, and statistical methods).

(8) Finally, 8 out of the 12 participants were interviewed by a study coordinator to gather feedback for general preferences, concerns, and thoughts about the study.

5.3.4 Generation of Test Questions

The query question set and the gold standard answer are divided into two categories: (1) Clinical and (2) Biostatistics. Two domain experts in the clinical setting created the query questions for the clinical category and one biostatistician in the research setting created the query questions for the biostatistics category. Domain experts were given an open amount of time to create the query questions and gold standard for each of the 21 test papers (Figure 5-5). To address biases with question wording, the two domain experts worked together to eliminate ambiguities and differences in response. A second biostatistician was asked to proofread the questions generated by the main biostatistician.

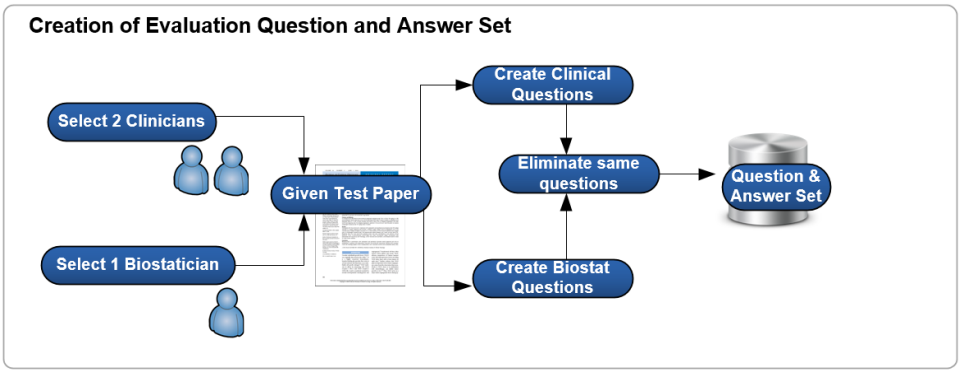


Figure 5-5. Process to create clinical and biostatistical test questions

A sample of clinical questions are as follows:

1. What is/are the study's objective(s)? What is the clinical reasoning as to why this study was created? By the end of the paper, does the paper answer this/these objectives?
2. What is the medication name, dosing strength, and frequency? Does the treatment regimen account for dose interruption or reduction during the study?
3. Describe the target population (i.e., total sample population) and the control/comparator group.
4. List all Grade 3 and above adverse events (or side effects) for the control and/or comparator group(s) and the experimental group?
5. What are the causes of death (if any) in the control/comparator group(s) and experimental arms? Are the causes of death the same?
6. What is the proportion and number of patients that dropped out in the control/comparator group(s) and experimental arms? Are the proportions the same?
7. List the outcome measures that help determine the intervention's clinical relevance (i.e. quality of life markers, survival metrics, etc.)?

8. If this is a survival study, how many more months/days does the intervention prolong life?

A sample of biostatistical questions are as follows:

1. What is the proportion of patients with tumor stage IIIB in each group (by gene expression, treatment) and overall?
2. The following questions relate to study design: What is the objective of the study? How many experimental arms are there? What is the phase of the clinical trial? What types of analyses are performed?
3. What is the median survival or progression free survival (PFS) and hazard ratio in each group (of gene expression, treatment group and overall?
4. What is the response rate of treatment groups? List the time points for all response rates of treatment groups given.
5. What is the top 3 adverse events in the treatment groups? For each adverse event, how many patients experienced that adverse event?
6. What is the context for the most significant statistical result (i.e., p-value)? Describe (a) the age, demographic of population, (b) interventions details, of each population.
7. Describe how the most significant statistical test was calculated? Describe (a) variables, (b) test statistics used, (c) sample size.

See Appendix B for a more comprehensive sample of actual questions posed to the test subjects.

5.3.5 Dependent Measures and Statistical Analyses

During usability sessions part I and part II, participants were instructed to answer the query questions for clinical trial reports demonstrating their comprehension of the clinical trial study using the status quo without the representation for Part I and with the representation for Part II, and to record the time required to answer each question. The dependent variables were: (1) self-reported completion time, and (2) graded score as determined by a domain expert. Answers to query questions for Part I and Part II were collected as free-text responses and graded for correctness on a scale from 1-3, where 1 is incorrect, 2 is partially incorrect, and 3 is correct.

Grading scores were assigned by two domain experts and one biostatistician, in one of two methods. For clinical questions, the two domain experts each graded the participant answers for Part I and Part II while being blinded from each other. Discrepancies were resolved jointly. For biostatistics questions, an answer key was generated by a domain expert. An experienced grader obtained the answer key and graded all participant answers.

Following the usability sessions, a post-questionnaire was prepared to assess the affinity and usefulness of the representation, to gather impressions on the adequacy of its contents, and to provide feedback on design, interface, and additional functionalities (Appendix C). Finally, I conducted a 30-minute long semi-structured interview with a group of participants. Throughout the interview process, I documented meeting notes and partially transcribed these notes. Following interviews, despite possible misunderstandings that arise in the initial survey answering, no participants were allowed to re-submit survey scores.

The participants used the status quo (Part I) first, then the status quo with the representation (part II) afterwards to answer questions demonstrating their comprehension of the clinical trial. Time required to answer the questions was self-reported. Summary statistics for time spent and accuracy were calculated for Part I and Part II overall, and stratified for each complexity level. A one-sided student's *t*-test was used to compare accuracy and time spent using the status quo vs. using the status quo method with the representation. In addition, summary statistics were calculated for participant characteristics, and the Likert scale survey.

5.3.6 Participant Characteristics

Twelve participants were involved in the study. Participants ranged in experience from one to five years. 50% of the participants (6 out of 12) had read a clinical trial report before and took on average 27 ± 8.4 minutes to read it completely. General participant characteristics are presented in Table 5-8. For confidence measures, scale values are 1 = not confident to 10 = very confident. For courses, values indicate number of college-level undergraduate or graduate level courses.

Characteristic	Mean	SD
Confidence with understanding of cancer mechanisms	4.0	2.4
Confidence with knowledge on NSCLC	3.3	2.3
Confidence in understanding knowledge within reports	6.6	2.2
Confidence in understanding statistical methods	6.7	2.5
Confidence in assessing the quality of statistical tests	6.3	2.6
Courses on biology	6.1	5.9
Courses on statistics	4.8	4.6

Table 5-8. Characteristics of participants in Experiment 2.

5.3.7 Results

Similar to Experiment 1, in Experiment 2, non-significant differences were again found in accuracy between the representation condition and the status quo condition for each clinical trial study (Table 5-9). When stratifying by complexity level, it was observed that the point estimate of the medium complexity report had decreased accuracy as compared with reports of low and high complexity. In Part I, the mean accuracy for low complexity reports is higher than medium and high complexity reports (low 94.4% and 100%, vs medium 75.0% and 66.7% and high 80.0% and 80.0%) (Table 5-10). Level of complexity within reports suggests a difference between the representation and the status quo, favoring the representation. In particular, the representation of the medium level of difficulty papers seemed to bring about the largest benefit.

The accuracy can be affected due to an increase in complexity of the study design and greater amount of content for both the representation and status quo method. Future evaluations can investigate this trend in an attempt to significantly increase the results for accuracy in the representation over the status quo. In summary, given the representation and status quo, participants can answer query questions with faster time and similar accuracy as compared with the status quo, alone. My results suggest an instrumental role of representations in assisting biostatisticians and clinicians in their assessment of quality and evidence-based medicine.

System	Accuracy (%)	SD	Time (min)	SD
Paper Only	69.5%	29.6%	54	32
Paper + Representation	75.5%	21.3%	34	17
P-value	0.15		0.000003	

Table 5-9. Measures of performance as a function of overall accuracy and overall time

Complexity	System	Accuracy (%)	SD	Time (min)	SD
Low	Paper Only	70.9%	47.7%	40	17
	Paper + Representation	100.0%	0.0%	31	16
	p-value	0.211		0.016	
Medium	Paper Only	75.0%	15.4%	61	29
	Paper + Representation	68.8%	13.9%	37	16
	p-value	0.252		0.001	
High	Paper Only	62.4%	33.3%	62	43
	Paper + Representation	72.2%	27.2%	36	20
	p-value	0.5		0.005	

Table 5-10. Measures of performance as a function of overall accuracy and overall time stratified by complexity level

5.4 User Preferences

The previous section discussed the utility of the system to answer task-related queries, and this section discusses participants' preferences and usefulness of the representation to the participant. My evaluation results are divided into two parts: (1) the results from a Likert scale survey about the usability and satisfaction, preferences, and likelihood of using the representation again; and (2) the open free-text comments about the study design organized by themes.

