TRIAL BANKS:
AN INFORMATICS FOUNDATION FOR EVIDENCE-BASED MEDICINE

A DISSERTATION
SUBMITTED TO THE PROGRAM IN MEDICAL INFORMATION SCIENCES
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ida Sim
December 1997

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

_____

Glenn D. Rennels, Principal Adviser
(Section on Medical Informatics)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

_____

Douglas K. Owens, Co-Principal Adviser
(Section on Medical Informatics)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

_____

Philip W. Lavori
(Department of Health Research and Policy)

Approved for the University Committee on Graduate Studies

_____

Abstract

Millions of dollars are spent annually on the conduct of randomized clinical trials, a type of experiment widely regarded as yielding the most valuable evidence for improving our understanding of medicine. Yet the results of many large and important clinical trials are published only as text-based articles in the clinical literature, articles that both practitioners and clinical researchers have difficulty finding, interpreting, and applying to clinical care. The result is an inefficient transfer of evidence from the research world to the clinic, and a waste of precious resources.

It is, however, not only the deficiencies of randomized-trial reports that contribute to this evidence-transfer problem; our difficulties with using randomized-trial evidence stem from problems that involve the entire lifecycle of trials — from their design, registration, standardization, and publication, to the synthesis of their results. Thus, I propose a comprehensive *trial-centered*, rather than article-centered, solution to the evidence-transfer problem: a trial-bank system.

In the **trial-bank system**, trial investigators will report randomized trials not only as text articles in traditional medical journals, but also as entries into standardized, structured, electronic databases, or **trial banks**. Such dual-format publication already exists: Bioinformatics researchers publish their genomic sequencing results in GenBank — a structured database administered by the National Institutes of Health — and discuss the implications of their work in a prose article. The journal that publishes the work appends to the article the GenBank accession number of the sequence. Readers of the article can thus immediately access and analyze the reported sequence data via the World Wide Web.

Abstract

If trial investigators authored their randomized trials directly into trial banks, we could expect several significant benefits. First, trial-bank–authoring software can help authors to describe their trials accurately and completely according to community-defined standards. Second, trial-bank–presentation software can customize the display of the evidence to suit the needs of various users; this flexibility contrasts with the present day, one-size-fits-all trial report. Third, trial banks will be up-to-date knowledge bases that can facilitate systematic reviews, meta-analysis, and information retrieval at the point of care. Trial banks can also facilitate the development of expert systems that reason about clinical trials, because individual expert systems will no longer be required to maintain their own knowledge bases of trials. Fourth, trial banks can anchor an informatics infrastructure in which all the evidence for clinical practice — including trials, systematic reviews, decision analyses, and practice guidelines — are integrated with the computer-based patient record to support evidence-based medicine at the point of care.

Because there will probably be many trial banks worldwide, we should strive to minimize the duplication of trial-bank entries. We should also strive to maximize access to trial-banks, so that systematic reviews of randomized trials can be as comprehensive as possible. To achieve these goals, it is vital that trial banks worldwide be interoperable. That is, users should be able to access trial banks worldwide as if they were one single trial bank — an integrated trial-bank system. To achieve interoperation, trial banks must have a common understanding of the clinical-trial concepts that are to be shared among the trial banks. Such a computer-based, common understanding of a domain can be encoded as a **conceptual model** that abstractly defines the meaning of the concepts in a domain, and the relationships among those concepts. Conceptual models can be encoded in natural language (e.g., English), first-order logic, or in a number of database definition languages.

The centerpiece of my thesis work is the design, implementation, and evaluation of a clinical-trials conceptual model for the interoperation of trial banks. I devised the **competency-decomposition** approach for compactly describing my conceptual-modeling work. With this approach, I state explicitly that the trial-bank system should help its users to per-

form the four core tasks of evidence synthesis: (1) trial retrieval; (2) trial critiquing; (3) quantitative computation; and (4) the interpretation of trials in their scientific, socioeconomic, and ethical context. I then specified the concepts that the clinical-trials conceptual model must include if it is to support these target tasks. The resulting competency decomposition can act (1) as a design specification to guide the construction of new trial banks, (2) as documentation for the competencies of an implemented clinical-trials conceptual model, and (3) as a yardstick for the evaluation of whether or not a conceptual model can indeed support the four core tasks of evidence synthesis.

I demonstrate that my design specification for a clinical-trials core conceptual model is reasonable by comparing its data requirements to those of 18 published trial-critiquing instruments. I show that my implementation of a clinical-trials conceptual model has sufficient conceptual coverage to be competent for three of the four core tasks of evidence synthesis for a broad range of randomized-trial types. I also show that health-services researchers were able to use a web-based presentation system for my clinical-trial knowledge base to complete a trial-critiquing questionnaire about a published trial.

With the advent of digital publication, we have a window of opportunity to design our publication systems such that they support the transfer of evidence from the research world to the clinic. An trial-bank system is a first step towards a comprehensive information infrastructure for assisting medical practitioners with applying the most-up-to-date scientific evidence to clinical care. This dissertation presents foundational work for the design and construction of an interoperating trial-bank system that will help us achieve the day-to-day practice of evidence-based medicine.

Acknowledgments

I took Ted Shortliffe's course on Medical Informatics as a second-year Stanford medical student in 1986. After seven years that were dominated by clinical training, I returned to medical informatics and entered the MIS degree program in 1993. It has since been a tremendous journey through a fascinating landscape of ideas.

Thanks are due to many for making these four-plus years such a delight. I thank Ted Shortliffe for the vision and dedication he brings to the Stanford MIS training program. I will draw inspiration from his example as I train my own MIS students. I thank Glenn Rennels, my principal advisor, for sharing with me his enthusiasm, his experience, his open criticism, and his common sense. Glenn provided the informatics grounding for my work in more ways than one: His thesis work on Roundsman inspired this work, and his solid grasp of what constitutes rigorous informatics research clarified many a research dilemma for me.

Although this is a dissertation in Medical Informatics, the problems that I address come from the domain of Health Services Research. I was extremely fortunate that I was able to complement my informatics work with my concurrent fellowship in health services. For this, I have Doug Owens and Mark Hlatky to thank. Doug's research-fellowship program at the Palo Alto Veterans Affairs hospital was a wonderful clinical and professional home for me. Doug has also been a constant source of support and career assistance. I am grateful to Mark for the opportunity to work with a first-class team on his Cardiac Arrhythmia

Patient Outcomes Research Team. Mark's supportive mentoring has given me a valuable start on my own academic career.

One of the first-class members of the PORT team that I have had the privilege to work with is Phil Lavori. Phil has a knack for explaining statistics with folksy lyricism. I thank Phil also for sharing his perspectives on the place of clinical-trial evidence in the wider scientific and ethical context of health care.

My years here were rounded out by hallway, cubical, and Beckman-cafeteria conversations with many other people. I thank Les Lenert for his no-nonsense career advice; Kathy McDonald for being a colleague and a friend; Anne Dembitzer for her New Yorker's instinct for what matters; my fellow students in medical informatics — particularly Wanda Pratt, Ramon Felciano, Malcolm Pradhan, Maria Tovar, and Gillian Sanders — for their camaraderie; and my fellow fellows in health services research for always asking tough questions.

This work was supported by the Veterans Affairs Health Services Research and Development Service through both its Ambulatory Care Practice and Research Fellowship, and its Medical Informatics Fellowship; and by the Agency for Health Care Policy and Research grant HS-08362. I thank the Veterans Affairs Cooperative Studies Program and Dr. Ken James for allowing me access to the records of the SPINAF trial. The Stanford computing resources for this work were supported by the CAMIS resource grant LM-05305 from the National Library of Medicine, and the computing resources for my RCT Presenter system were provided by the Artificial Intelligence Center at SRI International.

This is the first of many projects for which I anticipate thanking my husband, Peter Karp. I cherish his constant and unstinting love, support, and understanding. I value also his expert contributions to this work. Finally, I dedicate this thesis work to my parents, who took on risks in emigrating from Hong Kong to Canada, and more risks in deciding to send me to school at Stanford. I appreciate all that they have done, and I cannot thank them enough.

Table of Contents

Chapter 2 The Clinical Literature

Chapter 3 The Trial-Bank System

Chapter 4 Conceptual Modeling

Chapter 5 The Core Conceptual Model

List of Tables

List of Figures

Chapter 1

The Evidence-Transfer Problem

Millions of dollars are spent annually on the conduct of **randomized clinical trials**, a type of experiment widely regarded as yielding the most valuable evidence for improving our understanding of medicine. The results of these trials, however, are often difficult to find, interpret, or apply to clinical care. In this dissertation, I explore the reasons behind this evidence-transfer problem, and I propose a new approach for reporting and disseminating randomized-trial results to increase the return on our society's large investment in these studies.

1.1 Clinical Evidence: From the Literature to the Clinic

The practice of Western medicine used to be one of apprenticeship. Anecdote, past experience, and authority were the guides to delivering care. In 1992, a group of medical educators coined a new term and revamped the thinking about the evidential basis for clinical practice (EB Medicine Working Group, 1992). They argued that medicine should be **evidence based**: To the extent possible, doctors should deliver care that is justified by scientific evidence, rather than by the traditional triad of anecdote, experience, and authority. Randomized clinical trials, in which groups of patients are treated in controlled situations

and their outcomes carefully recorded, were lauded as a major source of evidence for this new, modern, practice of medicine. The call was for clinicians to keep abreast of the clinical research results reported in medical journals, and to base their everyday practice on a sound consideration of those results (Haynes, 1986).

It has been a few years since the initial rally to evidence-based medicine, and many clinicians now consider evidence-based medicine to be a laudable but impractical ideal. Finding, retrieving, interpreting, and applying just a single article to a particular clinical case can take a prohibitive amount of time, skill, and work. To use a completely evidence-based approach, doctors would have to appraise critically the evidence from tens or even hundreds of relevant articles. It is simply not realistic to expect busy practitioners to undertake such a daunting task. Rather, the more recent expectation is that clinical and methodological specialists will synthesize the evidence reported in the clinical literature, and will describe the relevant results in articles aimed at practitioners. The practitioners will then read these reviews, and will apply the expertly synthesized evidence to care at the clinic (Williamson, 1989; Guyatt, 1993; Cook, 1997). The task of reviewing and interpreting the barrage of new scientific evidence now rests squarely on the shoulders of the evidence synthesizers — a group that includes authors of reviews and textbooks, and makers of clinical guidelines and health policy.

Unfortunately, this newer, more realistic view of evidence-based medicine also is laudable but unrealized. One major reason that evidence synthesizers are not keeping up with the literature is that the quantity of evidence in the clinical literature is overwhelming, and a careful review of the literature takes tremendous time, work, and expertise. Because the literature is published as text, the information-management power of the computer cannot be fully harnessed to assist with the time- and labor-intensive tasks that evidence synthesizers must perform routinely. There are four **core tasks** of evidence synthesis; for all four tasks, the synthesizer could be greatly assisted by the computer:

1. Retrieve all randomized trials that are relevant to a particular medical decision.
2. Critique each trial, by judging the trial’s internal validity and generalizability.

3. Synthesize the results from all relevant trials, using the statistical technique of meta-analysis to combine quantitative results when appropriate.
4. Interpret the totality of the results in the scientific, socioeconomic, and ethical context of health care.

The broad hypothesis of my dissertation is that, if randomized clinical trials are reported into structured databases, called **trial banks**, that are shared throughout the world, then the scientific evidence from those clinical trials will be transferred more easily to the frontline practice of medicine. Before discussing the evidence-transfer problem and its solution in detail, I itemize the long-term objectives and the specific aims of this work.

1.1.1 Long-Term Objectives

The following are the 5- to 10-year objectives of this work.

1. My work will be a principled foundation for a network of structured databases of clinical-trial information that will allow computers to assist medical practitioners with applying randomized-trial evidence to clinical care correctly and expeditiously. These structured databases, called trial banks, will constitute a **trial-bank system**. Trial investigators will themselves write into these trial banks, as an integral step in publishing their trials in academic journals, in applying for government or other funding, or in seeking regulatory approval for therapeutics.
2. My work will help ensure that all the trial banks in this trial-bank system are **interoperable**, using a **core conceptual model** of clinical trials. A clinical-trials core conceptual model is a computer-understandable encoding of clinical-trial concepts necessary and sufficient for supporting the core tasks of evidence synthesis. With this core conceptual model, the trial-bank system will appear as one to users, since all trial banks worldwide will be accessible using a common protocol.

1.1.2 Specific Aims

The following are the specific aims that I achieved in this dissertation.

1. I developed a **design specification —** a blueprint of the desired contents — for a **core conceptual model** of clinical trials. I demonstrate that the design specification is reasonable and extensible.
2. I implemented **Ocelot-CCM**, a conceptual model that is based on the design specification for the clinical-trials core conceptual model. I show that Ocelot-CCM meets the design specification for a broad range of randomized trials, and that the model is extensible.
3. I built the **RCT Presenter** system: a web-based interface for browsing a trial bank that is based on Ocelot-CCM. I evaluated the use of this system by health-services researchers to perform a trial-critiquing task.

1.1.3 Significance

With the advent of widespread electronic publication, we have an opportunity to publish randomized trials directly into an informatics infrastructure that is expressly designed to help clinicians manage and apply the valuable evidence from these trials. I have drawn ideas from the database and knowledge-engineering disciplines, and from my expertise in evidence-based medicine, to define the critical components of this evidence-based–medicine informatics infrastructure. A shared conceptual model of clinical trials is one such critical component. My work presents a principled design specification for this critical component, and my implementation of a shared conceptual model lays the groundwork for deploying a full-scale trial-bank system.

1.2 Evidence-Based Medicine

The seminal 1992 article by the Evidence-Based Medicine Working Group stated that evidence-based medicine involved “problem defining, [followed by] searching, evaluating,

and applying original medical literature" to the care of a particular patient (EB Medicine Working Group, 1992). The Working Group assumed that the practice of evidence-based medicine results in health outcomes better than those resulting from the practice of traditional, authority-centered medicine, although this assumption has never been verified experimentally. Nevertheless, I accept without further question the assumption that evidence-based medicine is the goal toward which we should strive.

1.2.1 Need for Evidence Synthesis

The sheer volume of information makes it almost impossible to apply evidence from the clinical literature to a particular clinical decision. The scientific evidence may bear on a clinical decision in three possible ways, although clinicians lack the literature-management tools to help them decide which of these scenarios pertain in a particular case.

1. *Good studies, easy answer* — In the ideal evidence-based–medicine scenario, well-conducted randomized trials exist that consistently show the superiority, or the inferiority, of one intervention over another. For example, prophylactic lidocaine has never been shown in randomized trials to reduce mortality in acute myocardial infarction compared to placebo (Antman, 1992). This prophylactic use of lidocaine is clearly not supported by the scientific evidence.
2. *Reasonably good studies, but interpretation requires advanced methodology* — In a more complicated, although more common, scenario, a mixture of high-quality and lower-quality trials yields equivocal support for any one course of action. Both clinical and biostatistical expertise is needed to synthesize the evidence, and reasonable specialists may differ in their interpretations of the evidence. An example of this scenario is the randomized-trial evidence on the ability of the drug *amiodarone* to reduce the risk of sudden death in patients who have preexisting heart disease. Individual trials have yielded seemingly contradictory evidence, and the proper interpretation appears to involve differences in the experimental methods of the trials (Sim, 1997).

3. *Lower-quality studies, offering little useful information* — It has been estimated that less than 50 percent of clinical practice is supported by even modest scientific evidence.[1] Whether or not a given clinical practice belongs to this 50 percent is often difficult to discern. An example is the prevalent use of the drug sotalol to reduce the risk of sudden death in patients who have preexisting heart disease. The clinical literature yields scant support for this clinical practice. Few randomized trials on sotalol's protective efficacy have been reported, and the nonrandomized studies have methodological weaknesses that preclude any trustworthy conclusions; yet it took me and my colleagues on the Cardiac Arrhythmia and Risk of Death project weeks to establish that this reasonably common clinical practice is not evidence based.

As these examples show, the scientific evidence in the clinical literature is not in a form that can be applied directly to clinical care. Evidence must be synthesized before it can be applied, and this synthesis is the "basis for our understanding of reality, the basis for our decisions, and a determinant of our future" (Eddy, 1992, p. 1).