5.4.1 Questionnaire Results

75% participants (9 out of 12) preferred the representation to the status quo. Participants rated the usefulness of the representation with an average of 7.0 ± 1.5 (where 10 is completely essential, 5 is neutral and 1 is useless), and the satisfaction of the visualization of the representation at the current state with an average of 7.0 ± 1.5 (where 10 is completely satisfied, 5 is neutral, and 1 is unsatisfied). The likelihood of participants using the representation again is reported on average as 7.6 ± 2.1 (with 10 being will use the representation again, 5 is neutral, 1 being will not use the

presentation again). Among the participants that preferred the representation over the status quo, participants rated the usefulness of the representation with a median of 8, and the satisfaction of the visualization at the current state with a 7. The likelihood of participants using the representation again is reported with a median of 5. Among the participants that preferred the status quo over the representation, participants rated the usefulness of the representation with a median of 8, and the satisfaction of the visualization at the current state with a 5. The likelihood of participants using the representation again is reported with a median of 5.

5.4.2 Free Comments

In general, participants appreciated the representation as a way to provide an overview for complicated information, including viewing participant flow, and quickly identifying data points and statistical methods. All participants agreed that the representation contained advantages, such as increasing speed in retrieving information. When information is immediately clear, it was fast to answer the task. Selected free-text comments are reviewed below by theme:

Washout period: Two participants provided comments regarding their recollection of the paper during Part II of the evaluation. Both participants recalled specific details about the question task when using on the status quo method, and preferred using their method of arriving at the answers they recalled with the status quo method. If information was not presented in the representation, participants sometimes remembered that it did exist in the paper. Participants stated:

> *"I was able to recall how I answered the multiple choice questions, like 'Was statistical significance achieved?'"*

"Anything I struggled to identify the first time, I recalled [using my notes from when I read the paper]."

"Some of the more ambiguous questions (e.g., future directions) or questions for which the answers were not listed in the paper, I arrived at more quickly because I remembered how I had previously resolved these ambiguities [using the paper]."

While remembering specific facts may be considered a disadvantage with respect to a washout period; overall, this may be considered advantageous with respect to recalling information within previously read papers. Another participant stated:

"The representation confirmed that I was right [the first time I answered the question]."

Poor and inaccurate instantiations*:* Two participants addressed issues with poor and inaccurate annotated-populated instantiations. If the answer was not adequately answered using the representation, participants were skeptical of whether or not it was absent in the representation because it was missed by the annotator but present in the paper, or not present in the paper at all.

Learning curves from expert and novice readers: Despite expertise in interpreting clinical trials, there was a learning curve in comprehending the representation. For one participant who self-rated him/herself as very familiar with reading clinical trials, the addition of a new system had an unnecessary learning curve, when he/she was already fluent in reading clinical trials in the status quo form. Another participant who was not as familiar with reading clinical trials, found information was not presented clearly in the representation, without realizing that it is usually ambiguous in the paper report. For example, he/she was confused why participant flow information was not presented more clearly in the paper, and was quick to point out sample size

numbers did not add up in the representation. Another point of confusion for the participant unfamiliar with reading clinical trials was the lack of consistency of presenting concrete time points for observed events.

Chapter 6 - Summary

This chapter summarizes the results of this study (section 6.1), compares study results to other work in the field (section 6.2), identifies limitations with this work (section 6.3), and discusses future directions (section 6.4)

6.1 Summary of the Dissertation

This dissertation describes a representation that models clinical trial summaries within the context of experimental design steps. The approach introduced a novel hybrid representation that utilizes the process model and data grid, in an effort to describe the collection and/or constraints of a data variable. Once the representation was created, it was implemented into prototype applications to answer queries drawn from users who are interested in evidence-based medicine. Specifically, the representation is intended to support queries related to understanding the context of reported observations and analysis methods based on the details of the experimental design. The novel representation could then be visualized in a consistent manner across a diverse sample of clinical trial reports. A standard representation would lead to a familiarity in navigating and querying important details related to understanding statistical significance of scientific discoveries.

The specific contributions of this dissertation are summarized as follows:

- A representation with the ability to express detailed context for reported observations (e.g., quantitative descriptions). A backtracking algorithm transverses the nodes in the process model following the semantics of the linkages. Each node in the pathway from the first

node to the node of interest provides a different part of the context. The query results in a compilation of context aggregated from each node, which is returned to the user.

- A representation that is generalizable beyond the studied domain of non-small cell lung cancer (NSCLC) trial reports. The process model can be built to accommodate any level of detail, and the data grid is adaptable and assembled from familiar ontologies for a given application. The representation developed in this dissertation is based off a situational ontology for NSCLC, however, the methods provided can be used to generate a situational ontology for a different disease domain. The rationale is that clinical trials within specific trial designs follow similar steps, regardless of disease domain.
- A representation that is intuitive and easy-to-understand. The usability of the representation and its impact on time-savings was demonstrated via the evaluation in Experiment 1 and 2. The results of Experiment 1 showed that users with the representation maintained similar accuracy, and required on average less time when answering CONSORT-like questions than users with the status quo method. The results of Experiment 2 confirmed the results of Experiment 1 for a set of typical query questions. It showed that users with the representation and status quo method answered query questions faster and with similar accuracy as compared with the status quo alone.

6.2 Contributions to the Field

The need for formalizing information contained within clinical trials research papers has been previously recognized and is motivated by a number of driving applications: 1) the need for editors, peer reviewers, and readers to understand how the trial was performed and to judge whether the findings are likely to be reliable; 2) the need for decision support for evidence-based medicine; 3)

the need to create comprehensive disease models; and 4) the need for more sophisticated (accurate) retrieval systems. The specification for defining a good representation is evolving from many complementary efforts. This work accompanies a number of existing efforts to characterize clinical trial studies including the following major efforts.

ClinicalTrials.gov – the ClinicalTrials.gov registry includes a large breadth of studies, containing over 100,000 records and meta-tags for describing clinical trials studies [Zarin 2005 and Laine 2007]. The database is motivated by issues related to patient recruitment, and thus, includes meta-data related to the trials purpose, intervention, recruitment criteria, research arms, primary outcome measures, locations and contacts. Similar to this dissertation, one application is a web based interface to identify a particular clinical trial study. However, metadata tags are less descriptive than the system developed in this dissertation and the application focus is mainly directed towards patient recruitment and/or matching patient cases to trial studies. The quality of the study is not characterized in the clinicalTrails.gov effort and quantitative information is not present, or in an unusable form.

CDISC – The Clinical Data Interchange Standards Consortium is a nonprofit organization committed to the development of industry standards to support the electronic exchange of clinical trials data and metadata [Kush 2012]. The organization provides one overarching standard model for the data interchange of healthcare information and clinical trial/research data at the individual patient level. The standard is motivated by the need to integrate data sets from different institutions and to get improved estimates of the probabilities of expanded state spaces. This dissertation can aid this effort by improving the richness of the representation used to characterize patient data participating in a clinical trial. In particular, the representation allows patient observations to be

completely specified for each stage of a clinical trial, including how/why patients were recruited, the arm in which they participated, the exact specification of the intervention (e.g., drug administration details) and how observations were made. It has been commented by Professor Hyun J. Grace Kim that the representation studied in this dissertation could be a powerful rich approach to gather and collect measurements during the data collection stage of a clinical trial.