Because practitioners are frequently pressed for time during their patient encounters, evidence synthesis is rarely performed at the point of care. Practitioners may also lack the clinical expertise to evaluate studies outside their own specialty. Furthermore, many practitioners do not feel comfortable with their skills in evaluating the evidence in the literature (Williamson, 1989), although several projects to educate clinicians about these methods have had success (Bennett, 1987). Thus, evidence synthesis is not a task that full-time clinicians do often or well (Eddy, 1990). The task is properly one that is performed by people who have expertise in evidence synthesis. Even if high-quality, timely reviews were widely available, however, it is still questionable whether readers would translate the evidence into appropriate changes in practice (Lomas, 1991).

---

1. The proportion of clinical practice that is supported by scientific evidence is difficult to define and to measure. This estimate is by the Committee on Clinical Practice Guidelines of the Institute of Medicine (Field, 1992, p. 34).

Experts in evidence synthesis are familiar with **meta-analysis**, a statistical technique first developed in educational and social research, and now increasingly popular in medicine (Moher, 1995).[2] In the traditional qualitative review, an expert reads the literature and argues for a particular interpretation of the evidence based on experience and scholarship. The expert may, however, be biased in which articles she reviewed, or by a strong prior belief in a particular interpretation of the evidence. The gold standard for a review article is now the **systematic review**, in which an explicitly defined protocol is followed in identifying and retrieving studies for a comprehensive synthesis of the evidence (Chalmers, 1995). Meta-analysis is appropriate and feasible for a subset of systematic reviews. In a meta-analysis, synthesizers augment their experience and scholarship with statistical methods for combining the quantitative results of the studies.

1.2.2 Difficulties with Accomplishing the Four Core Tasks

Whether or not the systematic-review approach is used, evidence synthesizers must perform all four core tasks of evidence synthesis, explicitly or implicitly, with each review. Each of these core tasks poses logistical problems for evidence synthesizers, and even more problems for full-time clinicians.

1.2.2.1 Retrieval of Relevant Trials

Medline, the National Library of Medicine's electronic bibliographic index to 3700 medical journals, adds 31,000 new citations each month with daily updates, for a total of over 8.5 million records (National Library of Medicine, 1996). Although practitioners in one study were found to have two scientific questions on average at every clinic visit, fewer than one-third searched the literature electronically; the other two-thirds perceived the literature to be unmanageable (Covell, 1985). Electronic searching often misses relevant articles, while retrieving many irrelevant articles. Trial results may be reported in more than one article, thus leading to erroneous multiple counting of the evidence. Conversely, results of completed trials may remain unpublished, leading to erroneous

2. In the Medline records for 1996, 471 articles were coded with the publication type "meta-analysis."

omissions of evidence. People who do not have access to the full text of articles via electronic retrieval systems must physically retrieve the identified articles from a library, an undertaking that is not trivial.

1.2.2.2 Critique of the Trials

Clinical trials are scientific experiments. As such, the interpretation of clinical-trial results is dependent on a thorough understanding of the methods used in conducting the trial. Critiquing a trial can be decomposed into two tasks: (1) judging the internal validity of a trial, and (2) judging the generalizability of the trial. The **internal validity** of a trial is the extent to which the experiment was conducted such that its findings are likely to reflect the true state of the world, rather than reflecting experimental bias. The **generalizability** of a trial is the extent to which the findings, regardless of their validity, are applicable to the situation in which the evidence is to be applied.

The clinical literature does not routinely report sufficient information for proper critiquing of a clinical trial. To judge the internal validity of a trial properly, an evidence synthesizer needs more details about the conduct of a trial than are commonly reported in the literature. To judge the generalizability of a trial properly, an evidence synthesizer needs more details about the setting of a trial and its enrolled patients than are commonly reported in the literature.

1.2.2.3 Synthesis of Quantitative Results

The summary outcomes of a clinical trial are influenced by the play of chance, or sampling error. The more patients who are enrolled in a study, the less dominant is this sampling error. A major premise behind the statistical technique of meta-analysis is that many trials with few patients, and thus with low statistical precision, can be pooled to yield the equivalent of a large trial with higher statistical precision. The judgment on whether a given set of trials can be pooled legitimately is a complex one, and requires a careful exploration of the clinical and methodological differences among the trials. Again, these differences are neither routinely nor uniformly reported in the clinical literature. A further difficulty with combining the quantitative results from clinical trials is that the numbers

must be abstracted from the printed page and physically transferred to a computer system — a step that is tedious and is the source of many transcription errors.

1.2.2.4 Interpretation of Trials in Context

Every clinical trial is conducted on a background of prior research and in the context of the epidemiology and socioeconomic impact of the disease under investigation. Despite the need for evidence synthesizers to interpret clinical trials within their wider context, trials are published as though they were stand-alone pieces of evidence. Only several highly selected and possibly biased references (Wessely, 1997) are published with each trial report. Sometimes, the report is accompanied by an editorial — but often not by a systematic review of prior literature. To establish the proper interpretation context for a clinical trial, evidence synthesizers must perform many bibliographic searches and must even resort to word of mouth for uncovering relevant prior work.

Establishing the proper interpretation context is even more difficult for clinicians who are seeking to apply randomized-trial findings to the care of a particular patient. In the clinic, the context of care is paramount to proper application of the evidence. Is the patient clinically similar to the trial's patients? Is the reported treatment available locally? What would be the cost to the patient, and to the health plan? The clinician is in even more of a quandary if no systematic review exists to help the clinician place these trials in their scientific context. Until clinical scientific evidence is custom delivered for each particular decision-making context, front-line practitioners will continue to have difficulty applying scientific evidence at the point of care.

1.3 The Trial-Bank System

As long as scientific evidence is published as text, we will not be able to exploit fully the computer's power for managing information: Computers cannot and will not soon be able to read the clinical literature. In this age of digital publishing on the World Wide Web, and of the promise of smart computer systems, the benefits of publishing clinical scientific evidence into smart, Internet-accessible, electronic databases are obvious. What is less

obvious are the critical determinants for a practical and efficient publishing system that will systematically address the difficulties of using the clinical literature.

To start with, we should publish randomized clinical trials into structured databases, or trial banks, because randomized trials yield the highest-quality experimental evidence (Friedman, 1985), and because their highly regular structure eases their standardized representation in databases. In this new form of publishing, authors will themselves report their trials directly into trial banks completely and accurately, with the aid of authoring software. They will submit to journals their trial-bank entries in conjunction with the prose articles describing their trials.

Figure 1.1 is a schematic of the proposed trial-bank system. The hallmark of this system is that it is **interoperating**. Interoperating trial banks appear as one to a user. To retrieve trial information from an interoperating trial-bank system, a user submits only one query that is automatically routed and mapped simultaneously to all trial banks in the system, regardless of the physical location of the trial banks. The user does not need to know where the information about individual trials is stored; the information is shared across the system.

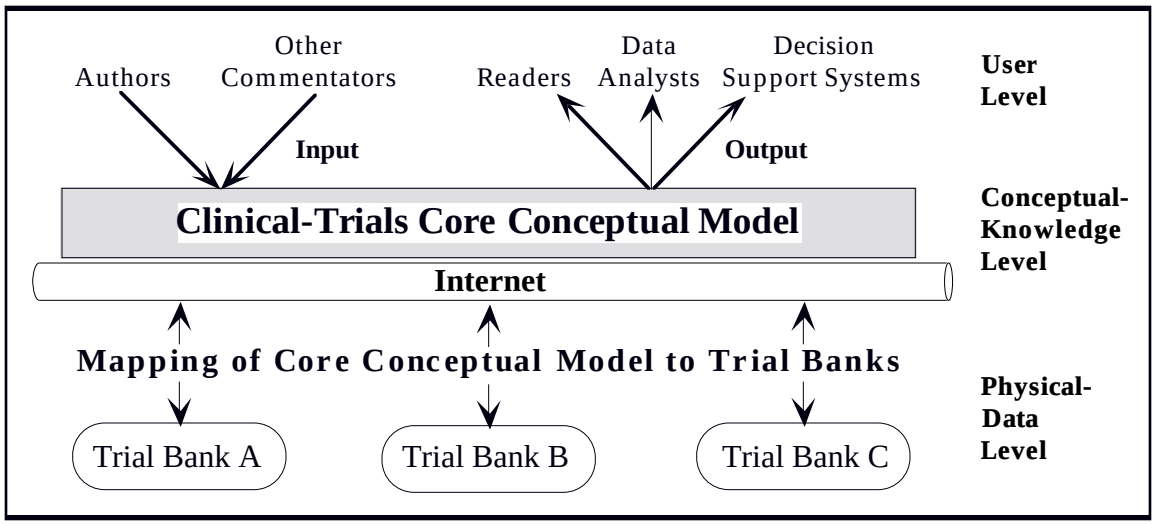

Figure 1.1. The interoperating trial-bank system. Multiple trial banks can be searched and analyzed as one by users. Both input and output from the trial banks are mediated by the core conceptual model of clinical trials — an abstract model that standardizes the communication of clinical-trial concepts within the trial-bank system. The Internet is the physical network for the trial-bank system.

To achieve interoperation, the trial-bank system should express the input and output of all trial banks in a shared **conceptual model** that abstractly details and represents the meaning of all clinical-trial concepts that can be shared. Because the shared conceptual model in the trial-bank system should support the core tasks of evidence synthesis, I call the shared model the clinical-trials **core conceptual model**. The design and implementation of this clinical-trials core conceptual model is one of the critical determinants of an ideal trial-bank system, and is the subject of my thesis research.

What will having a trial-bank system allow us to do that we cannot do now? The answer is diagrammed in Figure 1.2, which shows the trial-bank system as an integral component of

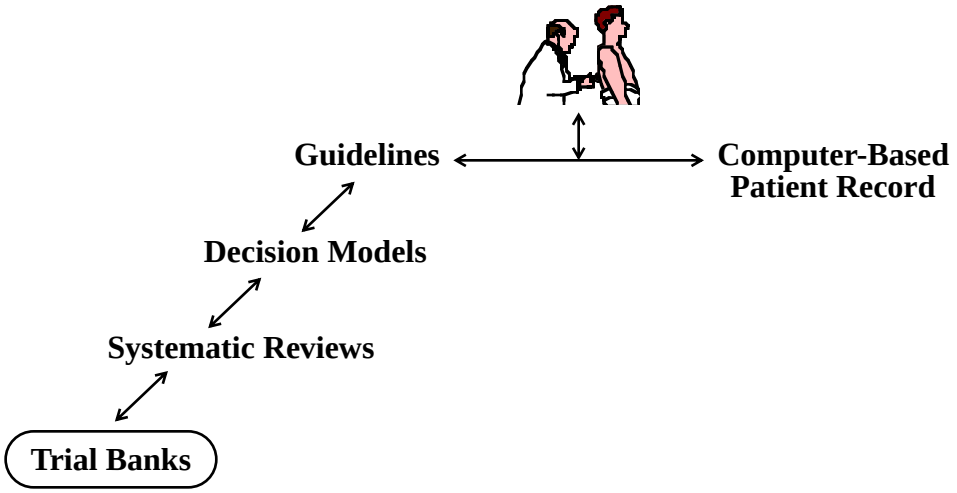

Figure 1.2. The trial-bank system in evidence-based care. Randomized trials anchor a chain of increasingly synthesized and processed evidence. This chain of evidence culminates in practice guidelines that should be customized to individual patients and settings before being applied to clinical care. Guidelines could be customized using the computer-based patient record.

an extensive information infrastructure for supporting point-of-care, evidence-based, clinical decision making. Evidence-based medicine in the era of the trial-bank system will be as follows:

1. Using the trial-bank system, evidence synthesizers systematically identify, retrieve, and combine evidence from related randomized trials worldwide. They perform quantitative meta-analyses when appropriate The system automatically links each trial to relevant information such as editorials, letters to the editor, and prior and subsequent related studies. Thus, the trial-bank system facilitates all the core tasks of evidence synthesis. The resulting systematic reviews are stored in structured databases as well, with hyperlinks to the appropriate trial-bank entries.
2. Decision analysts incorporate the evidence from systematic reviews into online interactive decision models that are closely integrated with online utility assessment tools.[3] Analysts or expert systems can use these interactive models to incorporate dynamically new data retrieved from trial banks or from structured databases of meta-analyses, and to perform sensitivity analyses with custom parameters.
3. Guideline developers are guided by the analytic framework of decision analysis to formulate practice guidelines that combine the totality of scientific evidence with wider policy considerations. The guidelines are also online, and are linked extensively to relevant policy-related information, as well as to the relevant trial-bank entries, systematic reviews, and decision models.
4. At the point of care, the practicing clinician starts the evidence-based decision-making process by using the computer-based patient record to identify relevant practice guidelines. A coordinated group of expert systems tailors the evidence base with parameters specific to the patient and the context of care: one system determines which randomized trials are applicable based on patient-record data; another system recalculates meta-analytic summary point estimates of effect using this restricted set of trials; yet another expert system uses online cost

3. **Decision analysis** is a methodology for decision-making according to the axioms of rational thought (Raiffa, 1968). A health **utility** is a number between 0 and 1 that reflects a subject's preference for a particular health state, where 0 is equivalent to death and 1 to perfect health.

> databases and utility-assessment tools to tailor decision-model parameters such as costs and patient utilities. A final expert system tailors the practice recommendation to the clinician. Armed with this recommendation and an audit trail of the supporting evidence, the clinician decides upon an action with full knowledge of the state of the evidence supporting that action. At any time, the clinician can further explore online the entire evidence base for the decision.

Granted, there are many methodological and technical hurdles to realizing this vision of computer-supported evidence-based medicine. The torrent of clinical evidence being published nevertheless leaves us no option but to harness the information-management power of the computer if we are ever to practice evidence-based medicine as it was originally conceived. The trial-bank system is the seed for a shared information infrastructure that will help practitioners to surmount the daunting information-management challenges of evidence-based clinical decision making.

1.4 Design Specification for the Core Conceptual Model

The implementation of the proposed trial-bank system — not to mention its extension to decision models and to practice guidelines — involves large-scale engineering and a fundamental change in the reporting of clinical trials. The scope of my thesis research is far more limited. My first specific aim was to define and evaluate a blueprint for a conceptual model of randomized trials that allows trial banks to share all the concepts necessary for accomplishing the four core tasks of evidence synthesis. To achieve this specific aim, I devised a new method called **competency decomposition**.

1.4.1 Specification Using Competency Decomposition

Conceptual models are finite in size, and a conceptual modeler must therefore choose which domain concepts to include in a model, and, just as important, which not to include. These choices are never inherently right or wrong; it is not more correct in any absolute

sense for a conceptual model of clinical trials to include the concept of cost outcomes than not to include it. Rather, the appropriateness of the modeling choices must be judged with respect to the tasks that the model is intended to support. Thus, a proper design specification for a conceptual model of clinical trials must itemize the tasks that the model is to support, and must itemize and justify the clinical-trial concepts needed to accomplish those tasks.

Table 1.1 shows a partial design specification for the core conceptual model of clinical trials. The target task — called a competency for reasons explained later — is quantitative

Table 1.1 Quantitative synthesis competency decomposition.

Competency	Method	Method-Associated Subcompetency	Data Requirement of Clinical-Trials Model
I. Calculate summary statistic, for pairwise comparisons	A. OR[a]	1. Calculate OR	a. Complete 2 X 2 contingency table
II. Quantitative meta-analysis	A. Mantel–Haenszel, using OR	1. Calculate OR for each trial	a. Same as I.A.1-2.a
		2. Calculate meta-analytic summary	a. ORs for all the trials

a. odds ratio

meta-analysis using the Mantel–Haenszel method, with odds ratios as the summary statistic. This method of meta-analysis requires that we have the odds ratios for each of the trials that we are combining (data requirement II.A.2.a), and that implies that we must have a complete 2 X 2 contingency table for the outcome that we are meta-analyzing for each trial (data requirement II.A.1.a). We can use this framework to specify the data requirements for tasks accomplished with more than one method, and we can even extend the framework to specify the procedural knowledge required for a method (e.g., the Mantel–Haenszel formula for subcompetency II.A.2). Chapter 5 details the full design specification for the core conceptual model of clinical trials. I based the decomposition of

the competencies on a thorough review of the clinical-trials interpretation literature, and on my own experience meta-analyzing randomized trials. Each of the four core tasks of evidence synthesis is decomposed separately. The design specification for the high-level task of trial critiquing consists of 12 subcompetencies, 42 subsubcompetencies, and 145 data requirements.