CONSORT - the CONSORT (Consolidated Standard for Reporting Trials) statement, discussed in Chapter 2, defines a set of guidelines and suggestions to aid RCT authors in deciding what to report [Hopewell 2008, Moher 2010, Altman 2001]. It is motivated by issues related to improving the critical appraisal and interpretation of RCT reports. It includes a flow diagram and a 21-point checklist of required items necessary to inform the reader about what the researchers did during the trial and what they learned from it—their methods, results, and analysis [www.consort-statement.org]. CONSORT has received powerful backing from journal editors including JAMA, Annals of Internal Medicine, the British Medical Journal and at least 70 other leading journals. The efforts of this dissertation could provide extensions to the CONSORT specification including the integration of the hybrid process model into the checklist item for experimental design procedure.

Global Trial Bank Project / Human Studies Database Project – The Global Trial Bank (GTB) is a nonprofit organization formed under the auspices of the American Medical Informatics Association whose goal is to speed the dissemination, understanding, synthesis, and translation of clinical trials to improve healthcare for humans [Sim 2007, Sim 2010]. The project has attempted to further refine the representation of information specified generally in the CONSORT statement. A comprehensive schema (ontology) of RCT concepts has been defined which standardizes the

representation such that improved computations can be performed (querying, deduction, inferencing, etc.). In this dissertation, the characterization and contextualization of quantitative information could add to the functionality of this effort, especially in regard to assessing the quality and transportability of a study between populations.

NeuroScholar / Research Maps – NeuroScholar [Khan 2006, Burn 2006], an open source software platform, provides a way to extract knowledge from various sources (i.e., images, lab notes) and create links (associative, causal, *etc*.) between the pieces of knowledge to show how the extracted knowledge fragments relate to one another. The goal is to synthesize the experimental and observational evidence for a given disease target of investigation. The resulting text fragments, or "knowledge statements," are then saved and synthesized in order to obtain a holistic view of the domain. Similarly, Research Maps, discussed in Chapter 2, attempts to synthesize causal statements discussed in the scientific literature [Silva 2015]. While both NeuroScholar and Research Maps provide a synthesized summary for fragments of information, however, they do not characterize the strength of associations between causal hypotheses. This dissertation can complement these efforts by providing details of the statistical methods used in a study and clarify exactly the context for reported observational frequencies.

These efforts attempt to improve the sharing of data and knowledge, to enforce consistent information coverage required to assess and interpret RCT studies, and to improve the documentation of how various pieces of data are related. These and other efforts are important and complementary efforts toward the goal of improving the utility of information currently stored in free-text research papers.

6.3 Limitations of this Dissertation

This is an exploratory dissertation on establishing the specifications for a representation on clinical trial studies reported in the scientific literature, specifically to support details related to the context for observational data and statistical calculations. There were a number of limitations however in the study that require further consideration before large scale application of the methods can be executed. These limitations are summarized as follows:

Situational Ontology Completeness: The system requires a comprehensive situational ontology for the application domain. In this dissertation, many elements of the ontology were borrowed from existing knowledge sources. However, a large number of entries related to drugs, properties, property states, intervention methods, etc. had to be manually included in the ontology. Additionally, organizing the concepts into a logical semantic model for the domain is challenging, requiring the development of definitions for semantic frames and relations between frames, in general. Future directions for this research could employ knowledge acquisition methods based on natural language processing to expedite identification of unique concepts within a large corpus of research papers from a selected domain. As various aspects of the ontology are generic to all domains (e.g., study design and statistical methods), it is likely that ontology development will be incrementally more scalable as a greater number of domains are covered.

Instantiation Tools: In this dissertation, the annotation tools to instantiate the representation for a given PDF research paper were not being tested or evaluated for large-scale deployment. While instantiating a single research paper was time and energy intensive and could benefit from a well-developed tool, it was not the main purpose of this dissertation to develop tools. Future directions

could employ a number of improvements in this area including: 1) the development of text classifiers to localize the text within a clinical trial report to specific aspects of the representation (e.g., hypothesis, study design, intervention details, analysis methods, *etc*.); 2) natural language processing methods to automatically annotate quantitative and other observational details; and 3) a helper program interface along with a systematic line of questioning to progress the annotator user through each and every aspect of the representation without knowledge about the underlying semantics, storage details, and underlying structure. Inventories of common process models could be provided in the annotation program to serve as a starting point for instantiating the process model for a given paper.

User Interface: In this dissertation, the user interface to the representation could be significantly improved through providing custom views and templates of different ontological class objects and their values. To avoid visually disorganized interface, a hierarchical tree could be employed to collapse attributes for larger frame objects (e.g., Demographic Class of properties). An example of a custom template can include a worksheet to drag-and-drop data for immediate appraisal. For example, graphing of multiple survival curves from different arms of the study should be provided as a standard function to facilitate comparison of outcomes. Another limitation of this current work is the ability to view only one clinical trial report at a time. Operations for comparison between studies cannot be performed. An effort to develop an integrated visualization for multiple clinical trial reports is underway. With the integration of data, issues arise pertaining to appropriate ways for dealing with conflicting data and assigning relative weights to data.

Generalizability: While this dissertation concentrated mainly on randomized clinical trial studies from lung cancer studies, an informal evaluation on generalizability was conducted on students

from BE 226 (academic years 2013 and 2014). Students instantiated representations for their own chosen domain outside lung cancer studies. Most students stayed in the domain of cancer, ranging to brain cancer to skin cancer. In general, efforts were quite successful in characterizing their own selected research paper. However, the representation needs to be formally tested on a much larger and a more diverse sample of research papers, including the many variations in research designs and scientific areas of investigation.

Several limitations relate to Experiment 1 and 2 of the evaluation methods. Figure 6-1 labels the biases at various parts of the study design.

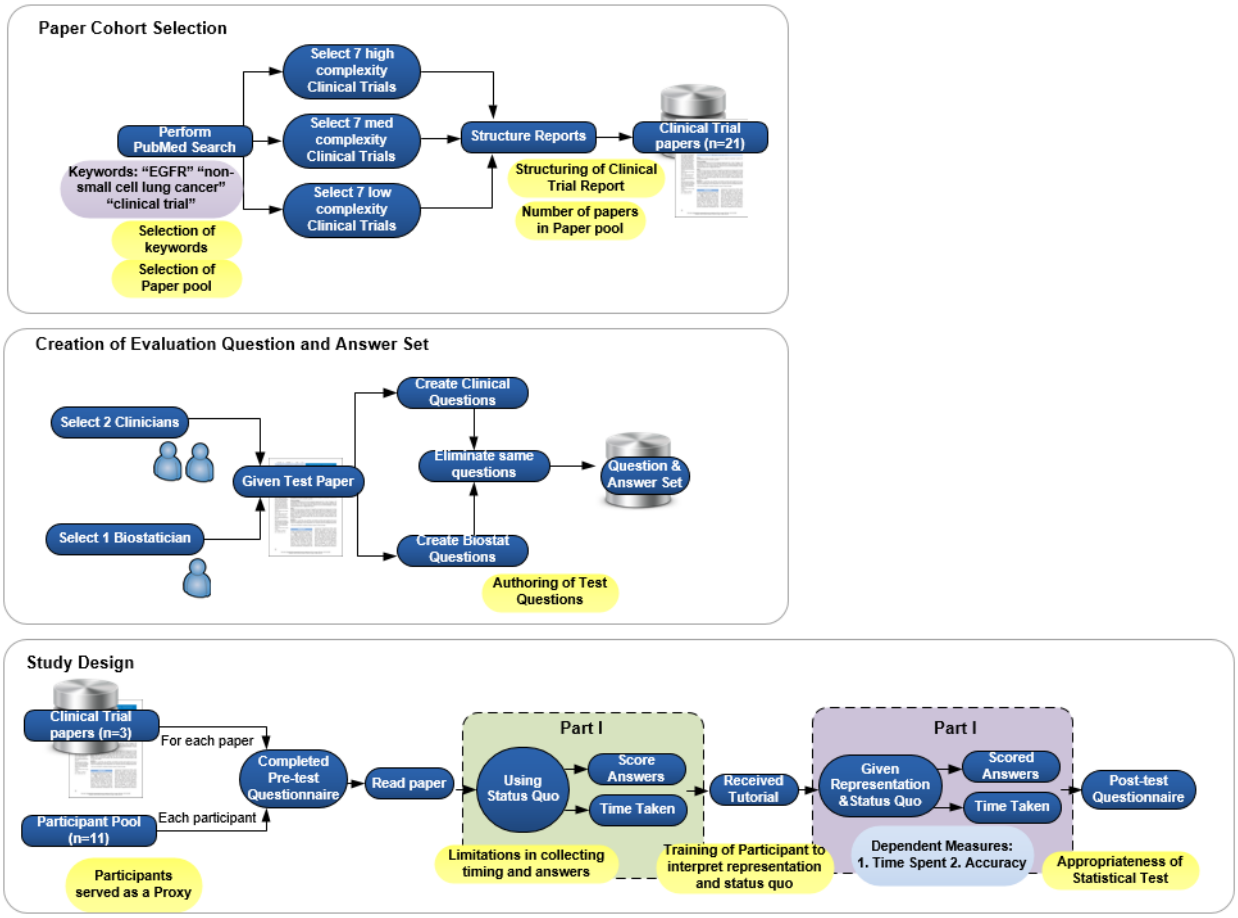


Figure 6-1. Biases associated with Experiment 2 of the evaluation

Evaluation Subject Pool: Test subjects were drawn mainly from student pools in either medical informatics or bioengineering. Students were assumed to be a proxy for clinical investigators and/or evidence-based medicine practitioners.

Bias in Cross-over specific to Experiment 2 design: A modified-crossover design was used for Experiment 2 that involves the collection of data from a sample at two time points. The purpose of the modified crossover design was to document the changes in the dependent variable due to the addition of an intervention, and not the changes over time. I assumed that baseline in the sample population just prior to assessment at both time points was equivalent. However, there is a chance that the washout period was not adequate and the results exhibited a carryover effect, or recollection of the task. In the case of a carryover effect, participants can perform better simply by repeating the task a second time. The carryover effect was assessed anecdotally when participants were surveyed afterwards about the extent of what they remembered during Part II. Most participants answered saying they did not remember much from Part I of the modified crossover design. Within participants that remembered, the trend showed participants recalled answers to the questions they struggled with and devoted a large amount of time to. While crossover designs provide a way to control for confounding factors by providing a more efficient comparison of treatments, a crossover design contains inherent design flaws related to whether improvement in performance can be attributed to the intervention or recollection of the tasks.

Authoring and Grading of Test Questions: Another limitation stems from the design of the task questions for Experiments 1 and 2; and the grading of answers. The goal of the task questionnaires and common query questions was to accurately measure comprehension in a sample population. Because no standard list of questions exists to test comprehension of clinical trials, questions were

modified from standard reporting guidelines to determine the types of information necessary for comprehension. To protect from further bias during modifications, the final list of questions was confirmed by domain experts to determine if answering questions display understanding for Experiment 1. For Experiment 2, common clinical queries were generated individually and answered by each domain expert, and common biostatistical queries were double-checked by an outside biostatistician. In addition to limitations in generating task questionnaires, there is no standard metric to grade the answers from participants, and responses can vary greatly. For Experiment 2, participant responses were manually coded by a grader, and similar responses were grouped together into several categories. To address the lack of a standard metric to grade answers, a rubric was generated for each category and answers was randomly checked by at least 2 graders to assure a level of agreement between the codes given to a response. Extra precautions were taken to ensure that questions were designed in a systematic way and answers were reproducible and valid.

6.4 Future Direction

There are several possible areas of expansion for this dissertation work.

Other types of clinical studies: Given the many varieties of clinical trial designs possible, the representation should be revised to extend its applicability to include other types of investigations, such as natural observational studies. I believe that any type of observation that can be linked to a process model can be adapted to the representation. Incorporating new studies would entail additional entries in the situation ontology, and methods described in this dissertation may be used to develop a supplemental ontology for each different trial design.

Functionality: There are several comments noted from expert users (i.e., biostatisticians) regarding how the representation could be ideal for facilitating comparison of trials. This might be particularly relevant for various meta-analysis efforts. One common task in meta-analyses is to assess the bias within a study. To support queries related to how known confounders were addressed (e.g., controlled for) within a particular research study, the representation can provide links from the ontology to causal models to reveal specific variables to further investigate. Another task includes an assessment of similarity of selected trials. The representation can be used to display and compare the global variable list, and similarities between variables can be gauged. Queries can be built on the representation to assist with and automate the process.

Application Areas: Recommendations from various individuals introduced to the representation have suggested experimenting with the representation to present journal club research articles, for teaching experimental study design and analysis, and for providing evidence for a management approach during clinical tumor boards.

6.5 Concluding Remarks

Numerical data is the key to assessing the contributions of the clinical trial. However, these contributions are locked within published reports that are unstructured and often require extensive manual review to gain a deeper understanding of the study itself. A significant amount of effort is needed to identify and organize information scattered throughout published reports, requiring clinicians and researchers to organize this information mentally. A representation is necessary to help summarize essential elements and connect relevant elements together.

The contribution of this dissertation is a representation that characterizes and places numerical data in precise context of how it was generated. This study demonstrated that the representation is intuitive, and provides significant time savings when answering common queries asked by clinician and biostaticians. While an immediate goal is quality assessment, the eventual goal is to create a disease model for inferring diagnosis or the best therapeutic strategies, and/or predicting prognosis. These disease models require a sufficiently rich bridge representation to unambiguously extract the information from clinical trial studies. The representation can be considered a step towards creating a unifying bridge representation.

Chapter 7 - References

1. Adams-Campbell LL, Ahaghotu C, Gaskins M, Dawkins FW, Smoot D, Polk OD, Gooding R, DeWitty RL. Enrollment of African Americans onto clinical treatment trials: study design barriers. J Clin Oncol 2004;22:730–734.
2. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, and Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134(8): 663-94.
3. Ambrose PG, Hammel JP, Bhavnani SM, Rubino CM, Ellis-Grosse EJ, Drusano GL. Frequentist and Bayesian pharmacometric-based approaches to facilitate critically needed new antibiotic development: overcoming lies, damn lies, and statistics. Antimicrob Agents Chemother. 2012 Mar;56(3):1466-70.
4. Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ. 2013 Jan 21;346:e8668.
5. Berger JO and Sellke T. Testing a point null hypothesis: The irreconcilability of P-values and evidence. Journal of the American Statistical Association. 1978;82:112-139.
6. Berry D, Wathen JK, Newell M. Bayesian model averaging in meta-analysis: vitamin E supplementation and mortality. Clin Trials. 2009 Feb;6(1):28-41.
7. Blake C. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. J Biomed Inform. 2010 Apr;43(2):173-89.
8. Bodenheimer T. Primary care—will it survive? N Engl J Med. 2006;355(9):861–864.
9. Bogdan-Lovis E, Fleck L, Barry HC. It's NOT FAIR! Or is it? The promise and the tyranny of evidence-based performance assessment. Theor Med Bioeth. 2012 Aug;33(4):293-311.
10. Brank J, Groberlnik M, Mladenic D. A survey of ontology evaluation techniques. SIKDD 2005 at multiconference IS 2005, 17 Oct 2005, Ljubljana, Slovenia.
11. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Soldatova LN, Stoeckert CJ Jr, Turner JA, Zheng J; OBI consortium. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010 Jun 22;1 Suppl 1:S7.
12. Brugger W, Triller N, Blasinska-Morawiec M, Curescu S, Sakalauskas R, Manikhas GM, Mazieres J, Whittom R, Ward C, Mayne K, Trunzer K, Cappuzzo F. Prospective molecular marker analyses of EGFR and KRAS from a randomized, placebo-controlled study of erlotinib maintenance therapy in advanced non-small-cell lung cancer. J Clin Oncol. 2011 Nov 1;29(31):4113-20.

13. Burns G, Cheng W-C. Tools for knowledge acquisition within the NeuroScholar system and their application to anatomical tract-tracing data. J of Biomedical Discovery and Collaboration. 2006;1(10).