The intellectual heritage of this competency-decomposition approach comes from the **task-decomposition** approach of Chandrasekeran and colleagues (Chandrasekaran, 1993) and of other researchers, and from the **competency-questions** approach of Gruninger and colleagues (Gruninger, 1995). From the task-decomposition approach, I borrowed the idea of decomposing tasks hierarchically into subtasks and into methods.[4] From the competency-questions approach, I borrowed the ideas of designating a target task of a conceptual model as a *competency* of that model, of indexing the data requirements of that model to its competencies, and finally, of using this framework to evaluate the actual competencies of conceptual models.

1.4.2 Evaluation of the Design Specification

The full design specification is given in Appendix A. The bulk of the data requirements for the clinical-trials core conceptual model arises from the competency decomposition of the trial-critiquing task. To demonstrate that this competency decomposition is reasonable, I compared its data requirements to those of 18 published trial-critiquing questionnaires that reflect the state of the art in trial critiquing. Because there is no gold-standard method for critiquing a randomized trial, these 18 instruments “differ from one another in almost every respect” (Moher, 1995), and they run the gamut in what trial information they require. Overall, 95 percent of the data required by the trial-critiquing instruments are also required by my trial-critiquing competency decomposition. The reasons why the remaining 5 percent are not required by the competency decomposition are judgment calls that are explained in Chapter 7. Conversely, 74 percent of the data requirements of the trial-critiquing competency decomposition were also required by one or more of the 18

4. **Methods** are actions that lead to the accomplishment of a task.

trial-critiquing instruments. The 26-percent mismatch occurs because my competency decomposition is expressly designed to support all reasonable evidence-synthesis tasks, including tasks that are in the trials-interpretation literature but are not in the 18 instruments.

The competency decompositions of the remaining core tasks of evidence synthesis — information retrieval, quantitative computation, and the interpretation of trials in context — are much smaller than the one for trial critiquing, and their reasonableness is evident at face value (Chapter 7 and Appendix A). In addition to showing that my design specification for a core conceptual model of clinical trials is reasonable, I also show through argument that the design specification is extensible to new tasks and to new methods (Chapter 5).

1.5 Ocelot-CCM Core Conceptual Model

To construct a concrete conceptual model, we must encode the concepts required by the design specification in a knowledge-representation language. A natural language such as English is a candidate knowledge-representation language, but computers cannot, and will not soon be able to, read unrestricted natural language. Of the computer-understandable knowledge-representation languages, the classic one — first-order logic — is the most expressive, but it is difficult to build and maintain a knowledge base with it. Furthermore, formal logic is not commonly understood by those people who will likely be building the trial-bank system. The relational data-definition language is commonly used for structured databases, but its expressivity falls short of what is ideal for a core conceptual model for interoperating trial banks. The most appropriate class of languages for encoding a core conceptual model of clinical trials is the object data-definition languages. I have chosen to use the object-based Ocelot language,[5] which is sufficiently expressive for the trial-bank tasks, yet is compact, concise, and understandable by many people.

---

5. In this dissertation, the database term *object based* and the knowledge-engineering term *frame based* are synonymous. Ocelot is commonly known as a frame-based language.

Ocelot-CCM is the clinical-trials core conceptual model that I built according to the design specifications introduced in Section 1.4. Ocelot-CCM is a class hierarchy consisting of 128 frames (or objects) with 430 unique slots (or attributes), of which 27 percent take another frame as an instance. Thus, Ocelot-CCM is a small but rich conceptual model.[6]

In the fully implemented trial-bank system, a controlled medical vocabulary is essential for interoperating the clinical content of the trials. For example, trial banks must standardize on one of the terms H2-BLOCKER, H2-ANTAGONIST, or ANTI-HISTAMINE if they are to share information on drug types. In this dissertation, neither the names of the concepts in Ocelot-CCM nor the terms used to instantiate[7] the concepts belong to a controlled medical vocabulary. Incorporation of a controlled vocabulary is a high priority for future work.

1.5.1 Expressivity Characteristics

Like other object-based conceptual models, Ocelot-CCM cannot represent nonmonotonicity (e.g., that a person whom we thought was dead is actually still alive), uncertainty (e.g., that we are not sure whether or not a person is dead), negation (e.g., that the negation of life is death), and disjunction (e.g., that a person is either alive or dead, but cannot be both). Ocelot-CCM expresses logical rules, but expresses only simple temporal relationships. The implications of these expressivity characteristics are that Ocelot-CCM can represent neither crossover nor Bayesian trial designs, nor can it support trial simulation.

1.5.2 Structure and Content

The objects (or frames) in Ocelot-CCM can be partitioned into the following trial-feature groups: administration, statistical design, publications, subjects and recruitment, treatment assignment, intervention, follow-up, outcomes definition and measurement, and results.

---

6. Ocelot-CCM can also be considered to be a *data schema*, an *ontology*, or a *class definition*.

7. An *instance* of a concept is a particular example of that concept. For example, CIMETIDINE is an instance of the generic concept H2-BLOCKER, and CIMETIDINE instantiates H2-BLOCKER.

These objects allow a broad range of trials to be captured in Ocelot-CCM, from two-armed randomized trials, to cohort studies, to trials with run-in and washout periods. Trial interventions can be drugs, procedures, or behavioral counselling; trial outcomes can be dichotomous or continuous; and the analysis can be traditional statistical tests or regressions.

1.5.3 Evaluation of Ocelot-CCM

The evaluation of Ocelot-CCM uses the design specification (Appendix A) as the yardstick for determining the tasks that Ocelot-CCM can support, and for determining Ocelot-CCM's conceptual coverage. For each of the data requirements in the trial-critiquing competency decomposition, examples of the data required — called **criterion instances** — were collected from published trial reports, and from the trial design and execution records of the VA Cooperative Studies Center's Stroke Prevention in Nonrheumatic Atrial Fibrillation (SPINAF) trial. When I attempted to enter these 152 criterion instances into Ocelot-CCM, Ocelot-CCM successfully captured 93 percent of them. This success demonstrates that Ocelot-CCM contains the clinical-trial information necessary and sufficient for accomplishing 56 out of 62 (90 percent) of the lowest-level competencies specified in the design specification. The competencies that were not supported can be supported with only minor revisions to the model. Ocelot-CCM's conceptual coverage of the domain of clinical trials is summarized in Table 1.2 (page 19).

The evaluation of Ocelot-CCM's conceptual coverage involved the instantiation of two complete randomized trials (Ezekowitz, 1992; Singh, 1995). Using the generic knowledge-base editing tool GKB-Editor (Karp, 1995), I took approximately 10 hours to enter the SPINAF trial into Ocelot-CCM directly from SPINAF's design and execution records. This preliminary experience suggests that direct authoring of trials into trial banks will require an amount of time and work that will be neither trivial nor prohibitive for trial investigators.

Feature Dimension	Range
Design	Randomized trials with more than two treatment arms; nested randomization; run-in or washout periods; factorial trials; prospective cohort studies
Subjects	Patient, MD, etc.; depends on clinical vocabulary
Intervention	Drugs (fixed or stepped dosages, or titrated to effect); surgical and radiological procedures; medical devices; behavioral change interventions.
Endpoint Type	Clinical (e.g., laboratory results); death
Data-Aggregation Level	Summary or individual patient level
Result Type	Dichotomous; continuous; ordinal; categorical; proportions; parametric and nonparametric summaries; comparative statistics
Statistical Method	Contingency tables; t–test; Kaplan–Meier; regression; others

Table 1.2 Summary of the clinical conceptual coverage of Ocelot-CCM. Ocelot-CCM can capture all of these trial features.

1.6 RCT Presenter

The construction of an entire, interoperating trial-bank system is beyond the scope of my thesis work. I have, however, built a single trial bank that can be browsed over the web. RCT Presenter is a proof of concept artefact, and its empiric evaluation yields findings that complement the more abstract evaluations of the design specification and of Ocelot-CCM.

1.6.1 Architecture

RCT Presenter consists of two components: (1) **RCT Bank**, a structured database built in the Ocelot knowledge-representation system and whose conceptual model is Ocelot-CCM; and (2) a web site programmed in Lisp and running on the CL-HTTP web server (Mallery, 1997) that responds to RCT Bank queries with dynamically generated web pages.

RCT Bank contains complete descriptions of two randomized trials — the CHF-STAT trial (Singh, 1995) and the SPINAF trial (Ezekowitz, 1992) — and partial descriptions of five others. These are the same trials used in the evaluation of Ocelot-CCM's conceptual cover-

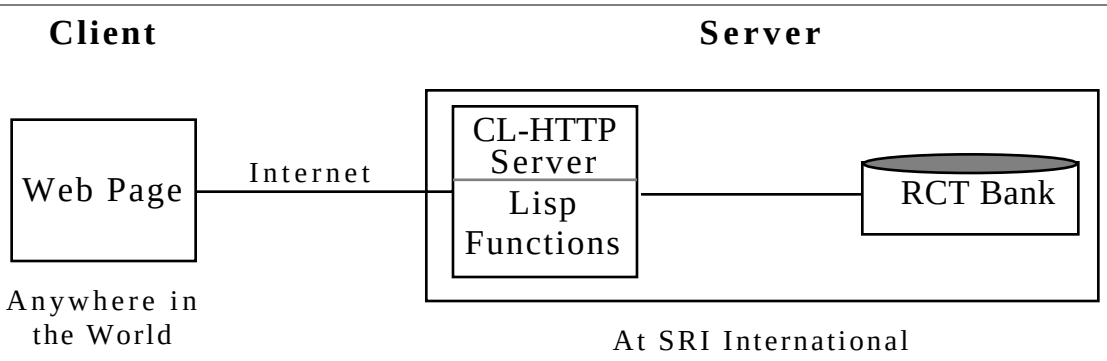

Figure 1.3. RCT Presenter system architecture. The RCT Presenter system follows the client–server model.

age (Chapter 7). The interface allows users to select a trial to browse, to browse that trial in hyperlinked or in linear fashion, and to generate custom tables of up to four attributes across multiple trials. Trial critiquing is expressly supported with the availability of two trial-critiquing questionnaires online (Detsky, 1992; Begg, 1996), with each item of the questionnaires hyperlinked automatically to the appropriate trial information. Users can thus retrieve, with a simple click on a questionnaire item, exactly the information that they need to critique a trial.

1.6.2 Pilot Evaluation of RCT Presenter

A convenience sample of 11 health-services research fellows and faculty used an early version of RCT Presenter to rate the quality of the CHF-STAT trial (Singh, 1995) using the Detsky instrument (Detsky, 1992). All the subjects were at least partially familiar with the critical appraisal of randomized trials. They required 14 minutes on average to complete the 15-item questionnaire, which required information on such trial attributes as what the exclusion criteria were, and whether or not the outcome assessors were blinded to the treatment the patient received. Of the 12 items with a definitive answer, the subjects answered 10 of them correctly more than 80 percent of the time; problems with the interface were responsible for some of the subjects' difficulties in answering the questionnaire. The subjects rated the system's ease of use, usefulness of content, and format of presentation highly (average 4.4 out of 5, where 5 is ideal), and they agreed that publishing randomized trials into trial banks would be a good idea. However, several

subjects voiced a deep mistrust of information on computers. They wished for reassurance of the veracity and quality of the information either through validation by a trusted journal, or through a declaration that the trial descriptions were entered by the authors themselves. The implications of this pilot evaluation for trial-bank publishing are several. First, the core conceptual model contains sufficient information to allow health-services researchers to complete a short but representative trial-critiquing questionnaire from the literature. Second, different users should be catered to with different browsing interfaces. Third, due consideration must be given to mechanisms to assure readers of the accuracy, fairness, and quality of trial-bank entries.

1.7 Guide for the Reader

The remainder of this dissertation elaborates on the ideas and work presented in this chapter. In Chapter 2, I discuss the specific problems of using the clinical literature in medicine, and I review current approaches to tackling these problems. In Chapter 3, I present a detailed description of the proposed trial-bank system, including a discussion of various legal, social, and economic considerations. Chapters 4 and 5 focus, respectively, on the design and construction of conceptual models in general, and of Ocelot-CCM in particular. In Chapter 6, I describe the architecture and implementation of RCT Presenter, as well as its pilot evaluation. Chapter 7 presents the evaluations and implications of this work. I conclude in Chapter 8 with a summary of my contributions, and with a discussion of the prospects for deploying a large-scale trial-bank system.

Chapter 2

The Clinical Literature

Large, randomized, clinical trials are one of our most valuable sources of scientific knowledge. The clinical literature is the main channel for disseminating the results of these important studies, yet the results, obtained at great expense, must often all be communicated in a single, paper-based, text article of 4000 words or less. Within these tight form and length constraints, authors and editors attempt to satisfy the varied needs of the users of the literature — users who are basic scientists, practicing clinicians, educators, and evidence synthesizers. The authors and editors do not succeed. In its attempt to serve too many audiences at the same time with short, single-format, text-based articles, the clinical literature is too often wrong for everyone. The result is an inefficient transfer of evidence from the research world to the clinic, and a waste of precious resources.

In this chapter, I discuss the difficulties with using reports of randomized trials, and I discuss the piecemeal nature of the current approaches to these difficulties. Current approaches all address the phases of a trial's life-cycle in isolation: its design, registration, standardization, publication, synthesis, or application to particular patients. If we are to have a comprehensive strategy for transferring randomized-trial evidence from the literature to the clinic, we need an approach that spans a trial's entire life-cycle; we need an approach that is trial-centered, rather than article-centered.

2.1 Central Role of the Clinical Literature in Evidence-Based Medicine

The clinical literature is the primary venue for reporting new medical evidence. The size of the literature is growing inexorably; in 1993, the National Library of Medicine catalogued over 22,000 active medical serials, of which they considered 16,000 to be journals (National Library of Medicine, 1993). Both industry and academia seek the imprimatur of quality that comes from publishing in the peer-reviewed literature. Yet this vast repository of knowledge is "underused" (Huth, 1989) and has "loose connections" with clinical practice (Haynes, 1990).

In this chapter, I argue that the clinical literature is underused because, in its present form, it fails to satisfy the needs of any of its intended users. In Chapter 3, I discuss how the trial-bank system that I propose can more effectively address the needs of the users of the clinical literature.

2.1.1 High-Quality Evidence of Randomized Trials

The clinical literature comprises many kinds of articles. Some articles report on basic-science experiments; some are tutorials on the management of clinical conditions; still others present original data from clinical research, or review previously published research. Each kind of article is intended for a different audience; thus, we must tease apart and discuss separately the problems encountered by users of each kind of article.

I focus on the problems of using a small subset of the clinical literature: the reports of randomized clinical trials.[1] I do so because of technical reasons, which I discuss in Chapters 4 and 5, and because randomized trials are less subject to confounding factors than most other study designs. In randomized trials, patients are assigned randomly to receive either an experimental treatment or a control treatment. Any differences in the final outcomes of

1. About 1 percent of articles published each year are reports of randomized trials (Meinert, 1984).

the two groups can, in theory, be attributed solely to the treatment received, because confounding factors are randomly and thus equally distributed between the groups. Other study designs, such as the case-control study or the case series, are subject to many types of confounding that often complicate the interpretation of the results. A major criticism of the randomized trial is that their results are of limited generalizability, because neither the subjects of, nor the care during, a randomized trial are representative of everyday practice. Nevertheless, randomized trials can be expected to yield the most internally valid findings, and they are considered to be one of the most reliable sources of clinical scientific evidence (Friedman, 1985). Randomized trials are accorded a Level I quality of evidence — the highest — for the support of guidelines issued by the Canadian Task Force on the Periodic Health Exam, and by the United States Public Health Services Task Force.

However, many clinical questions of vital importance are not amenable to being tested with a randomized trial. For example, it would be neither ethical nor practical to randomize school-aged children to tobacco smoking to determine whether lung-cancer rates are higher in long-term smokers than in nonsmokers. In such cases, the methodologically weaker observational studies have an important role (Black, 1996).