14. Cañas AJ, et al. Concept maps: integrating knowledge and information visualization. In: Tergan SO, Keller T, editors. Knowledge and Information Visualization Searching for Synergies. Berlin/Heidelberg: Springer; 2005. p. 205-219.

15. Center for Biologics Evaluation and Research (CBER). 2010 http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm071072.htm

16. Centers for Disease Control and Prevention. National Center for Health Statistics. CDC WONDER On-line Database, compiled from Compressed Mortality File, 1999-2012. 2014;20(2R).

17. Ceusters W, Smith B and Goldberg L. A terminological and ontological analysis for the NCI Thesaurus. Methods of Information in Medicine. 2005;44:489-507.

18. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981 May;2(1):31-49.

19. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of out-comes in randomized trials: comparison of protocols to published articles. JAMA. 2004 May 26;291(20):2457-65.

20. Chen Q, Chen MH, Ohlssen D, Ibrahim JG. Bayesian modeling and inference for clinical trials with partial retrieved data following dropout. Stat Med. 2013 Apr 26.

21. Chen R and Snyder M. Promise of personalized omics to precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2013;5(1):73-82.

22. Chootrakool H, Shi JQ, Yue R. Meta-analysis and sensitivity analysis for multi-arm trials with selection bias. Stat Med. 2011 May 20;30(11):1183-98.

23. Clark GT, Mulligan R. Fifteen common mistakes encountered in clinical research. J Prosthodont Res. 2011 Jan;55(1):1-6.

24. Clinical Data Interchange Standards Consortium. 2012. http://www.cdisc.org/

25. Colwill JM, Cultice, Kruse RL. Will generalist physician supply meet demands of an increasing and aging population? Health Aff. 2008;27(3):232–241.

26. Coultas D. Ethical considerations in the interpretation and communication of clinical trial results. Proc. Am Thoracic Soc. 2007;4:194-199.

27. Davis R, Shrobe H, and Szolovits P. What is a knowledge representation? AI Magazine.1993;14(1):17-33.

28. Davis RB, Mukamal KJ. Hypothesis testing: means. Circulation. 2006 Sep 5;114(10):1078-82.

29. de Carvalho EC, Jayanti MK, Batilana AP, Kozan AM, Rodrigues MJ, Shah J, Loures MR, Patil S, Payne P, Pietrobon R. Standardizing clinical trials workflow representation in UML for international site comparison. PLoS One. 2010 Nov 9;5(11):e13893.

30. Druzdzel MJ and van der Gaag LC. Building probabilistic networks: where do the numbers come from? IEEE Transactions on Knowledge Engineering.2000;12(4)481-486.

31. Dumas M and ter Hofstede AHM. UML activity diagrams as a workflow specification language. In Proc of the UML Conference. 2001.

32. Eisenberg JM. Ten lessons for evidence-based technology assessment. JAMA. 1999 Nov 17;282(19):1865-9.

33. Gigerenzer G. Mindless statistics. Journal of Socio-Economics. 2004;33:587-606.

34. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ. 2008 Jun 28;336(7659):1472 4.

35. Goodman SN and Royall R. Evidence and scientific research. American Journal of Public Health. 1988;78:1568-1574.

36. Goodman SN. Towards evidence-based medical statistics. 1: The P-value fallacy. Ann Intern Med. 1999;130:995-1004.

37. Guarino N, Oberle D, Staab S. What is an Ontology? International Handbooks on Information Systems. 2009;Part 1:1-17.

38. Gupta SK. Use of Bayesian statistics in drug development: Advantages and challenges. Int J Appl Basic Med Res. 2012 Jan;2(1):3-6.

39. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evidence Baswed Medicine. 2007;11(6)162-164.

40. Held L. A nomogram for P values. BMC Med Res Methodol. 2010 Mar 16;10:21.

41. Hellems MA, Gurka MJ, Hayden GF. Statistical Literacy for Readers of Pediatrics: A Moving Target. Pediatrics. 2007;119:1083-1088.

42. Holland PW. Statistics and causal inference. Journal of the American statistical Association. 1986 Dec 1;81(396):945-60.

43. Hoogendam A, de Vries Robbé PF, Overbeke AJ. Comparing patient characteristics, type of intervention, control, and outcome (PICO) queries with unguided searching: a randomized controlled crossover trial. J Med Libr Assoc. 2012 Apr;100(2):121-6.

44. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, Shulz KF, CONSORT Group. CONSORT for Reporting Randomized Controlled Trials in Journal and Conference Abstracts: Explanation and Elaboration. PLoS Med 2008;5(1): e20.

45. Horsfield MA. Magnetization transfer imaging in multiple sclerosis. J Neuroimaging. 2005;15(4Suppl):58S-67S.

46. Horwitz RI. The experimental paradigm and observational studies of cause-effect relationships in clinical medicine. J Chronic Dis. 1987;40(1):91-9.

47. Hsu W, Speier W, Taira RK. Automated extraction of reported statistical analyses: Towards a logical representation of clinical trial literature. Proc AMIA Fall Symp, 2012. p.350-9.

48. Hubbard R, Lindsay RM. Why P-values are not a useful measure of evidence in statistical significance testing. Theory & Psychology. 2008;18(1):69-88.

49. Hubbard, R. and Armstrong, J. S. (2006). Why we don't really know what statistical significance means: Implications for educators. Journal of Marketing Education. 28(2):114-120.

50. Hyland K. Hedging in Scientific Research Articles. John Benjamins B.V., Amsterdam, Netherlands, 1998.

51. Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles Written Communication 13 (2): 251-281.

52. ICH E9 1998 http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html.

53. Ilic D, Tepper K, Misso M. Teaching evidence-based medicine literature searching skills to medical students during the clinical years: a randomized controlled trial. J Med Libr Assoc. 2012 Jul;100(3):190-6.

54. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294(2):218-228.

55. Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;2(8):696-701.

56. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, Moher D. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA. 1998 Jul 15;280(3):278-80.

57. Johnson DH, Fehrenbacher L, Novotny WF, Herbst RS, Nemunaitis JJ, Jablons DM, Langer CJ, DeVore RF 3rd, Gaudreault J, Damico LA, Holmgren E, Kabbinavar F. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. J Clin Oncol. 2004 Jun 1;22(11):2184-91.

58. Khan AM, Hahn JD, Cheng WC, Watts AG, Burns GA. NeuroScholar's electronic laboratory notebook and its application to neuroendocrinology. Neuroinformatics. 2006;4(2): 139-62.

59. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001 Jul 7;323(7303):42-6.

60. Kim MY, Goldberg JD. The effects of outcome misclassification and measurement error on the design and analysis of therapeutic equivalence trials. Stat Med. 2001 Jul 30;20(14):2065-78.

61. Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak. 2010 Sep 28;10:56.

62. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences. Publication Date: June 13, 2012. ISBN-10: 1412974453. ISBN-13: 978-1412974455. Edition: Fourth Edition

63. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med. 2001 Dec 4;135(11):982-9. Erratum in: Ann Intern Med. 2008 Aug5;149(3):219.

64. Korkontzelos I, Mu T, Ananiadou S. ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials. BMC Med Inform Decis Mak. 2012 Apr 30;12 Suppl 1:S3.

65. Kush RD. Current status and future scope of CDISC standards. Clinical Data Interchange Standards Consortium, October 2012.

66. Laine C, Horton R, DeAngelis CD, Drazen JM, FrizelleFA, Godlee F, Haug C, Hebert PC, Kotzin S, Marusic A, Sahni P, Schroeder TV, Sox HC, Van der Weyden MB, and Verheugt FW. Clinical trial registration--looking back and moving ahead. N Engl J Med. 2007;356(26): 2734-6.

67. Lehmann HP, Goodman SN. Bayesian communication: a clinically significant paradigm for electronic publication. J Am Med Inform Assoc. 2000 May-Jun;7(3):254-66.

68. Leon-Novelo LG, Zhou X, Bekele BN, Müller P. Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories. Biometrics. 2010 Sep;66(3):966-74.