2.1.2 Problems with Randomized-Trial Reporting

Randomized-trial reports are a small but highly influential subset of the clinical literature. In this section, I discuss the particular problems that attend the use of these reports. Several of these problems — inaccurate retrieval, and publications that are not accurate, clear, or timely — are common to the use of all of the clinical literature. Other problems of particular relevance to the use of randomized-trial reports include incomplete and nonstandard reporting, and flaws in study methodology.

2.1.2.1 Problems with Study Retrieval

One of the first difficulties confronting a user of the clinical literature is finding the relevant studies. There are two common measures of the accuracy of electronic bibliographic searching. **Recall** is the percentage of all the relevant documents in a document collection

(e.g., Medline) that a search retrieves. The concept is analogous to the sensitivity of a test, but there is no objective gold standard for whether a particular document is relevant to a particular query. **Precision** is the percentage of documents retrieved by a search that are relevant to the search query. This concept is analogous to the positive predictive value of a test. **Retrieval performance** is a function of how accurately a query is expressed, how richly the information in the document is represented, and how well the query can be matched with the document representation.

Numerous studies have documented the poor retrieval performance of Medline, the most widely used electronic index to the clinical literature. Dickersin reviewed studies on Medline searching, and found that recall was 51 percent overall ((Chalmers, 1995), p 18). Many irrelevant articles were retrieved: The mean precision of the searches was 9 percent, with a median of 33 percent and a range of 2 to 82 percent. In all cases, the gold standard was hand searching of relevant journals. More advanced Medline search strategies that use free-text words or word truncation may achieve better precision and recall than these searches.

One explanation for poor retrieval with most keyword-based electronic searching is that many common queries simply cannot be expressed. In Melvyl Medline, a search interface to Medline, we cannot express the query “retrieve trials where mortality was a primary outcome.” Another explanation for poor retrieval with index-based bibliographies is that the information content of the documents is neither richly nor accurately captured in the index terms. For example, because the primary outcome of a trial is not separately indexed in Medline, a Medline citation cannot state that a trial has mortality as its primary outcome. Also, the interrater reliability for assigning major subject headings to Medline records is only 61 percent, and the reliability for assigning major and minor subject headings combined is only 34 percent (Funk, 1983). Thus, neither queries nor documents are captured accurately in most Medline searching.

A more difficult problem in study retrieval is that foreign-language documents and the gray — or shadow **—** literature are difficult to identify and retrieve. The **gray literature** includes conference proceedings, dissertations, technical reports, the unpublished

literature, and commercial publications. Because these publications are often not included in electronic bibliographies, they are difficult to identify. Furthermore, many of these publications are not routinely available at a local library. Although this gray literature is not peer reviewed and is most likely of lower average quality, evidence synthesizers who systematically exclude this literature are introducing bias into their reviews. We need more research on whether or not such biases are consequential.

Foreign-language articles are often routinely excluded from evidence syntheses because of the trouble and expense of translating them, even though their reporting standards and quality are comparable to those of the English literature (Moher, 1996). Systematic exclusion of these sources of evidence can lead to biased evidence syntheses (Gregoire, 1995). Any proposal for improving information management for evidence-based medicine should also address these problems of inaccurate and incomplete study retrieval.

2.1.2.2 Problems with Completeness

The definition of what constitutes complete reporting depends on who will be using the report. Different users require different information. Practicing clinicians may want to know only whether and how a trial's findings are applicable to their daily work. Evidence synthesizers, on the other hand, often need detailed and specific information for judging a trial's internal validity and generalizability. I surmise that the trial-information needs of an evidence synthesizer is a superset of the trial information needs of the practicing clinician. Thus, I discuss the completeness of randomized-trial reporting with respect to the information needs of evidence synthesizers only.

Completeness for trial critiquing — Often, much of the information necessary for trial critiquing is not reported, such that it is difficult to judge whether or not a randomized trial is internally valid or generalizable. DerSimonian found widespread deficiencies in reporting such information in top-flight journals (DerSimonian, 1982). More recently, Schulz found that only 9 percent of surveyed reports described sequence-generation and allocation-concealment procedures (they are needed for judgments of internal validity) (Schulz, 1994), and Moher reports that only 32 percent of randomized trials with negative results

reported a sample-size calculation (Moher, 1994). Pocock found serious statistical-reporting problems in three major journals (Pocock, 1987), one of which has since instituted routine statistical review of submitted papers (Gore, 1992). These and many other commentators call consistently and insistently for vastly improved reporting of the methods and statistics of clinical studies, and of randomized trials in particular. However, few of the commentators acknowledge that only evidence synthesizers are likely to be interested in these details, whereas the bulk of a journal's readers are more likely to be practicing clinicians who consider these details irrelevant to their needs. If studies continue to be published as a one-size-fits-all text article, "combining accurate, complete reporting with easy readability" will most certainly continue to be a challenge (Rennie, 1994).

Completeness for evidence synthesis — The evidence-synthesis process involves retrieving relevant trials, critiquing each trial for its internal validity and generalizability, and then combining the quantitative trial results using a statistical method called *meta-analysis*, when appropriate. The detail of trial information needed for meta-analysis exceeds even that which is needed for trial critiquing.

Meta-analysis of clinical studies is evolving from a strictly quantitative aggregation of trial results to a study of studies. The objective of this newer approach is to generate insight and new hypotheses by exploring the influence of clinical, methodological, and statistical heterogeneity among the trials on their observed outcomes (Thompson, 1994). For example, I have found that the observed benefit of amiodarone for preventing sudden cardiac death correlates with the kind of control used in the randomized trial (placebo versus usual-care controls), but not with the kind of heart disease that the patients had (Sim, 1997). Trials with non-placebo controls had systematically larger observed effects, probably as a result of biases arising from post-randomization treatment differences. Based on these findings, we discounted the results of trials using non-placebo controls in our evidence synthesis, and this led to clinical implications different from those we found by synthesizing the results of all the trials.

Thus, meta-analyses can be misleading if critical sources of heterogeneity are not explored. Another hurdle to performing a good meta-analysis is that meta-analyzing aggregate, summary-level data may yield results different from the gold standard of meta-analyzing patient-level data (Stewart, 1993; Jeng, 1995). To perform high-quality meta-analyses, meta-analysts need trial reports that have clear and thorough descriptions of clinical, methodological, and statistical information — preferably with data at the individual patient level.

Completeness for avoiding publication bias — For the purposes of evidence synthesis, completeness of reporting also involves the complete cataloging of all planned and completed trials. The evidence from completed trials is not always published, thus creating the **file-drawer problem**. Scherer found that only 51 percent of randomized trials presented as abstracts at national opthamology meetings were subsequently published in full, with the larger and positive studies being published preferentially (Scherer, 1994). Somewhat surprisingly, trials with negative results remain unpublished because the trial investigators submit them less frequently for publication, rather than because editors reject them (Dickersin, 1992). This finding implies that at least part of the solution to the publication-bias problem must be targeted to trial investigators (Section 2.2.3).

Evidence syntheses that review only published data can be biased as a result of this tendency to publish preferentially trials with positive results. For example, Simes performed two meta-analyses on whether or not combination chemotherapy improves survival in patients who have advanced ovarian cancer (Simes, 1986). When only published trial results were used in one meta-analysis, combination chemotherapy appeared to offer significant benefit; in the meta-analysis using both published and unpublished results, in contrast, no benefit was found. Based on the expense and risk to patients that randomized trials entail, commentators are increasingly considering the problem of publication bias as one of scientific misconduct and human-rights violation (Chalmers, 1990).

On the flip side of the file-drawer problem is the problem of **redundant publication**. When authors publish the results of the same trial more than once, their findings can be double-counted in evidence synthesis, leading to an incorrect synthesis (Tramer, 1997). For the literature as a whole, duplicate publications "overburden busy reviewers, fill the medical literature with inconsequential material, and distort the academic reward system" (Kassirer, 1995). At present, the only mechanism for avoiding redundant publication is trust in the integrity of authors.

2.1.2.3 Problems with Accuracy and Clarity

Accurate trial information correctly reflects the true state of the trial and its findings. Unfortunately, not all information in the literature is accurate (e.g., there are incorrect citations of trial results in some meta-analyses (Teo, 1993)). **Fraudulent information** is another form of inaccurate information that the literature identifies and culls poorly. In an exploration of a famous case of fraud, Friedman found that only 15 of 60 fraudulent articles were retracted, and only seven of these were indexed under the heading "Retraction of Publication" in Medline (Friedman, 1990). In another study, retracted papers were cited only about 35 percent less frequently than a comparable group of papers that had not been retracted (Pfeifer, 1990).

Problems with **clarity of reporting** are close cousins to problems with accuracy. Dickersin found that 26 percent of the opthamology trials she reviewed were unclear about whether or not treatment allocation was randomized. In 40 percent of these trials, the authors clarified, when tracked down and asked, that the allocation was indeed randomized (Chalmers, 1995, p. 28). Another form of ambiguous reporting occurs when trial information is internally inconsistent (i.e., when the reported data contradict themselves). Internally inconsistent data are often published despite editorial review. Two design papers for the Arrhythmics versus Implantable Defibrillators (AVID) trial gave conflicting descriptions of an exclusion criterion (5 days after acute myocardial infarction in the American Journal of Cardiology article (AVID Investigators, 1995), 7 days in the American Heart Journal article (Greene, 1994)). Because of the prevalence of ambiguous reporting, the standard of practice in meta-analysis is for at least two independent reviewers to

abstract information from each trial report. Disagreements are then usually resolved by consensus with a third reviewer.

2.1.2.4 Problems with the Quality of Study Methodology

There is no gold standard for what constitutes a well-designed and well-executed study (Greenland, 1994). Nonetheless, biostatisticians and other commentators have long been concerned about the quality of the design, execution, and analysis of clinical studies (Fletcher, 1979; Hemminki, 1982; Williamson, 1986; Emerson, 1990; Altman, 1994). Glantz found, in 1980, that about one-half of the articles that used statistical methods used them incorrectly (Glantz, 1980). Schulz documented statistical evidence that treatment assignment in putatively randomized trials was subverted (Schulz, 1994), and that this subversion could lead to an exaggeration in observed outcomes (Schulz, 1995).

There also exists no consensus on a methodology for assessing the quality of a trial. Compounding the confusion is the problem that many quality-scoring instruments do not distinguish between the quality of a trial's reporting and the quality of the trial itself (Moher, 1995). In a comparison of six scales used to assess the quality of 16 randomized trials, where scores were normalized to a maximum of 100, Walsh found that the pairwise differences in scores ranged from 13 to 73 points, with a mean difference of 44 points (Walsh, 1994). However, there is some agreement on the relative importance of the scored criteria: Detsky's shorter scale gave the same quality rank-ordering of trials as Chalmer's classic scale (Chalmers, 1981; Detsky, 1992). Overall, the ratings of published trials usually hover in the mid-range of scales. Ratings of trial quality over time are difficult to interpret, because of changing trial-reporting standards and changing definitions of trial quality.

2.1.2.5 Problems with Standardization

Problems with the standardization of randomized-trial reporting are of two types. One is the lack of a standard vocabulary for indexing reports, so that *mortality*, for example, is sometimes indexed as *death*. Medline uses the Medical Subject Headings (MeSH) for indexing, but, as I discussed on page 26, these headings are not assigned reliably, and we still lack true standardization of indexing terms.

The other type of standardization problem concerns both trial design and trial reporting. Combining evidence from many studies is easiest when all the studies measure and report the same outcomes — for example, 1-year mortality in patients who are given aspirin chronically after an acute heart attack. There exists, however, no mechanism for authors to coordinate the outcomes that they plan to measure. Incomplete trial reporting further contributes to the standardization problem when reports do not describe all data that were collected during the trial. Evidence synthesizers are often left wondering whether an outcome was measured but not reported, or was not measured at all.

2.1.2.6 Problems with Timeliness

Reports on new advances in the treatment and diagnosis of illness are now a staple of the lay press. Frequently, the evidence is not formally published in the clinical literature until weeks or even months after popular dissemination (Steinbrook, 1990), because of delays attributable to peer review and to paper-based publication. Physicians are left to practice medicine by press release. The clinical literature often fails to deliver timely, well-reported evidence in precisely those situations that interest the media, and the public, the most.

2.1.2.7 Influence on Usability

All the problems discussed thus far affect directly the usability of randomized trial evidence. I discuss separately its usability for practicing clinicians and for evidence synthesizers.

Usability for practicing clinicians — Practitioners consider much of the clinical literature — including randomized-trial reports — difficult to use, and irrelevant to their work (Justice, 1994): electronic searching returns many irrelevant articles; it takes time to retrieve articles from the library; it takes time and skill to read reports of original studies; and the evidence cannot be translated easily into clinical action. The consequence is that advances in clinical science diffuse poorly to the very people who should be putting them into practice. In one study, about 50 percent of physicians surveyed were unaware of at least one significant advance in the literature that was relevant to their practice (Williamson, 1989).

One of the most efficient ways for practitioners to keep abreast of research is to read evidence syntheses rather than reports of original research (Section 1.1). Before the advent of systematic reviews, review articles were usually idiosyncratic discussions of the literature, and were often based as much on personal opinion as on evidence. Antman showed that "review articles often failed to mention important advances or exhibited delays in recommending effective preventive measures. In some cases, treatments that have no effect on mortality or are potentially harmful continued to be recommended by several clinical experts" (Antman, 1992). Antman concluded that the clinical literature must serve the needs of evidence synthesizers better if we are to have more accurate reviews.

Usability for evidence synthesizers — Like practitioners, evidence synthesizers also consider the clinical literature difficult to use. The Cochrane Collaboration, an international group of meta-analysts devoted to systematic reviews of the effects of health care (Chalmers, 1994), has had to develop policies for addressing the problems of the proper retrieval of relevant trials, the reliable abstraction of trial information, the quality scoring of trials, the standard indexing of trials, and effective methods for contacting trial authors for information and clarification (Sackett, 1996). The tremendous time and work required to complete an evidence synthesis belies the usability of the literature for evidence synthesis.

2.1.3 Implications for Getting the Evidence to the Clinic

The transfer of evidence from the research world to the clinic is inefficient and, in large part, ineffective. None of the myriad audiences of the clinical literature — basic scientists, practicing clinicians, educators, and evidence synthesizers — are satisfied. In particular, the literature fails to provide practitioners with the sound, relevant, and timely information that they need to practice evidence-based medicine.

2.2 Current Approaches to Improving the Clinical Literature

These problems with using the clinical literature are well known and well documented, and approaches abound for rectifying them. In this section, I group these approaches into eight general strategies, and I discuss notable examples of these strategies. I then argue that these approaches, albeit worthwhile, are piecemeal. In Chapter 3, I propose and describe a comprehensive solution: the trial-bank system.

2.2.1 Approaches to Improve Trial Retrieval

The problem of retrieving relevant information accurately from a large information pool is a general one. The popularity of the World Wide Web has spawned an information-retrieval problem in the large that researchers in the information sciences, library science, and artificial intelligence are working actively to solve. Information retrieval in medicine is a popular research area. Of note, much of this research assumes that the content, form, and medium of the clinical literature are givens. In the approaches that I discuss here, only the final one (Section 2.2.1.4) involves changing the clinical literature itself.

2.2.1.1 Postpublication Processing

One strategy to improve retrieval performance is to improve the semantic richness of document representation (see Section 2.1.2.1). The approaches that I discuss here — improved indexing and context markup — provide more extensive or more accurate descriptions of the content of articles. These approaches are implemented after an article has already been written and published. Therefore, they cannot correct — and indeed are limited by — the literature's shortcomings in completeness, accuracy, clarity, and the quality of study methodology. Many postpublication approaches are also labor intensive.

Improved indexing — The inconsistency of Medline indexing is problematic (Section 2.1.2.1). The National Library of Medicine's (NLM's) MedIndEx expert system was designed to help humans to index Medline bibliographic citations with greater

consistency, but its success has not yet been demonstrated (Humphrey, 1992). Automated, concept-based indexing of bibliographic citations is another computer-based approach. It too has failed so far to yield retrieval more accurate than that yielded by indexing with standard MeSH terms (Hersh, 1993).

One of the most successful approaches to improving indexing is the Medline Retagging Project of the Baltimore Cochrane Center. Through electronic and hand searching of journals, Cochrane Collaboration members worldwide identify randomized or controlled trials. The Baltimore Cochrane Center relays this information to Medline, which then adds the MeSH heading “randomized controlled trial” or “controlled clinical trial” to the citations. From 1995 to October of 1997, this project submitted 27,612 citations to NLM for retagging as “randomized controlled trial” (Center, 1997). This work has improved the recall of randomized trials (Johnson, 1995), and that is important for systematic reviews of the evidence.

Context markup — To improve the precision of retrieval, the context markup approach overlays a structure of **contexts** onto articles in the clinical literature (Purcell, 1996). For a randomized-trial report, for example, sentences would be marked as belonging to contexts such as “Background,” “Exclusion/Withdrawal,” and “Experimental Findings.” In her thesis work, Purcell found that retrieval precision using context markup was slightly higher than with full-text searching, but only in an experiment where recall was artificially held constant at 100 percent for both searches. Like other postpublication approaches, this approach is limited by the existing content problems of the clinical literature; lack of clarity and missing data cannot be repaired by context markup or improved indexing.

2.2.1.2 Advanced Search Techniques

We could also improve retrieval performance by improving the accuracy of queries, or by improving the algorithm that matches queries with the information source. Librarians specialize in formulating precise and accurate queries. They also help searchers in the tradeoff of recall versus precision. For example, an evidence synthesizer would favor high recall

over high precision, because an evidence synthesis would be biased if relevant trials were not included (Section 2.1.2.2). In contrast, a busy practitioner may demand a high-precision search that yields few irrelevant articles. Search techniques that allow users to gracefully trade precision for recall and vice versa are most desirable.

Preformed search strategies. — Medical librarians use their familiarity with Medline indexing terms and strategy to achieve retrieval performance more accurate than that of novice searchers (Haynes, 1990). Preformed search strategies that capture a librarian's expertise could therefore improve novice searching. An example is the optimized search strategies devised by Haynes and associates to improve the retrieval of methodologically sound studies from the clinical literature (Haynes, 1994). Preformed search strategies can also be devised for particular clinical concepts, such as in Hepatopix (Powsner, 1989) and Psych Topix (Powsner, 1992). Users cannot, however, easily change the search to emphasize better recall or higher precision.

Word-frequency–based statistical approaches — Word-frequency–based methods match one or more keywords representing a query to the words in the documents of a document collection. Documents with a higher match rate are assumed to be more relevant to the query. Two examples are (1) string matching of keywords to an inverted file of a document, as used in standard full-text searching; and (2) matching a vector of keywords to a multidimensional vector of all the words in a document, as in vector-based retrieval (Salton, 1991). The vector-based approach has consistently performed as well as, or better than, other retrieval approaches, including the semantic approaches discussed next.

Semantic approaches — The word-frequency–based statistical approaches do not use the meanings of the keywords — the **semantics —** to decide whether or not a document is relevant. Concept-based approaches use an explicit model of the knowledge of a domain to improve the accuracy and the semantic richness of both queries and document representations. Examples include Verity Corporation's TOPIC system, and the SAPHIRE system (Hersh, 1993). These systems have not been shown to perform more accurately than the

vector-based approaches, however, and they require significant knowledge-modeling work.

Researchers in natural-language processing struggle to make computers extract the meaning from prose documents. They have achieved success in limited domains, such as with radiology reports (Hripcsak, 1995), but no solution is in sight to the problem of reading and understanding the clinical literature in general. Thus, information-retrieval engines will not soon be sending us their reading recommendations.

Machine-learning approaches — Machine-learning approaches combine semantic and statistical approaches to decide whether or not an article is relevant to a query. Supervised machine-learning approaches use a training set of queries and their associated relevant documents to derive a relevance discriminant rule. For example, we can give a neural network a keyword-based query and a document collection whose relevant trials have been marked. The neural-network algorithm will then find a network that reproduces the relevance ratings of the training set. If the network's performance can be duplicated with a second test document collection, the network is said to be trained, but only for that particular query. These supervised machine-learning techniques are impractical, because they are computationally intensive, require large training sets, and must be trained separately for each query (Salton, 1991). Unsupervised machine-learning techniques do not require training sets and may offer practical performance benefits (Hearst, 1996).

2.2.1.3 Informative Interfaces

When a search returns a large number of possibly relevant articles, the presentation of the search results can either greatly hinder or greatly improve the effective retrieval performance of the search. Confusing or ugly displays may obscure relevant information; conversely, well-designed interfaces can exploit visual and other cues to help users zero in on relevant articles, thus augmenting precision and recall. The burgeoning field of **information design** is concerned with the selection and presentation of large collections of related information — for example, use of three-dimensional visualization to present oncology

protocols (Cole, 1995), or novel displays of the characteristics of retrieved articles (Hearst, 1995). Pratt is exploring methods for using explicit models of domain knowledge to group search results into meaningful categories automatically (Pratt, 1997).

2.2.1.4 Structured-Text Reporting

Structured-text reporting is the only information-retrieval approach that I discuss that involves changing the clinical literature itself. Structuring text reporting has been recommended for abstracts (Ad Hoc Working Group for Critical Appraisal of the Medical Literature, 1987), for review articles (Mulrow, 1988), for clinical practice guidelines (Hayward, 1993), and for randomized trials (SORT, 1994; Begg, 1996). A structure is imposed onto the text. For example, structured abstracts are usually sectioned into paragraphs labelled with headings such as “Background,” “Methods,” and “Conclusion.” (See the box on page 140 for an example of a structured abstract.)

Structured-text reporting was introduced mostly to improve the content of research reporting, rather than to improve information retrieval. I defer the discussion of the efficacy of structured-text reporting for improving the reporting of randomized trials until Section 2.2.2. As for structured-text reporting’s effect on study retrieval, only the effect of structured abstracts has been evaluated; retrieval was not clearly improved (Wilczynski, 1995). Because articles with structured abstracts tend to be assigned more indexing terms than do articles without structured abstracts, any improvement in retrieval may be attributed more to increased indexing than to structuring of the abstracts (Harbourt, 1995).

2.2.2 Approaches to Improve the Content of Trial Reporting

In this section, I discuss the more institutional approaches to fixing the deficiencies in trial reporting. I defer discussion of recommendations from individual commentators to Chapter 5. For those approaches that attempt to supplant free text, readability concerns are paramount.

2.2.2.1 Supplementary Approaches

Supplementary approaches are those that make available to readers information that is not in the published, text-based article. For example, many journals contract with the National Auxiliary Publications Service to store "extensive tables of important data" (NEJM, 1996) that readers can order in microfiche or photocopy form. The contents of these data repositories are not standardized, and usually do not include information on trial methodology — the information that is generally missing and needed for evidence synthesis. Another supplementary approach is taken by Sanders and colleagues, who published a decision analysis in a traditional journal and posted the decision model itself on the web (Sanders, 1996). Use of the web to augment research reporting is a theme to which I return in Chapter 3.

2.2.2.2 Prescriptive Approaches

In contrast to the supplementary approaches to improving the content of trial reporting, prescriptive approaches stipulate how the articles themselves should be reported, and what types of data the articles should contain. These approaches reflect a growing editorial activism to improve the clinical literature.

Structured abstracts — Structured abstracts were the earliest, widely instituted form of structured reporting. Abstracts were supposed to describe the key features of a study — the key features being those needed by clinicians for selecting high quality, relevant articles of that study type (Haynes, 1990). The objective of the structured-abstracts proposal was to provide more accurate and useful abstracts, to assist peer reviewers, and to allow more precise computer-based literature searches. Evaluations of structured abstracts to date show that authors adhere only partially to the recommendations (Narine, 1991; Froom, 1993) — mainly because of space constraints — and that errors are common (Pitkin, 1997). I discussed the effect of structured abstracts on electronic retrieval in Section 2.2.1.4.

Structured trial reporting — The Standards of Reporting Trials (SORT) Group extended the structured-abstracts approach to the reporting of randomized trials. They proposed that

every randomized-trial report include, in a standardized order, information about 32 items, 24 of which they deemed "essential" (SORT, 1994). The Asilomar Working Group independently came up with a similar list (Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature, 1996). Only one trial has been published in the SORT format (Williams, 1995). The author thought that the resulting article was longer and more disjoint, and therefore less readable, than a normal article. He also thought that the SORT items captured in "exquisite detail" the information for judging internal validity, whereas the information for judging generalizability was nearly ignored (Rennie, 1994). Published comments from readers have been minimal, but we can surmise the prevailing sentiment by noting that the SORT Group has abandoned its highly prescriptive approach to improving randomized-trial reporting.

In 1996, the SORT group and the Asilomar Working Group jointly promulgated the CONSORT statement (Begg, 1996). This statement suggests that authors use a checklist to show editors where the recommended 21 items of information can be found in their manuscript. In addition, the statement recommends that authors include a flow diagram showing the progress and follow-up of patients through the trial. Several major journals, such as the *Journal of the American Medical Association* (JAMA), the *BMJ*, and the *Annals of Internal Medicine,* have adopted the CONSORT statement. It is too early to discern whether or not the CONSORT approach will significantly improve the content of trial reporting; many reporting checklists have been proposed before (e.g., Meinert, 1984; Bailar, 1988), with little demonstrable effect on overall trial-reporting quality.

Prospective meta-analysis — Several investigators have proposed that protocols of meta-analyses be published so that designers of randomized trials can coordinate the definition of their outcomes and can facilitate any subsequent meta-analysis (2nd International Cochrane Collaboration Summary Panel, Hamilton, Ontario, 1994). I am not aware of any examples of this approach being put into practice.

2.2.3 Approaches to Improve the Proportion of Trials Reported

The editors of over 100 medical journals have recently announced an "amnesty" for unpublished trials, in the hopes that researchers worldwide will notify these journals of their unpublished trials, and that evidence from these trials may someday be recovered for evidence synthesis (Smith, 1997). Although dramatic, the amnesty approach is not a promising long-term solution to the problem of publication bias. Trial investigators have little incentive to register their unpublished trial during an amnesty. Also, unpublished trials include trials that are ongoing or are being analyzed. A one-time call for unpublished trials will net far more of these in-progress trials than of completed ones (Hetherington, 1989).

In a more systematic attempt to combat the file-drawer problem (Section 2.1.2.2), Dickersin and other researchers propose that all planned randomized trials be registered at conception into trial registries (Dickersin, 1992). Trial registration could be mandated by funding agencies, or by institutional review boards. The existence of all trials would thus be known, regardless of whether or not their eventual findings were positive. Many such registries now exist, and there is even a standard list of attributes that the registries should maintain (Easterbrook, 1992). However, these registries are often incomplete, outdated, and poorly disseminated, and often include neither the results of the trials nor the information necessary for judgments of a trial's internal validity and generalizability. Still, the trial-registry idea is a powerful one; it was the initial impetus for the trial-bank system that I propose in Chapter 3.

2.2.4 Approaches to Improve Results Dissemination

A clinical research article is disseminated effectively if its target audience knows of and can access the article in a timely fashion. On the other hand, the article's *information* is disseminated effectively only if the target audience actually changes its clinical practice appropriately on the basis of the new information. Many dissemination approaches address the dissemination of articles; far fewer approaches tackle the challenge of disseminating appropriate behavior change through the literature.

2.2.4.1 Electronic Publication

Electronic publication can improve the effective dissemination of articles through a combination of technologies. For example, electronic notification services (e.g., MedConnect, 1996) can notify a subscriber quickly of new articles that match that subscriber's individual interest profile. If the articles are available in electronic form on-line, the subscriber can then retrieve them quickly and easily. Rapid dissemination is particularly important for findings that are discussed in the popular media. With electronic publication, the delay between a manuscript's acceptance and its publication is much shorter than with paper-based publication.

However, electronic publication does not always ensure effective dissemination. The On-line Journal of Current Clinical Trials (OJCCT) used to be available only through a modem and PC link. Because of its poor accessibility, awareness of its articles was poor and the journal was not even indexed in the Institute for Scientific Information's Journal Citation Reports (1996). Thus, publishing articles as bytes instead of pages is no panacea for effective dissemination.

2.2.4.2 Systematic Reviews

As I discussed in Section 1.2.1, it is unrealistic to expect front-line practitioners to synthesize the information in the literature individually, and then to change their behavior accordingly. Rather, the effective dissemination of the *information* within articles depends critically on the dissemination of systematic reviews to these practitioners. This insight is the basis of the Cochrane Collaboration's charter to disseminate systematic reviews on major clinical topics (Chalmers, 1994). The dissemination of timely, high-quality reviews from this and other groups is hampered, however, by the tremendous amount of time and labor that is currently required to complete a systematic review. A major aim of my work is to build a trial-reporting infrastructure that will ease the task of completing high quality, timely evidence syntheses.

2.2.5 Approaches to Assist Information Management

There are many computer-based approaches to improve the usability of the clinical literature. In this section, I discuss database and expert-system approaches to managing the information in the clinical literature.

2.2.5.1 Database Approaches

Database approaches use computers primarily to store and retrieve information, rather than to reason about the information. As it is in all postpublication approaches to improving the usability of the clinical literature, the quality of the information in the following systems is limited by the quality of the original trial reports.

Cochrane Collaboration software — Several projects are underway in the Cochrane Collaboration to build databases of trial information.[2] A registry of randomized and controlled trials is part of the Cochrane Database of Systematic Reviews CD-ROM. This registry includes basic information on each trial's design and bibliographic citation. The Collaboration is also developing software to coordinate the maintenance of this central registry with the registries of specialized review groups, such as the Stroke Review Group. The main purpose of these Cochrane registries is to provide sufficient trial information for identifying potential trials for Cochrane reviews, rather than to provide sufficient trial information for performing a systematic review. Therefore, reviewers must still retrieve the articles describing the trials, and they typically use a commercial bibliographic program and the Cochrane Collaboration's RevMan (Review Manager) program to complete a Cochrane systematic review.

Postpublication trial-report databases — Two database systems for storing reported, randomized trials have been described in the literature (Morris, 1992; Strang, 1994). Both are intended to help meta-analysts manage trial information; however, both systems

2. Information about Cochrane registry activities comes from discussions on the Cochrane Trials Registry Development Group electronic mailing list.

assume that data abstractors will enter only published reports into the databases, in which case these systems will perpetuate all the problems of trial reporting and publication bias that I discussed in Section 2.1.2.

2.2.5.2 Expert Systems

Expert systems exploit the computer's reasoning capabilities more than its data-storage capabilities. Classic examples of expert systems for randomized-trial reasoning include Roundsman and THOMAS, but neither system was used outside the research setting. One reason that these expert systems were never fielded is that their knowledge bases of trials were all manually entered, and therefore were small and were quickly outdated. Such a cumbersome mechanism for maintaining a system's all-important knowledge base limits severely the practicality of these systems.

Roundsman — The Roundsman system stored information about 24 breast-cancer trials, and used a rule-based approach to determine the generalizability of trial results to particular patients (Rennels, 1987). Given a particular patient, the system automatically generated a custom-tailored prose discussion on how the evidence from the most relevant of the 24 trials applied to that patient. Where appropriate, the prose discussion commented on trial-reporting problems.

THOMAS — The THOMAS system modelled randomized trials as influence diagrams to help physicians perform Bayesian statistical analyses of the trial results (Lehmann, 1991). It asked users to describe a trial, and to provide their prior knowledge and beliefs about the clinical domain. The system then calculated a posterior distribution of the trial's interpretation, and helped the user to understand the implications.

2.3 Mismatch of the Problems and the Solutions

Several of the current approaches to improving randomized-trial reporting that I have discussed address more than one of the literature's problems. Table 2.1 displays three of the

more versatile approaches. We see an incomplete overlap — a mismatch — of the problems and the solutions. Instead of this piecemeal approach to the multifaceted problems of randomized-trial reporting, we need a more comprehensive and coordinated solution.

Problem with Randomized-Trial Usability	Structured-Text Reporting	Trial Registries	Electronic Reporting
Retrieval		√	√
Completeness	√	√	
Accuracy and clarity	√		
Study methodology	√		
Standardization	√	√	
Timeliness			√

Table 2.1 Incomplete overlap of the problems and the solutions. Problems addressed even partially by the approaches are shown with a checkmark. No one approach addresses all the problems, showing that a comprehensive solution to the randomized-trial usability problem is not yet at hand.