69. Levin A. Reporting standards and the transparency of trials. Annals of Internal Medicine. 2001;134(2):169-172.

70. Light M, Qiu XY, and Srinivasan P. The language of bioscience: facts, speculatons, and statements in between. In BioLINK 2004: Linking Biological Literature, Ontologies, and Databases. 17-24.

71. Lin M, Lucas HC Jr., and Shmueli G. Too big to fail: Larger samples and false discoveries. Robert H. Smith School Research Paper No. RHS 06-068, June 15, 2011.

72. Marden JI. Hypothesis Testing: From p Values to Bayes Factors. Journal of the American Statistical Association. Dec., 2000;95(452): 1316-1320.

73. McAlister FA, Graham I, Karr GW, Laupacis A. Evidence-based medicine and the practicing physician. J Gen Intern Med. 1999;14:236-242.

74. McCance I. Assessment of statistical procedures used in papers in the Austrialian Veterinay Journal. Aust Vet J. 1995; 72:322-8.

75. Miller VA, Riely GJ, Zakowski MF, Li AR, Patel JD, Heelan RT, Kris MG, Sandler AB, Carbone DP, Tsao A, Herbst RS, Heller G, Ladanyi M, Pao W, and Johnson DH. Molecular characteristics of bronchioloalveolar carcinoma and adenocarcinoma,

bronchioloalveolar carcinoma subtype, predict response to erlotinib. J Clin Oncol. 2008;26(9):1472-8.

76. Mills EJ, Ioannidis JP, Thorlund K, Schünemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA. 2012 Sep 26;308(12):1246-53.

77. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010 Mar 23;340:c869.

78. Moher D, Jones A, Lepage L; CONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA. 2001 Apr 18;285(15):1992-5.

79. Mosenifar Z. Population Issues in Clinical Trials. Proceedings of the American Thoracic Society. 2007;4(2):185-188.

80. Neill PO, Sohal AS (1999) Business Process Reengineering: A Review of Recent Literature. Technovation 19:571–581.

81. Novak, JD, Gowin DB. Learning how to learn. Cambridge University Press, New York. 1984.

82. Pan G, Ke S, Zhao J. Comparison of the efficacy and safety of single-agent erlotinib and doublet molecular targeted agents based on erlotinib in advanced non-small cell lung cancer (NSCLC): a systematic review and meta-analysis. Target Oncol. 2013 Mar 21.

83. Pan Q, Schaubel DE. Proportional hazards regression in the presence of missing study eligibility information. Lifetime Data Anal. 2013 Jun 22.

84. Payne PRO, Eneida AM, Justin BS (2007) Modeling Participant-Related Clinical Research Events Using Conceptual Knowledge Acquisition Techniques. AMIA Annu Symp Proc 593–597.

85. Pearl J. Causality: Models, reasoning, and inference. Cambridge University Press, New York, 2000.

86. Pirker R, Pereira JR, von Pawel J, Krzakowski M, Ramlau R, Park K, de Marinis F, Eberhardt WE, Paz-Ares L, Störkel S, Schumacher KM, von Heydebreck A, Celik I, O'Byrne KJ. EGFR expression as a predictor of survival for first-line chemotherapy plus cetuximab in patients with advanced non-small-cell lung cancer: analysis of data from the phase 3 FLEX study. Lancet Oncol. 2012 Jan;13(1):33-42.

87. Price KA, Azzoli CG, Krug LM, Pietanza MC, Rizvi NA, Pao W, Kris MG, Riely GJ, Heelan RT, Arcila ME, Miller VA. Phase II trial of gefitinib and everolimus in advanced non-small cell lung cancer. J Thorac Oncol. 2010 Oct;5(10):1623-9.

88. Roland M-C. Publish or perish: Hedging and fraud in scientific discourse. European Molecular Biology Organization Reports. 2007;8(5)424-428.

89. Rubin DB. Bayesian inference for causality: The importance of randomization. InThe Proceedings of the social statistics section of the American Statistical Association 1975;233-239.

90. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology. 1974;66(5):689.

91. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996 Jan 13;312(7023):71-2.

92. Sackett DL, Straus SE, Richardson WS, Rosenberg W, and Haynes RB. Evidence-based Medicine: How to Practice and Teach EBM. (2nd Ed.) New York: Churchill Livingstone; 2000.

93. Sackett DL. Evidence-based medicine and treatment choices. Lancet. 1997 Feb 22;349(9051):570; author reply 572-3.

94. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. BMC Medicine. 2010;8:18.

95. Sellke, Thomas; Bayarri, M. J.; Berger, James O. Calibration of p values for testing precise null hypotheses. The American statistician 55:11, 62-71

96. Senn SJ. Falsificationism and clinical trials. Stat Med. 1991 Nov;10(11):1679-92.

97. Sequist LV, Martins RG, Spigel D, Grunberg SM, Spira A, Jänne PA, Joshi VA, McCollum D, Evans TL, Muzikansky A, Kuhlmann GL, Han M, Goldberg JS, Settleman J, Iafrate AJ, Engelman JA, Haber DA, Johnson BE, Lynch TJ. First-line gefitinib in patients with advanced non-small-cell lung cancer harboring somatic EGFR mutations. J Clin Oncol. 2008 May 20;26(15):2442-9.

98. Sestini P and Rossi S. Exposing the P value fallacy to young residents, at the 5th International Confrence of Evidence-based Health care Teachers and Developers, Taormina, October 29 2009.

99. Sharma SV, Bell DW, Settleman J, Harber D. Epidermal growth factor receptor mutations in lung cancer. Nature. 2007; Vol 7.

100. Silva AJ and Muller KR. The need for novel informatics tools for integrating and planning research in molecular and cellular cognition. Learn Mem, 2015. 22(9): p. 494-8.

101. Sim I, Carini S, Tu S, Wynden R, Pollock BH, Mollah SA, Gabriel D, Hagler HK, Scheuermann RH, Lehmann HP, Wittkowski KM, Nahm M, Bakken S. The human studies database project: federating human studies design data using the ontology of clinical research. AMIA Summits Transl Sci Proc. 2010 Mar 1;2010:51-5.

102. Sim I, Olasov B, and Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. Journal of Biomedical Informatics. 2004;37:108-119.

103. Sim, I. (2007). The Trial Bank Project http://rctbank.ucsf.edu

104. Sim I, Olasov B, and Carini S. The Trial Bank System: Capturing Randomized Trials for Evidence-Based Medicine. AMIA Annu Symp Proc. 2003; 2003: 1076.

105. Sim I, Owens DK, Lavori PW, Rennels GD. Electronic trial banks: a complementary method for reporting randomized trials. Med Decis Making. 2000 Oct-Dec;20(4):440-50.

106. Smyth RM, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. BMJ. 2011 Jan 6;342:c7153.

107. Soldatova LN, King RD. An ontology of scientific experiments. J R Soc Interface. 2006 Dec 22;3(11):795-803.

108. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ. 2001 Jan 27;322(7280):226-31.

109. Steves R, Hootman JM. Evidence-Based Medicine: What Is It and How Does It Apply to Athletic Training? J Athl Train. 2004 Mar;39(1):83-87.

110. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research--a review of common pitfalls. Swiss Med Wkly. 2007 Jan 27;137(3-4):44-9.

111. Tai-Seale M, McGuire TG, Zhang W. Time allocation in primary care office visits. Health Serv Res. 2007 Oct;42(5):1871-94.

112. Taylor AL, Wright JT Jr. Should ethnicity serve as the basis for clinical trial design? Importance of race/ethnicity in clinical trials: lessons from the African-American Heart Failure Trial (A-HeFT), the African-American Study of Kidney Disease and Hypertension (AASK), and the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). Circulation. 2005 Dec 6;112(23):3654-60; discussion 3666.

113. The Biomedical Research Integrated Domain Group (BRIDG) Model. <http://www.bridgmodel.org/>

114. The Clinical Data Interchange Standards Consortium (CDISC) 2013. <http://www.cdisc.org>

115. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol. 2000 Feb;53(2):207-16.

116. Tong M, Hsu W and Taira RK. A representation for standardizing numerical data from clinical trial reports. 2012 RSNA Scientific Assembly and Annual Meeting Bioinformatics Exhibit.