We can begin to see a more comprehensive solution if we recognize that the same entity — a randomized trial — is the central object of all the problems I have discussed. The articles in which trial results are reported are not themselves the problem; the problems of randomized-trial reporting start at the inception of the trials, and concern the trials' design, registration, standardization, publication, synthesis, and application to particular patients. Any comprehensive approach to dealing with these problems must be trial centered, rather than article centered.

2.4 Summary

In this chapter, I discussed the nature of the problems that we all face when using the clinical literature, and when using randomized-trial reports in particular. I described current approaches to these problems, and explained their inadequacies. In Chapter 3, I propose that trial investigators report their planned protocols and the completed results of their randomized trials into structured, electronic trial banks, in addition to describing their trials in

the traditional, text-based journals. I argue that this *trial-bank system* will be a comprehensive approach to transferring the evidence from randomized trials to the clinic efficiently and effectively.

Chapter 3

The Trial-Bank System

At present, valuable scientific evidence from clinical trials is transferred to the bedside neither efficiently nor effectively. Approaches to this evidence-transfer problem have until now been piecemeal. With recent advances in databases, networking, and knowledge engineering, however, we can envision a comprehensive, computer-based approach that would have been far less realistic only a few years ago. In this chapter, I present such an approach, called the trial-bank system, and discuss its potential benefits and hurdles to implementation.

3.1 A Comprehensive Approach to the Evidence-Transfer Problem

In Chapter 2, I discussed the multifaceted problems that we encounter in applying evidence from randomized trials to the practice of medicine. I noted that, in a comprehensive solution to the problems, the central informational entity should be the randomized trial itself, rather than the article in which a trial is reported. Because randomized trials are inherently structured entities, it should be possible to describe all their most important aspects in a structured, electronic database.

My proposed solution to the evidence-transfer problem, then, is that randomized clinical trials be reported not only as text-based articles, but also as entries into network-accessible, structured, and standardized electronic databases of clinical trials, or **trial banks**. Trial banks would be standardized collections of trial protocols and summary results, and these collections of information would complement the information in the traditional, text-based articles describing the trials. The broad hypothesis of this dissertation is that a standardized structure for computer-based randomized-trial reporting can be specified, and that we can derive concrete benefits from such standardized reporting.

3.1.1 Primary Goals of the Trial-Bank System

The primary goals of the trial-bank system are as follows:

1. To make electronically available standardized information about all planned and completed randomized trials
2. To make summary-level data on completed trials electronically available at the same time that the paper presenting the conclusions of this trial appears in electronic or paper-based print
3. To structure the information such that it can be understood by computers (i.e., it is machine parsable)
4. To provide mechanisms for ensuring the quality of the information
5. To ensure that the trial banks can be shared by people and computers worldwide

These goals are similar to those of a system of electronic data publishing in bioinformatics, a scientific field concerned primarily with the biological information encoded in genetic sequences (Cinkosky, 1991). In bioinformatics, authors who wish to publish genetic-sequencing work in a scientific journal (e.g., the *Proceedings of the National Academy of Sciences* (Burks, 1989)) must first submit their sequence to GenBank, a structured database administered by the National Institutes of Health (NIH) (Burks, 1992). If the submitted sequence passes GenBank's validation checks, it is assigned an accession number, which is then appended to the published text article. Readers of the printed article

can thus immediately access and analyze the reported sequence data. This publication of bioinformatics research in both prose and as a structured database entry is a form of **electronic data publication** (Cinkosky, 1991), which differs from electronic publication in that electronic publication is simply the publication of prose in digital form.

The direct, machine-parsable availability of sequence data has changed the nature of biological research in ways not even imagined a decade ago. For example, Boguski, et al, compared the gene sequence for ataxia-telangiectasia, cloned after 18 years of work, with entries in GenBank using a sequence-matching algorithm (BLAST) (Boguski, 1995). (Ataxia-telangiectasia is an inherited cerebellar, vascular, and immunologic disorder). They found that the gene was homologous with a yeast enzyme critical for cell growth and for DNA repair; thus, they discovered a strong hint about the nature of the defect in ataxia-telangiectasia. Furthermore, the yeast-gene product was related to human proteins that are intracellular targets for certain immunosuppressive agents, thus suggesting possible therapeutic approaches. These discoveries required only minutes of computation; without the GenBank information infrastructure, each discovery might have required many years of expensive research. The benefits that will accrue from such accelerated scientific research are likely to justify the heavy investment in GenBank and in other bioinformatics reporting systems.

The bioinformatics community's success with requiring authors to submit data directly to structured, electronic databases suggests that a similar system may work for the computer-based reporting of randomized clinical trials (Section 3.1.4). If clinical-trial researchers submitted trial information directly to trial banks, as sequence researchers submit to GenBank, how might these trial banks be used? What might the implications be for clinical research and patient care?

3.1.2 Potential Uses

Although I propose the trial-bank system primarily for improving the transfer of randomized-trial evidence from the literature to the clinic, the system could also serve many other purposes. I discuss briefly the potential for a trial-bank system to improve clinical science

and practice (Section 3.1.2.1), in part by improving trial recruitment (Section 3.1.2.2) and trial reporting (Section 3.1.2.3). Certainly, there may be other uses for a trial-bank system that we cannot even imagine currently.

3.1.2.1 Improved Clinical Science and Practice

Internationally shared, up-to-date trial banks can accelerate discoveries in clinical science by enabling researchers to build on valuable evidence from clinical trials quickly and efficiently. I discuss here several mechanisms by which this acceleration might occur.

Evidence synthesis — The trial-bank system is designed specifically to support the synthesis of evidence from randomized trials. In Section 3.1.5, I discuss why the trial-bank system will improve the quality of evidence syntheses. The easier that high-quality evidence syntheses are to perform, the more they will be performed, and the more we will learn about what completed randomized trials tell us about clinical medicine.

Future technologies that are enabled by a trial-bank system may also accelerate clinical science. We can imagine expert systems that perform for us some of the reasoning tasks of meta-analysis; intelligent agents could monitor trial banks for newly published trials and could then automatically initiate a cumulative meta-analysis (Lau, 1992) to update an existing meta-analysis with the new trials. The complete automation of meta-analysis is, I believe, a distant goal, because meta-analysis as it should be performed requires considerable factual and procedural knowledge of both clinical medicine and biostatistics; meta-analysis is fundamentally a tool for exploring how and why the differences among a collection of trials are correlated with the observed results (Thompson, 1994; Sim, 1996), rather than a substitute for large randomized trials (LeLorier, 1997). The construction of fully autonomous meta-analysis agents with the requisite knowledge will remain a research challenge for years to come.

Trial design — Evidence syntheses help the research community to know what the state of the science is, and consequently to direct future research to areas of greatest need and promise (Chalmers, 1993; Olkin, 1995). Trial banks, if they contain both the protocols of

ongoing trials and the results of completed trials, could facilitate this and other trial-design approaches. For example, institutional review boards and funding agencies could search trial banks to verify that the evidence from previous randomized trials are properly noted in funding requests. Reviewers could also ensure that resources are not wasted on duplicate trials, or could ask that trial designers coordinate the outcomes to be measured so as to ease future meta-analysis of the trial results — a prospective-cohort approach to meta-analysis (Simes, 1987; Cook, 1995). Expert systems may be able to use trial-bank information to assist researchers with designing clinical trials; for example, the target sample size could be based on the size of the responses seen in previous, relevant trials. Current trial-design expert systems, such as Design-a-Trial (Wyatt, 1994), could be enhanced by access to a vast, existing network of trial banks that describes thousands of randomized trials in computer-readable form.

In summary, standardized, computer-readable access to information about ongoing and completed randomized trials can help to ameliorate the squandering of precious research resources on poorly designed trials bemoaned in “The Scandal of Poor Medical Research” (Altman, 1994). It could be argued that it is unethical to use patients and resources in clinical trials that are not likely to advance clinical science and practice. Assuming that this argument is accepted, the trial-bank system can promote clinical research that is both more efficient and more ethical.

Clinical practice — As I argued in Chapter 2, evidence syntheses form a critical link in the transfer of clinical-trial evidence from the literature to the clinic. In addition to strengthening this link by making high-quality evidence syntheses easier to perform, the trial-bank system may in the future help to improve clinical practice by encouraging the development of more evidence-based decision support systems.

The trial-bank system as I propose it will **interoperate** worldwide; that is, all trial banks will appear as one single, virtual trial bank to human and computer users. If we had such interoperating trial banks, expert systems such as Roundsman and THOMAS that interpret

clinical trials (Section 2.2.5.2) could be fielded more practically, because no system will have to maintain its own clinical-trial knowledge base. Rather, each expert system could tap into an integrated information infrastructure for clinical-trial evidence, allowing the practitioner to query on-line trial banks anywhere in the world, to use an electronic medical record to narrow trial-bank search results to a particular patient's case, or to access electronically stored quality-control guidelines. You can doubtless imagine many more uses of sharable trial banks for supporting evidence-based medicine at the point of care.

Data exploration — Trial banks will be rich repositories of both qualitative and quantitative data. We could use machine-learning techniques such as neural networks, genetic algorithms, and regression methods to discover automatically patterns in the data that may be worthy of further investigation.

3.1.2.2 Improved Trial Recruitment

A trial-bank system that contained protocols of open trials could facilitate large-scale patient recruitment. For example, research foundations and patient-advocacy groups could use the trial banks to help patients find and join appropriate trials. This situation would be more efficient that the current one, in which individual organizations are building their own trial registries (e.g., the National Breast Cancer Coalition, 1997). Furthermore, if patient records were electronic and computer readable, then expert systems could identify potential subjects by matching patient characteristics with the eligibility criteria of open trials in the trial banks. Prototypes of eligibility-matching systems have already been fielded and tested (Miller, 1995).

3.1.2.3 Improved Trial Reporting

The standardized, structured reporting of randomized trials in trial banks could vastly improve the quality of trial publication, and will enable flexibility in the presentation and dissemination of trial reports.

Quality — I detailed in Chapter 2 the argument that standardized, structured reporting of randomized trials could reduce the inaccuracies, ambiguities, and incompleteness of trial

reports. The trial-bank system implements electronically standardized, structured reporting of trials, thereby enhancing not just the quality of trial reports, but also our ability to monitor that quality.

Flexibility — Trial-bank users will be able to choose among different presentations of the same trial-bank entry. For example, evidence synthesizers could ask for a trial report that contains data elements different from those in reports for clinicians or lay readers. Since each trial-bank entry may be associated with more than one text-based article, different user groups, including lay readers, might even read custom-targeted articles. With appropriate security safeguards (Section 3.3.2.1), reports could also be generated automatically for regulatory review: The trial-bank–system design that I propose and define in Chapter 5 already includes most of the concepts in the International Conference on Harmonization standard for the reporting of pharmaceutical trials for regulatory approval (ICH, 1995). With flexibility in trial-bank presentation, we will no longer be in the irreconcilable position of reporting randomized trials in one and only one report while trying to satisfy the mutually incompatible needs of researchers, clinicians, and methodologists; randomized-trial reporting can be freed to serve the varied and individual needs of its many readers.

Through improved trial-bank search and retrieval (Section 3.1.5.1), intelligent agents will be able to monitor and deliver to us new trial-bank entries that we are likely to find interesting. Dissemination of trial results will be far more targeted, and therefore more efficient, than it is today.

3.1.3 System Architecture

In this section, I present an architecture for the trial-bank system and draw technical and sociological lessons from the bioinformatics experience with electronic data publication. Because the full implementation of the trial-bank system is well beyond the scope of my thesis research, I highlight the particular contributions of my work to the future of trial-bank publishing.

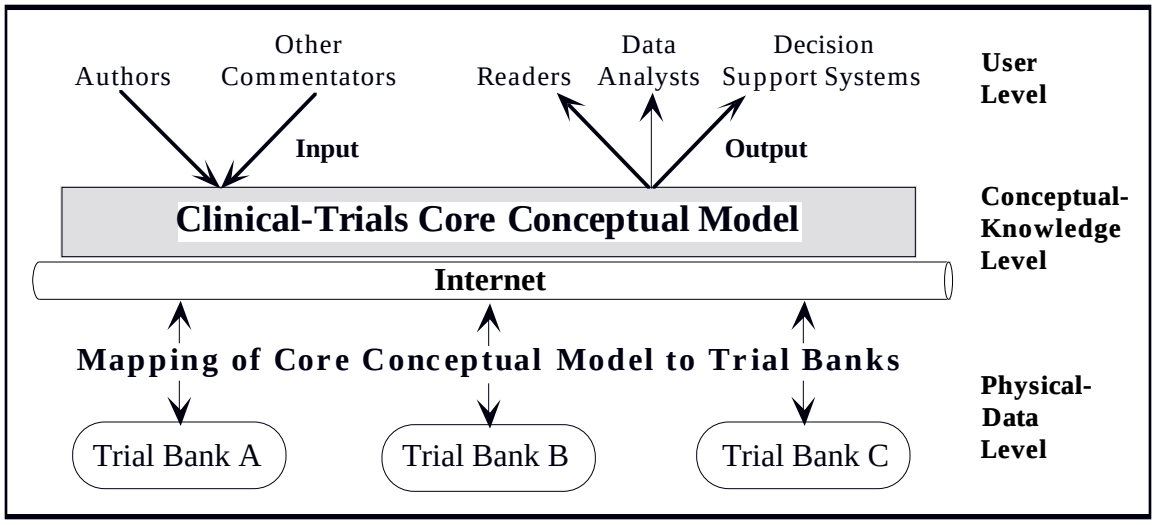

Figure 3.1. Architecture for sharable trial banks. The input to trial banks is from authors reporting randomized trials, and from commentators such as editors or letter writers. The output from trial banks can be used by clinicians or lay readers, by evidence synthesizers, and by computer-based expert systems.All use of the trial banks is integrated by the clinical-trials core conceptual model. The core conceptual model hides from users those implementation details that they do not need to know, and guides the standardized development of new trial banks.

The trial-bank system architecture can be partitioned abstractly into three levels (Figure 3.1). The bottom level is the **physical-data level**, which encompasses the actual trial banks. The top level is the **user level**, which encompasses human or computer users of the trial banks. The middle level, the **conceptual-knowledge level**, consists of an abstract conceptual model of the information that can be passed among the users in the user level, and among the trial banks in the physical-data level. The conceptual model defines not only the concepts that can be communicated, but also the kinds of statements that can be made about those concepts. For example, a conceptual model of randomized trials may define the concept “unit of randomization” while not allowing you to make the nonsensical statement that, “One of the units of randomization was lost to followup.” As shown in Figure 3.1, I propose that there be one shared conceptual model of clinical trials to guide the passing of information between any user and any trial bank. I discuss the construction of conceptual models in detail in Chapter 4; I describe my implementation of a core conceptual model in Chapter 5.

A scenario may convince you that the trial-bank system should have a shared conceptual model. Suppose that we wish to review the evidence from many trials, several of which are in, say, Morris' trials database (Morris, 1992), several in Strang's TSRS (Strang, 1994), and so forth (see Section 2.2.5.1 for discussion of these databases). In the absence of a shared model, we would have to contact Morris and Strang individually and to query their databases one by one. If, instead, we had a shared model that guided our interaction with both trial databases automatically, we could retrieve the data of interest without having to know the details of how Morris and Strang designed their databases. Having a shared conceptual model of the trial information that is being requested and retrieved is crucial to integrating the use of multiple trial banks. I defer the technical discussion of trial-bank sharing until Section 3.2.1.

3.1.4 Entry of Trials into Trial Banks

How trial reports will get into trial banks will have a heavy influence on the feasibility, timeliness, reliability, and usefulness of the trial-bank system. In this section, I discuss the bioinformatics community's experience with soliciting author participation in the GenBank project, and I identify three prerequisites for successful implementation of direct author submissions to a trial-bank system.

In the early 1990s, authors themselves submitted approximately 80 percent of the new sequences into GenBank. The bioinformatics journals' requirement that publication of sequence-research papers be contingent on the submission of data to GenBank "met with far greater acceptance on the part of authors than might have been anticipated" (Cinkosky, 1991). Now, close to 100 percent of GenBank's data are submitted directly by authors (Burks, 1992), who see the system as a natural mechanism for the communication of large quantities of scientific data. In the process, the quality of publicly released sequence data has been improved, because GenBank's *Authorin* data-entry tool automatically checks for many common sources of sequence-reporting errors.

This experience from bioinformatics suggests that there are three prerequisites for getting authors to agree to report their randomized trials into trial banks: (1) there must be user-

friendly data-entry software that help authors to submit their trials correctly; (2) authors must regard the system as being of overall benefit to the clinical-research enterprise; and (3) data submission must mesh with existing incentive and reward systems for authors. As these prerequisites are met, trial-bank data will increasingly be acquired as direct submissions from authors, rather than as transcriptions from previously published reports. We now discuss these prerequisites further.

3.1.4.1 Trial-Bank–Authoring Software

Trial-bank–authoring software should be easy to use for authors who have minimal computer experience. These data-entry programs should all enforce a standard, basic set of requirements for randomized-trial reporting, but individual trial banks may require the reporting of additional information depending on the editorial choices of their owners. For example, about 20 percent of medical journals require full disclosure of conflicts of interest (Wilkes, 1995). Trial banks associated with these journals may require disclosure of conflicts, whereas other trial banks may not. In Section 3.3.3, I discuss who might define such a basic set of reporting requirements, and how. Tutorials and interface-design challenges for trial-bank–authoring tools are topics for future research.

3.1.4.2 Overall Benefit

Section 3.1 sketched the potential benefits of the trial-bank system to the clinical-research enterprise. The overall benefit of the trial-bank system will be enhanced if authors themselves submit information about planned and completed trials directly and expeditiously into trial banks. If direct submission does not occur, standardization and automated error checking of trial reporting will not take place, but the benefits of shared trial banks can still be realized.

The demonstration of potential benefits from structured, electronic reporting of randomized trials is one of the specific aims of this dissertation (point 3 on page 4). I also discuss the time and material resources required for entering trials into trial banks (Section 7.3) and for maintaining and interoperating trial banks worldwide (Section 3.3.2.2). A

comprehensive cost–benefit analysis of trial-bank system must await the latter's full implementation.

3.1.4.3 Incentives and Rewards

We can assume that authors will submit trial reports directly to trial banks only if direct submission is either necessary or yields nontrivial rewards. Academic medical journals, as dispensers of professional credit, have great leverage over clinical-trial authors to encourage direct submission. Other potential points of leverage include the funding and the regulatory agencies. The registering of all funded randomized trials is mandated by the United Kingdom's National Health Service and has been considered by the NIH (Harlan, 1994). The Food and Drug Administration (FDA) and institutional review boards have also been suggested as points of leverage (Dickersin, 1990). Since one of the goals of the trial-bank system is comprehensive coverage of randomized trials (point 1 on page 48), I believe that any and all of these points of leverage should be used to encourage authors to describe their trials in trial banks. An implication of trial banks being maintained by many different groups is that it becomes doubly important to define a shared conceptual model of randomized trials for integrating trial banks worldwide (Section 3.1.3).

3.1.5 Advantages over Other Approaches

In Sections 3.1.5.1 to 3.1.5.5, I compare the trial-bank system to five of the approaches to improving evidence transfer that I discussed in Chapter 2. The main advantage of the trial-bank system is its comprehensive approach to the evidence-transfer problem.

3.1.5.1 Information-Retrieval Methods

Accurate document retrieval is currently hampered by the inability of computer-based methods to get at the semantics — the meaning — of the words in a trial report. For example, in the string-matching approach to document retrieval, we cannot express the query, "Find all trials in which mortality was a primary outcome." We must search instead with just the keyword "mortality," which will return all documents containing the word "mortality" regardless of how that word is used. Instead of this nonspecific approach, retrieval

methods for the trial-bank system could exploit the semantic content of the core conceptual model for more accurate trial retrieval. For the preceding example, we would search for the standard term *mortality* specifically in the trial-bank data field that corresponds to the shared concept *primary outcome*. Once we have retrieved the relevant trials, we will be free to browse the trial-bank entries, to read any associated manuscripts, or to request standard or custom reports of the trials (Section 3.1.2.3).

The design of the trial-bank system does not preclude the use of other — possibly more powerful — information-retrieval methods. For example, a retrieval method could use the concept hierarchy of a controlled medical vocabulary to implement concept-based retrieval. Also, natural-language processing could be used to search selected data fields of the trial bank. Furthermore, if all trial banks were mapped into an international, shared medical vocabulary, then automatic translation of trial-bank entries could decrease the language-based selection bias that exists in present-day evidence syntheses (Gregoire, 1995; Moher, 1996). Proving that retrieval in a trial-bank system is superior to current systems is a future research topic.

3.1.5.2 Structured Trial Reporting

The original SORT proposal for structured trial reporting asked authors to write their articles in a specific, standardized format (SORT, 1994). The reaction was sufficiently negative that the newer CONSORT recommendations ask authors only to describe the contents of their articles in a checklist to journal editors, and for the printed article to have five new subheadings (Begg, 1996). The core conceptual model that I propose as the foundation of the trial-bank system encodes these recommendations for structured, electronic reporting of randomized trials. An implemented trial-bank system will thus make this recommended trial information understandable to computers, directly retrievable, and amenable to being mixed and matched into custom displays for different user groups. The trial-bank system reincarnates the structured trial-reporting approach into its electronic form; it thereby magnifies the power of structured trial reporting by incorporating this approach into an electronic information infrastructure for the management of clinical-trials evidence.

3.1.5.3 Trial Registries

I envision that future trial registries will be a variant of full-fledged trial banks. That is, trial registries will be electronic databases that are based on the core conceptual model of randomized trials, but that contain only the protocols of trials. The core conceptual model will allow us to query multiple trial registries along with trial banks that contain primarily results, as though they were all a single database. The trial-bank system is thus designed to coordinate information about randomized trials from throughout a trial's lifecycle — from its design, to its protocol registration, to the reporting of its results. Such coordination contrasts with the current situation, where a trial's protocol information might be reported in a poorly publicized trial registry and as a design paper, and the results reported in one or more paper-based publications that are cross-referenced neither to each other nor to the trial registry.

Because an integrated information infrastructure for randomized-trial reporting would make it easy to follow a trial from registration to results reporting, the trial-bank system might improve the proportion of trials registered into trial registries. Increased trial registration is an important way to mitigate publication bias (Section 2.2.3).

3.1.5.4 Electronic Publication

The publication of all articles in electronic text, either solely or in conjunction with paper-based publication, is inevitable. The *BMJ*, for example, has just called for comments on web-based publishing of manuscripts (Delamorthe, 1996). Although the electronic publication of preprints has worked well for the physics community, the possibility of public misinterpretation of preliminary results is clearly a far greater concern in medical than in physics publication (Kassirer, 1995). The trial-bank system is meant to be an adjunct to whatever the future form of electronic publication may be. As I discuss in Section 3.3.2.1, reinventing academic medical journalism for the electronic age will remain a challenge even if the trial-bank system is never built.

3.1.5.5 Stand-Alone Trial Databases

Other proposals for storing randomized trials in databases have not considered direct submission by trial authors, so these stand-alone trial-report databases cannot correct the inaccuracy, ambiguity, and incompleteness of trial reporting that is such a problem in the current literature. Also, since the use of databases to store randomized-trial information is an obvious idea, we run the serious risk of suffering from a proliferation of stand-alone databases that have mutually incompatible database designs. The existence of the Morris (Morris, 1992) and Strang (Strang, 1994) trial-database proposals should serve as strong notice for establishing a shared, core conceptual model of clinical trials sooner rather than later, so that future trial databases will be compatible. Integrating existing incompatible databases with a post facto shared conceptual model is a thorny problem that the medical community should avoid strenuously; the bioinformatics community has recently had to invest millions of dollars in the European Bioinformatics Institute to attempt this post-facto integration of their legacy biological databases (Williams, 1995).

3.2 Trial-Bank Interoperation: Enabling Conditions

A key lesson from the bioinformatics community is that it is cheaper to design a system to be sharable from the outset than to make that system sharable after it has been built. For an interoperating trial-bank system to be realized in its entirety, we must first achieve several technical milestones. First, there must be a widely used standard for syntactic interoperation. Second, there must be a conceptual model of clinical trials for the semantic interoperation of trial banks. Third, there must be a widely used, standardized medical vocabulary for sharing the clinical meaning of the trials.

My thesis work concentrates on the second of these enabling conditions: the core conceptual model of clinical trials. The first, a standard for syntactic interoperation, is being developed in the commercial world. I will discuss the incorporation of a standardized medical vocabulary into the trial-bank system in Chapter 4.

3.2.1 Technical Considerations

Tremendous resources will be required to set up trial banks. Trial banks should thus be sharable, to avoid any duplication of invested work and money. The technical term for sharing computer databases is **interoperation**; interoperable databases are ones that can be used by other computer systems, regardless of the database's storage model (e.g., relational or object) or operating system (e.g., Windows or Unix). These databases may also be heterogeneous in their contents or in their definitions of what information they store, but an interoperating collection of heterogeneous databases will appear functionally as one database to a user. In this era of the Internet, interoperation often involves the sharing of such heterogeneous databases worldwide over computer networks. Database interoperation has two basic aspects:

1. *Interoperating the syntax* — Syntax is the *how* of saying something. For example, European and North American videotapes encode the image content using different formats, or protocols. A European videotape player will not be able to access the image content on a North American videotape.

   Likewise, if two databases encode their information using different protocols, they cannot access each other's information.

2. *Interoperating the semantics*— Semantics is the *what* that is said. In the videotape example, the semantics comprises the sound and images on the videotape. To share the semantic content of a videotape, it is necessary, but not sufficient, that we use the same syntax: We must use a European videotape player to watch a European videotape, but the semantics of an Italian movie will escape us if we do not understand Italian. Thus, to share the semantics of a communication, we must share not just a common syntax, but also a common language about the semantics — a common language for describing the meaningful content of communication.

   Likewise, two trial banks must share a common language for describing clinical trials if they are to share information about clinical trials.

Most commercial work on interoperation is concerned with establishing a standard syntax for interoperation. Examples include the COM/OLE (Component Object Model/Object Linking and Embedding) and CORBA (Common Object Request Broker Architecture) protocols. My thesis work, in contrast, is concerned with the semantic interoperation of trial banks. I have defined the computer-based equivalent of a common dictionary for trial banks to describe randomized trials. This common dictionary, which I have encoded as a clinical-trials core conceptual model, will be applicable no matter which method for syntactic interoperation may dominate the industry.

Any common computer-based dictionary of randomized trials should be sufficiently rich and flexible that trial-bank system administrators can extend it easily to include new standards and methods. For example, a clinical-trials core conceptual model should be sufficiently flexible to incorporate — without significant redesign — the reporting of Cox beta values for all survival curves if such reporting becomes standard. The core conceptual model should also be sufficiently rich to support the sharing of commonly reported, but not necessarily required, concepts. For example, the model should include the concept of conflicts of interest; as discussed in Section 3.1.4.1, some but not all trial banks will want to capture conflict-of-interest information. Detailed specifications for the core conceptual model are presented in Chapter 5.

Other technical advances that would ease the implementation of a trial-bank system include the automatic translation of conceptual models to commercial-grade database schemas, and the easy connection of commercial databases to the Internet. These advances are not far off. Although the technical hurdles to a complete implementation of the trial-bank system are considerable, a partial trial-bank system of sharable, component-based databases could be implemented now. Chapter 4 sketches the architecture of such a partial trial-bank system that can be built with current technology.

3.2.2 Operational Definition of a Trial Bank

We must have a clear definition of what qualifies as a trial bank if we are to design an architecture for sharing such entities. Our definition must be neither too narrow nor too

broad. With too narrow a definition of trial banks, we will not gain much synergy from sharing among them; with too broad a definition, we will have difficulties implementing any sharing. As discussed in Section 2.1.1, I restrict the scope of trial banks to randomized trials. I also restrict the purpose of the trial banks to the support of evidence synthesis and its four core tasks: trial retrieval, trial critiquing, quantitative computation, and contextual interpretation. Given these restrictions, I define the form and content requirements for a trial bank as follows.

3.2.2.1 Form

To qualify as a trial bank, a trials database must have a declarative **database schema**, which is an explicit description of what data the database stores and how those data are stored. A declarative database schema is required for interoperation, as I will describe in Chapter 4. This requirement excludes purely text-based collections of trial reports: Neither a file drawer of reprints nor a CD-ROM of randomized-trial articles would qualify as a trial bank. Neither would an EndNote or a ReferenceManager bibliographic database qualify. In contrast, structured, flat-file databases (e.g., Filemaker Pro), relational databases (e.g., Oracle), object-oriented databases (e.g., Illustra), and frame-based knowledge bases (e.g., Ocelot), all could be considered trial banks provided that they have the requisite content.

3.2.2.2 Content

Users of a trial-bank system must have a guarantee that every trial bank contains some well-defined minimum set of trial information. The exact contents of this minimum set are best justified by the tasks that the trial-bank system is designed to support. Chapter 5 defines this minimum information set as the clinical-trial information required by evidence synthesizers to accomplish the tasks of trial retrieval, trial critiquing, quantitative computation, and contextual interpretation of trial results.

The trial-bank system is not restricted to any particular clinical domain, because the core conceptual model that defines the information that can be shared in the trial-bank system models randomized trials independent of the clinical medicine underlying the trials

(Section 5.2.1). As long as a trial bank contains the minimum information set for all its trials, it may have any clinical breadth; it may be narrowly scoped, containing trials exploring only one disease (e.g., a hypertension trial bank) or only one intervention (e.g., a Cesarean-section trial bank), or it may include all of medicine. The medical vocabulary used in a trial bank will determine the trial bank's ability to represent clinical concepts (Section 5.2.1, page 117)

By the criteria presented in Sections 3.2.2.1 and 3.2.2.2, I have identified only three trial banks that have been described in the literature. I have already discussed the Morris (Morris, 1992) and Strang (Strang, 1994) databases. The Physicians Data Query (PDQ) from the National Cancer Institute (Hubbard, 1995), is the third trial bank. PDQ has a relational database that is primarily geared towards patient recruitment, but that does include trial-protocol information and, as of 1995, summary trial results.

3.3 Legal, Social, and Economic Considerations

The nontechnical challenges to the implementation of this proposed trial-bank system are as daunting as the technical challenges; the proposed changes touch on almost every aspect of the clinical-trials enterprise. In this section, I first consider which communities might take the initiative to implement a trial-bank system. Then, I identify the main stakeholders in the clinical-trials industry, and I specify their particular concerns. I conclude with a brief question-and-answer session.

3.3.1 Potential Effector Communities

The potential benefits of the trial-bank system will accrue to a diffuse group of people, yet a concerted and coordinated effort will be needed to build the system. In the bioinformatics system, the GenBank administrators — the National Center for Biotechnology Information (NCBI) — could not expect to keep up with the flood of sequence data unless the authors themselves submitted those data.[1] So, NCBI approached the editors of major biological journals, and together they crafted the policy that sequence data must be submitted

to GenBank before a paper presenting those data can be published. The effector communities for the GenBank system were thus a federal agency and the relevant journals.

The sociology of the clinical-trials community is, of course, different from that of the bioinformatics one. Nonetheless, in both cases, the main players are academia, government, and the journals. The academic biostatistics and evidence-based–medicine communities have long advocated improved trial reporting. Since academics are already involved in structuring and standardizing clinical-trial reports through initiatives such as the SORT proposal and the Cochrane Collaboration, they would be likely to participate in, if not to lead, any move towards use of trial banks. Clinical subspecialists and disease-specific activist groups may be involved more with disease-specific than with clinically generic trial banks.

As the American public clamors for more government accountability, federal agencies are being asked to report the results and implications of the research that they fund (Public Law, 1993). Again, bioinformatics-research reporting is an early example of the spirit of this policy. The National Center for Human Genome Research (NCHGR) has funded six DNA-sequencing laboratories with the mandate that all data be released "as rapidly as possible" on the web (Marshall, 1996). This policy may be a harbinger for mandated registration of NIH-funded trials — a policy that has been voiced but not implemented (Harlan, 1994). If the NIH should choose to expand its nascent clinical-trials registry, it could draw on the NCBI's experience with GenBank and on the National Library of Medicine's (NLM's) experience with librarianship to galvanize the trial-bank system.

The most likely effector communities for the trial-bank system then include academia, the government funding agencies, and the medical journals — especially the opinion-leading journals. In the United States, a mixture of public and private entities is likely to be involved, as discussed in Section 3.3.2.2. Worldwide, each country would have its own effector communities, ideally resulting in an internationally integrated trial-bank system.