117. Tong M, Hsu W and Taira RK. A Formal Representation for Numerical Data Presented in Published Clinical Trial Reports. Studies in Health Technology and Informatics. 2013;192:856-60.

118. Uner H, Nezami FG, Yildirim MB, Dong F, Wellner J, Bradham DD. Visit length in pediatric primary care: lessons from a pilot study. J Med Pract Manage. 2013 May-Jun;28(6):363-70.

119. Urick PN. Implementing community-based standards of care. J Manag Care Pharm. 2005 May;11(4 Suppl):S11-6.

120. Van der Vet PE, Mars NJI. Bottom-up construction of ontologies. IEEE Jul/Aug 1998 Vol 10, Issue 4, 513-526.

121. Walker E, Hernandez AV, Kattan MW. Meta-analysis: Its strengths and limitations. Cleve Clin J Med. 2008 Jun;75(6):431-9.

122. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc. 2011 Dec;18 Suppl 1:i116-24.

123. West CP, Ficalora RD. Clinician attitudes toward biostatistics. Mayo Clinic Proc. 2007;82(8):939-943.

124. Willemsen AM, Jansen GA, Komen JC, van Hooff S, Waterham HR, Brites PM, Wanders RJ, van Kampen AH. Organization and integration of biomedical knowledge with concept maps for key peroxisomal pathways. Bioinformatics. 2008 Aug 15;24(16):i21-7.

125. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA. 2007;298(9):1010-1022.

126. Wood BP. What's the evidence? Radiology. 1999 Dec;213(3):635-7.

127. Xie H. Bayesian inference from incomplete longitudinal data: a simple method to quantify sensitivity to nonignorable dropout. Stat Med. 2009 Sep 30;28(22):2725-47.

128. Yap TA, Vidal L, Adam J, Stephens P, Spicer J, Shaw H, Ang J, Temple G, Bell S, Shahidi M, Uttenreuther-Fischer M, Stopfer P, Futreal A, Calvert H, de Bono JS, Plummer R. Phase I trial of the irreversible EGFR and HER2 kinase inhibitor BIBW 2992 in patients with advanced solid tumors. J Clin Oncol. 2010 Sep 1;28(25):3965-72.

129. Zarin DA, Tse T, Ide NC. Trial registration at ClinicalTrials.Gov between May and October 2005. New England Journal of Medicine. 2005;353(26): 2779-2787.

130. Zheng K, Guo MH, Hanauer DA. Using the time and motion method to study clinical work processes and workflow: methodological inconsistencies and a call for standardized research. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):704-10.

Appendix

Appendix A. Example Task Questions and Answers

Task questions **Johnson et al.**

Comprehension about scientific claims

Please time responses to the nearest minute. Please record a start time, and end times to each question.

Start Time :_____

1. What is the objective or hypothesis of this trial and primary and secondary outcome measure?
   Objective: _____
   Primary Outcome:_____
   Secondary Outcome:_____

Time :_____

2. The abstract states: " Bevacizumab in combination with carboplatin and paclitaxel improved overall response and time to progression in patients with advanced or recurrent non-small cell lung cancer." To what degree did this treatment improve overall response?
   Name the Statistical Test:_____
   List the number of participants in the experimental arm: _____
   List the number of participants in the control arm: _____
   Name the method(s) of assessment (imaging, biomarkers, etc.):_____
   Time points assessments were taken: Circle.
   a. 1 hour after each cycle
   b. after cycles 3,6,10,14,18
   c. every 3 weeks
   d. greater than 4 weeks after initial documentation
   e. every 2 months until death or loss to follow-up
   f. other: _____
   Results of Statistical test:_____
   What is the significance of this stat test result?_____
   _____

Time :_____

3. From the experimental arm, how many patients discontinued treatment and why?
   _____

_____

Time :_____

4. How many patient experience positive/negative outcomes in this trial? Example of positive outcome: efficacy of drug, stable disease; example of negative outcomes: disease progression, death.

   _____

   _____

Time :_____

5. Fill in. Ex: "better" or "worst." The results for patients with non-squamous cell histology had _____ outcome than patients without.

End Time :_____

Break time!

Information Retrieval

Please time responses to the nearest minute. Please record a start time, and end times to each question.

Start Time :_____

6. List the control, and experimental arm(s) under Group Name, write down the number of participants and drug dosage for each group.

Group Name	Number of Participants	Dosage
_____	____	_____
_____	____	_____
_____	____	_____
_____	____	_____

Time :_____

7. What is the eligibility criteria, regarding cancer status stage? _____

Time :_____

8. How and when was primary outcome assessed? List method of assessment (imaging, , biomarkers, etc) and time point and frequency of assessment.

Method of Assessment	Time point/Frequency

Time :_____

9. Is there a difference in median TTP between high-dose and control? Was statistical significance achieved?

   _____

   _____

Time :_____

10. Is there a difference between survival for high dose and control? Was statistical significance achieved?

    _____

    _____

End Time :_____

Task answers **Johnson et al.**

Comprehension about scientific claims

Please time responses to the nearest minute. Please record a start time, and end times to each question.

Start Time :_____

1. What is the objective or hypothesis of this trial and primary and secondary outcome measure?
   Objective: Safety and Efficacy
   Primary Outcome Tumor response rate and TTP using Kaplan Meier Curves
   Secondary Outcome: overall survival and duration of response

Time :_____

2. The abstract states: " Bevacizumab in combination with carboplatin and paclitaxel improved overall response and time to progression in patients with advanced or recurrent non-small cell lung cancer." To what degree did this treatment improve overall response?
   Name the Statistical Test: 2 sided chi squared test
   List the number of participants in the experimental arm: 67
   List the number of participants in the control arm: 32
   Name the method(s) of assessment (imaging, biomarkers, etc.): Tumor Status OR Imaging
   Time points assessments were taken: Circle.
   a. 1 hour after each cycle
   b. after cycles 3,6,10,14,18
   c. every 3 weeks
   d. greater than 4 weeks after initial documentation
   e. every 2 months until death or loss to follow-up
   f. other: _____

Results of Statistical test: None

What is the significance of this stat test result? Overall response improved

Time :_____

3. From the experimental arm, how many patients discontinued treatment and why?

   Eleven patients discontinued treatment as a result of a nonfatal AE. Discontinuations occurred as a result of: hemorrhagic event (three patients) in the low-dose bevacizumab arm; a hemorrhagic event (one patient); Aspergillus lung abscess (one patient); aspiration pneumonia (one patient); thrombotic stroke (one patient); vertebral fracture (one patient); and peripheral neuropathy (paclitaxel-related; one patient) in the high-dose arm. In two cases, bevacizumab was discontinued following initiation of anticoagulant therapy. Bevacizumab was withheld from one patient with subclavian vein thrombosis.

Time :_____

4. How many patient experience positive/negative outcomes in this trial? Example of positive outcome: efficacy of drug, stable disease; example of negative outcomes: disease progression, death.

   Based on investigator, 85 experience disease progression, 19 control patients crossed over to single-agent bevacizumab, 9 patients died as a result of AE

Time :_____

5. Fill in. Ex: "better" or "worst." The results for patients with non-squamous cell histology had better outcome than patients without.

End Time :_____

Break time!

Information Retrieval

Please time responses to the nearest minute. Please record a start time, and end times to each question.

Start Time :_____

6. List the control, and experimental arm(s) under Group Name, write down the number of participants and drug dosage for each group.

Group Name	Number of Participants	Dosage
Control	32	0
Low dose Bevacizumab	32	7.5 mg/mL
High dose Bevacizumab	35	15 mg/mL
_____	____	_____

Time :_____

7. What is the eligibility criteria, regarding cancer status stage? stage IIIB, stage IV, recurrent NSCLC

Time :_____

8. How and when was primary outcome assessed? List method of assessment (imaging, , biomarkers, etc) and time point and frequency of assessment.

Method of Assessment	Time point/Frequency
ECOG Tumor Response/Imaging	After cycles 3,6,10, 14, 18
TTP	Every 2 months

Time :_____

9. Is there a difference in median TTP between high-dose and control? Was statistical significance achieved?
   Yes. Using the log-ranked test, Investigator: 7.4 vs 5.9, p=0.023. Independent Research Faciility:7.0 vs 5.9, p=0.185.

Time :_____

10. Is there a difference between survival for high dose and control? Was statistical significance achieved?
    Possibly. Using the log-ranked test, 17.7 vs 14.9, p=0.63

End Time :_____

Appendix B. Common Query Questions

Start Time:_____

Clinical Trial Objectives and Study Design – 3 questions

1. What is/are the study's objective(s)? What is the clinical reasoning as to why this study was created? (Clinical Question)

   _____

   _____

Start Time: _____

End Time: _____

2. By the end of the paper, does the paper answer this/these objectives? If no, describe which objective(s) was/were not answered. (Clinical Question)

   Yes No

   _____

   _____

Start Time: _____

End Time: _____

3. The following questions relate to study design: (Biostat Question)
   a. How many experimental arms are there? _____

      Please note if control is included. Ex: "3 total (includes 1 control am)"
   b. What is the phase of the clinical trial? _____
   c. What types of statistical analyses are performed?

      _____

      _____

Start Time: _____

End Time: _____

Interventions – 1 question

4. The following questions relate to the interventions: (Clinical Question)
   a. What is the medication name dosing strength and frequency?

i. Medication Name: _____
ii. Dosing Strength: _____
iii. Frequency:

_____

_____

b. What is the protocol for the experimental group(s)?

_____

_____

c. Does the treatment regimen account for dose interruption or reduction during the study? If yes, what was changed?

Yes No

_____

_____

Start Time: _____

End Time: _____

Recruited Population – 3 questions

5. Describe the target population (i.e., total population) in terms of: (Clinical Question)
   a. Age - Median Age and Range:

   _____

   b. Age - % < 65 years and % ≥ 65 years

   _____

   c. Gender - % Male and % Female:

   _____

   d. Ethnicity - % White, % Black, % Hispanic, % Asian %Native Hawaiian Pacific Islander % Other:

   _____

   e. Geographic location(s)/Institution(s):

   _____

Start Time: _____

End Time: _____

6. What is the proportion of patients with tumor stage IIIB in each group (by gene expression, treatment) and overall? (Biostat Question)

Group Name	% with tumor stage IIIB

Start Time: _____

End Time: _____

7. List each control/comparator group and characterize each control/comparator group mentioned as: (a) standard care control, (b) placebo control, (c) no medication control, or (d) other control. If other, describe. (Clinical Question)

Group Name	Characterization

Start Time: _____

End Time: _____

Results – 5 questions

8. What is the response rate of treatment groups? If given, list the time points for each response rate for the treatment groups. (Biostat Question)

Group Name	Response Rate	Time point, if any

Start Time: _____

End Time: _____

9. If applicable, what is the median survival or progression free survival (PFS) and hazard ratio in each experimental group? (Biostat Question)

Group Name	Median Survival or PFS	Hazard Ratio

End Time: _____

Start Time: _____

10. List all Grade 3 and above adverse events (or side effects) for a) the control and/or comparator group(s) and b) the experimental group? (Clinical Question)

_____

_____

Start Time: _____

End Time: _____

11. What are the causes of death (if any) in the control/comparator group(s) and experimental arms? For example, this can be seen in the participant flow decision nodes on the right of the process model. (Clinical Question)

    Causes of death in control/comparator group:

    _____

    _____

    Causes of death in experimental group:

    _____

    _____

    Causes of death in _____ group:

    _____

    _____

Start Time: _____

End Time: _____

12. What are the top 3 adverse events in the treatment group(s)? For each adverse event, how many patients experienced that adverse event? (Biostat Question)

Top 3 Adverse Event	No. of patients
1)	
2)	
3)	

Start Time: _____

End Time: _____

13. What is the proportion and number of patients that dropped out (a) in the treatment arm(s) and (b) in the control arm? (Clinical Question)

Group Name	No. and	% drop outs
a)		
Reasons:		
a)		
Reasons:		
b)		
Reasons:		

Start Time: _____

End Time: _____

Statistical Effect – 2 questions

14. For the most significant statistical test, describe how the test statistic (hazard ratio, t-test, chi-squared, etc.) was calculated? Describe (a) variables (i.e., overall response, time to progression), (b) test statistics used, (c) sample size, d) any multiple comparison adjustments. (Biostat Question)
    a. Variables tested:

       _____

    b. Name the statistical test:

       _____

    c. List number of participants in experimental arm: _____
    d. List number of participants in comparison/control arm: _____
    e. Was statistical significance achieved?

       Yes No

       _____

    f. Were there any multiple comparison adjustments?

       Yes No Not mentioned Other:

       _____

Start Time: _____

End Time: _____

15. What is the context for the most significant statistical result (i.e, p-value)? Describe <u>the analyzed population (not the total population in Q5)</u> in terms of (a) the age, demographic of population, sample size; (b) interventions; (c) methodology used to collect data; (d) time point and frequency of assessment. (Biostat Question)
    a. Age:

       _____

    b. <u>Gender and Ethnicity</u>:

       _____

    c. Sample size:

       _____

d. Intervention:

_____

e.

Method of Assessment	Time point/Frequency

Start Time: _____

End Time: _____

Clinical Effect – 1 question

16. The following questions relate to clinical effect:

a. List the outcome measures that help determine the intervention's clinical relevance (i.e., quality of life markers, survival metrics, etc.)? (Clinical Question)

_____

b. If quality of life is addressed, describe method used to measure quality of life? (Clinical Question)

_____

c. If this is a survival study, how many more months/days does the intervention prolong life? Please list how many more months/day for each measure (i.e., PFS, OS, TTP, etc.) (Clinical Question)

_____

Start Time: _____

End Time: _____

Appendix C. Pre- and Post-questionnaires

Pre-questionnaire Form

Department: _____

Level:

_____Undergrad Student

_____Graduate Student Year_____

_____Post-doc

_____Faculty

Experience with lung cancer disease and therapy

1. On a scale of 1-10, 1 being completely clueless and 10 being a domain expert, what is your understanding of the mechanisms of cancer? _____

2. How many classes on biology have you taken and have understood the material?

Undergraduate courses: _____ Graduate courses:_____

3. On a scale of 1-10, 1 being completely uncomfortable and 10 being very comfortable, how comfortable are you with your knowledge on non-small cell lung cancer (NSCLC)? _____

Experience with clinical trial reports

1. Have you read a clinical trial report before? Y / N

2. If yes, how long on average does it take you to read a clinical trial report? _____minutes

3. Please rank how well you understand clinical trial papers on a scale of 1-10, 1 being completely confused and 10 being ready to apply the knowledge. _____

Experience with statistics

1. On a scale of 1-10, 1 being completely uncomfortable and 10 being very comfortable, how comfortable are you with understanding statistical methods and results?

2. How many courses on statistics have you taken and understood the material?

Undergraduate courses: _____ Graduate courses:_____

3. On a scale of 1 to 10, 1 being completely uncomfortable and 10 being ready to design statistical experiments, how confident are you at assessing the quality of a statistical test or developing your own statistical tests?_____

Please comment on any additional experiences not mentioned in this questionnaire:

_____

_____

Post-Questionnaire

Preferences

1. Does the visualization show the purpose of the trials? Y/N
2. Does the visualization show the recruitment? Y/N
3. Does the visualization show the interventions, including details of dosage, if applicable? Y/N
3. Does the visualization show the data? Y/N
4. Does the visualization show the results, including statistical methods? Y/N
5. Does the visualization show the conclusions? Y/N
6. On a scale of 1-10, 1 being completely useless and 10 being completely essential, how useful was the visualization in helping you understand clinical trials? _____
7. What is your preference? Paper Report "Status quo"/Visualization
8. On a scale of 1-10, 1 being not using the visualization again and 10 being will use the visualization again, what is the likelihood that you will use this visualization? _____
9. On a scale of 1-10, 1 being totally unsatisfied and 10 being highly satisfied, how satisfied are you with the visualization in its current state? _____
10. Comments:

_____

_____