---

1. The GenBank data-accrual rate was 20 million nucleotides per year in 1991 (Cinkosky, 1991).

3.3.2 The Main Stakeholders

Many different groups of people participate in the production and use of clinical-trial information. The trial-bank system will work only if the concerns of each of the groups are stably balanced against those of the others. The membership — and therefore the concerns — of these groups often overlap: Clinical-trial authors can become their own publishers through the web, and authors are often also consumers of other researchers' trial results. As academic publishing continues to change, new concerns and interrelationships among the stakeholders will come about. In this section, I consider the concerns of the producers, disseminators, and consumers of clinical-trial information.

3.3.2.1 Producers of Clinical Trials

The producers of clinical trials include the people and the institutions that design, conduct, and report clinical trials, as well as the agencies that fund them. Individual authors of clinical trials are primarily concerned about individual intellectual-property rights, the hassle factor, and academic credit and fairness. Security is of particular concern to institutional producers of clinical trials.

Intellectual-property rights — Intellectual-property rights are a major concern for authors undertaking any form of electronic publication. Electronic information is easy to copy, modify, redistribute, steal, or otherwise misappropriate. This concern would be magnified if patient-level data were to be published in trial banks. To allay these concerns, I propose that trial banks store only summary-level data, rather than patient-level data. Trial banks would thus publish no more information than is already published now, and the legal issues would not be more complicated than those for text-based electronic publication of the same information. Since copyright law is lagging far behind the rapid developments in electronic publishing, the definition and protection of intellectual-property rights will continue to evolve (Connect, 1996).

The hassle factor. — A major concern of trial authors is the work required to report into a trial bank. A user-friendly reporting interface is a requirement. Preliminary results from

my work suggest that the time required to report a trial into a trial bank is on the order of 10 hours, which is neither a trivial nor a prohibitive amount of time (Section 7.3). Looking at the clinical-research enterprise as a whole, this increase in the authors' work will probably be well balanced by the benefits from a trial-bank system of more efficient dissemination and use of randomized-trial evidence.

Academic credit and fairness — At present, publishing is the coin of the realm for academic promotion. Linking trial-bank publication to publication in traditional journals would therefore help authors to secure academic credit for the extra work needed to report trials into trial banks. It is imperative that a trial-bank system be adaptable to any future system of dispensing and rewarding academic credit. For example, academic medicine is changing such that promotions are increasingly based far less on the number of publications, and far more on the particular contributions of a faculty member to a project such as a randomized trial. This change in promotion criteria dovetails well with a recent proposal to list authors not simply as a linear list of names but as names with their individual intellectual contributions specified, much like credit lines at the movies (Rennie, 1997). My clinical-trials core conceptual model (Chapter 5) already supports this credit-line form of authorship.

Trial banks must also be fair in how trials are portrayed. Therefore, authors must be allowed to make qualifying comments for their trial-bank entries. Conversely, authors will find it harder to obfuscate deliberately the reporting of poorly designed or poorly executed trials, or to publish the results from the same trial in more than one journal.

Security — Institutional producers of clinical trials are concerned about maintaining institutional control of their intellectual property, although this control may manifest itself differently for different types of institutions. Proprietary institutions — some drug companies or some health-maintenance organizations (HMOs), for example — may wish to keep trial results private. Other institutions — the NIH or other HMOs, for example — may wish to

release results into the public domain. As public and private funding becomes increasingly mixed, the data-release policy may not always be clear.

A trial-bank system must exploit all available security technologies to accommodate these institutional-level intellectual-property concerns. For example, a pharmaceutical-research unit could store its own trials in a password-protected trial bank that is secured behind the company's firewall. If the company's trial bank is compatible with the core conceptual model of clinical trials, the company could expedite a trial's publication by simply downloading the relevant parts of its secure trial-bank entry into, say, a medical journal's trial bank. Likewise, the company could use a secure communications protocol to report its trial directly into the trial bank of a regulatory agency such as the FDA. The FDA then must have its own security mechanisms for controlling when the public gets what access to its own trial bank. In summary, the administrators of any trial bank must pay careful attention to the granting of appropriate access to trial-bank contents.

3.3.2.2 Disseminators of Clinical-Trial Protocols and Results

As the primary disseminators of clinical-research results, academic journals have a pivotal influence on, and responsibility for, shaping the form and nature of clinical scientific discourse. Journals must balance their financial interests with their responsibility to ensure that the results of clinical trials are reported properly for interpretation and analysis. In this section, I discuss concerns about peer review and editorial input, profit, and control of the publishing process. For simplicity's sake, I discuss the interests of publishers and editors as one, because they are both concerned with increasing quality along with circulation.

Peer review and editorial input — In his survey, Wilkes found that 97 percent of North American medical-journal editors think that peer review is necessary for maintaining high standards within the medical profession (Wilkes, 1995). In the trial-bank system, editors and peer reviewers would still review manuscripts and trial-bank entries; peer review would still play a role in making the prose more readable (Roberts, 1994), the statistics more appropriate (Gore, 1992), and the reports more balanced in their conclusions

(Goodman, 1994). As it has in the bioinformatics community, automatic error checking by trial-bank–authoring software may reduce errors, and cross-checking of trial characteristics with previously published trial-bank entries may detect attempts at redundant publication. Thus, traditional peer review may well be made easier through use of a computer-supported, standardized format for reporting trials.

Because peer review is even now an ill-defined concept (Burnham, 1990; Knoll, 1990; Wilkes, 1995), it may, in any future trial-bank system, take on many variations. The peer review community is currently exploring peer review with unmasked reviewer identity, with reviewers blinded to author identity (Justice, 1997), and with open comments from the entire reader community via web-based interfaces (van der Weyden, 1997). All these variations could be implemented easily in conjunction with a trial-bank system. Indeed, researchers who devise and evaluate new strategies for entering, reviewing, and displaying trial reports using trial banks will add to our knowledge about effective peer review.

An unchartered territory for editors will be the definition of the proper relationship between a trial-bank entry and any accompanying written manuscripts. Should the protocol details be described in both? Would readers be misled if they queried only the trial-bank entry and did not read the manuscripts, or vice versa? (See Section 7.3.2 for a brief discussion of these issues.) How scientific work can best be reported and disseminated in the digital age is an open research question.

Profit — Academic medical publishing is undergoing vast and rapid change (McConnell, 1996). Journals are hosting web sites for discussions related to published articles (e.g., *Science*'s Next Wave web site (Science, 1996)), and may soon present on-line conferences on cutting-edge research. Because the importance of randomized trials will increase with the growing emphasis on evidence-based medicine, journals will continue to disseminate and sell clinical-trial results, no matter how they reinvent themselves.

I now present three extreme scenarios for how trial banks may fit into the trials-publishing industry. I assume that journals will continue to publish the prose articles that present the

findings of trials. I also assume that a trial's intended protocol may be published separately from its final results.

In the **journal-hegemony scenario**, journals are the sole conduit for all trials information. Each journal maintains its own trial bank, and provides enhanced user services such as custom-tailored interfaces to gain a competitive marketing edge. Each journal also augments the minimal core reporting standards shared by all trial banks in an attempt to differentiate its editorial strengths. For example, one journal may demand full disclosure of conflicts of interest, whereas another may demand extreme statistical rigor.

In the **free-market scenario**, unaffiliated third parties provide and maintain trial-bank services, by contract with journals, federal funding agencies, clinical research organizations (CROs), or any other group. As in the journal-hegemony scenario, these third parties are aggressive in differentiating their trial-bank services while adhering to the minimal reporting standards.

In the **public-agencies scenario**, public agencies control all trial-bank reporting, while journals still publish the traditional prose article. Trial banks would be jointly administered by various public groups. A present-day example of such an approach is AIDSTRIALS, a trial registry of over 700 trials that is cosponsored by the FDA, NIAID, NLM, and CDC (ACTIS, 1996).

The actual market structure for trial-bank publishing will be some combination of these three scenarios. Different scenarios will dominate in different countries, because the economics of small and developing-world journals are markedly different from those of the large Western journals. The open architecture that I propose will allow the trial-bank system to operate under a wide variety of market scenarios.

Control of the publishing process — A debate is beginning about what journals produce: "The arrival of the world wide web gives us a good opportunity [...] to decide exactly how we add value to the dissemination of scientific information" (Delamorthe,

1996). Will journals be willing to publish a prose article about a trial whose trial-bank entry is not owned by that journal? Will most articles be published directly by their authors, and only a few be vetted by traditional peer review (as in the high-energy–physics community (Cohen, 1997))? Will articles be reviewed by an author's peers on the web, instead of or in addition to being reviewed by someone from a journal's stable of reviewers (van der Weyden, 1997)? How will the Ingelfinger rule[2] be enforced, if it is at all? I have no answers to these questions; the only certainty is that there will be new relationships on the horizon among the producers, disseminators, and consumers of clinical-trial results.

3.3.2.3 Consumers of Clinical-Trial Results

Consumers of clinical-trial reports include evidence synthesizers (human and machine, public and private), care providers, researchers, other health professionals, and the lay public. The trial-bank system must provide worthwhile benefits to at least one of these user groups if it is to succeed. Huth's equation analyzes information systems from the perspective of the consumer (Huth, 1985). In this equation, consumers of clinical-trial reports would benefit from an increase in the value of the numerator and a decrease in the value of the denominator:

$$\text{Value} = \frac{\text{relevance} + \text{thoroughness} + \text{efficiency}}{\text{purchase cost} + \text{access cost}} \tag{3.1}$$

The value of the numerator stands to be increased by the trial-bank system. Using the core conceptual model to integrate trial banks, we will be able to retrieve relevant trials more easily, accurately, and thoroughly than we can today (Section 3.1.5.1). We can then custom-tailor the display of the retrieved information to show only what is relevant to us at the time, and we can quickly follow electronic links to other relevant material. We will enjoy the increased efficiency of seamlessly searching distributed, heterogenous trial banks, and access time will thus be decreased. A fully-implemented trial-bank system will thus pro-

2. The *Ingelfinger rule* is that public release of an article's information must be held until the date of publication of the article.

vide more relevance, thoroughness, and efficiency than does the clinical literature of today.

The trial-bank system's effect on the denominator of the Huth equation is unclear. In the Internet world, you will probably pay for only those articles you want, rather than for entire journals. Since subscriptions to paper journals now run tens to hundreds of dollars each, overall purchase costs in the trial-bank system may not necessarily be higher than today's subscription costs. However, direct costs to individual readers may increase if libraries cease providing blanket access to expensive journals. The access cost in Huth's equation is equivalent to access time, which I discussed in the preceding paragraph.

Finally, Huth's equation does not capture the potential benefits of effective transfer of clinical evidence to clinical practice. If new truths about clinical care are discovered faster through improved evidence synthesis, we will all be the ultimate beneficiaries. Any investment in building a trial-bank system leverages the investment that we have already made in funding randomized trials. The potential benefits of the trial-bank system are therefore large, and even heavy start-up costs may be justified.

3.3.3 Questions and Answers

Following is a list of questions and their corresponding, brief, answers. Where appropriate, I refer to detailed discussions elsewhere in this dissertation.

1. **Who will determine the scope and focus of the trial-bank system?** Although the trial-bank system is a distributed system, some standardization must exist to enable interoperation. This standard would be in the form of a shared, conceptual model of randomized trials. The uptake of technical standards is a complex process, and requires a tremendous amount of time and work. There are three general routes to achieving standardization: (1) an international body, such as the International Organization for Standardization (ISO), convenes a committee that eventually produces an international standard; (2) a community convenes its own committee, achieves a consensus from scratch, and submits that consensus to an

international standards organization; and (3) an individual or committee proposes a tentative standard, solicits input from the community, and then submits the modified version to an international standards organization. My thesis work is intended to provide a principled foundation for building a consensus on a shared, conceptual model of randomized trials. In Chapter 5, I present in detail my proposed core conceptual model for integrating the trial-bank system. In Chapter 7, I evaluate the strengths and weaknesses of the model.

2. **Who determines the minimal reporting standards?** Even with a standardized model for reporting randomized trials into trial banks, there remains the separate question of which subset of trial information every trial bank should contain about every one of its trials. If a central trial-bank consensus group exists, it will decide this minimal reporting standard in cooperation with journal editors, and with whomever controls the trial-bank submission process. Further requirements may be imposed by individual trial-bank administrators (see Section 3.1.4.1). If no central group exists, then many competing trial-reporting requirements will coexist. The overall reporting standards of the trial-bank system probably will combine top-down dictates and bottom-up innovation.
3. **How important is it that all trial-bank administrators agree on a core conceptual model of clinical trials?** The concrete implication of my analysis using Huth's cost–benefit equation (Equation 3.1 on page 71) is that the benefits of interoperation are likely to be worth the work required to build and implement the core conceptual model. The trials community must not shirk the task of achieving a consensus on trial-bank design to enable trial-bank interoperation, or else there will be much needless duplication of trials-database work.
4. **Will trial-bank entries include patient-level data?** In the system as I have proposed it, no patient-level data are reported, even though evidence syntheses would more accurate if primary data were available (see Section 2.1.2.2, page 28). Besides the difficulties of ensuring that individual patients cannot be identified in the data, it is unrealistic at present to expect trial investigators to report primary

data because they are so protective of that data. However, the core conceptual model can be extended to patient-level results if the opportunity arises (See “Extension to individual patient-level data” on page 124.).

5. **Will authors be willing to write directly into the trial banks?** Authors probably will have to be compelled, through a point of leverage, to submit their trial reports directly into trial banks. The points of leverage may be at the funding level, or at the journal-publication level. The incentives that would be necessary for such direct submission are discussed in Section 3.1.4. Preliminary information on the time required for submitting a trial-bank entry is presented in Section 7.3.

6. **Who will maintain the system?** The trial-bank system would use the Internet for its worldwide connectivity, and would follow industry standards for syntactic interoperation (see Section 4.1.2). Individual trial banks could be maintained by journals, by governments, or by private groups (see Section 3.3.2.2), and probably not with great expense. A commercial-strength database, and the development and maintenance of a web site, can cost as little as several thousand dollars; more typical costs would be several hundred thousand dollars per year, which would be exorbitant for the smaller journals.

   It is unclear who will maintain the core conceptual model of clinical trials for the semantic interoperation of the trial banks. Perhaps it will be maintained by a central trial-bank consensus group, if one exists, or perhaps it will be maintained by a loose consortium of trial-bank administrators.

7. **What will happen to trials published before the trial-bank system?** In 1996 alone, 8963 articles were indexed as a “randomized controlled trial” in Medline. The work required to enter retroactively all previously published randomized trials would be astronomical; our efforts should be directed to entering only those trials previously published that satisfy some quality threshold. Quality scoring of trials is, however, still an uncertain science (see Section 2.1.2.4).

8. **How will the trial-bank system take advantage of the most current technology?** Since there are so many uncertainties about how trial banks will be administered, the most prudent course of action is to build a flexible, open architecture for trial reporting that makes as few implementation assumptions as possible. Thus, there may be one or a thousand trial banks, only summary data or also individual patient data may be reported, and the majority of users may be humans or may be intelligent computer systems. The trial-bank system is designed to adapt to, and to take advantage of, future technologies, while being based on a clear understanding of how clinical-trial results should be used.
9. **Will the benefits of the trial-bank system justify its costs?** No system costs are justified if that system cannot be shown to offer benefits. I show in Chapter 7 that benefits can from reporting clinical trials into a structured electronic database. A formal cost–benefit analysis must await the implementation of a larger-scale trial-bank system.

3.4 Summary

Academic medical publishing is rapidly moving towards electronic publication of one form or another. However, the biggest dividends from “going digital” will come not from reporting clinical trials in electronic text — which is fundamentally identical in content and form to paper-based articles — but rather from publishing trials into structured, standardized, databases, or trial banks. We now have a precious window of opportunity to use new database, networking, and knowledge-engineering technologies to build an integrated, electronic, trial-reporting system that will be a central component of an informatics infrastructure for evidence-based medicine.

The trial-bank system as I have proposed it is large and complex, and its full implementation is beyond the goals of my thesis work. I have designed, built, and evaluated two products that are prototypes of key trial-bank system components: (1) a core conceptual model of clinical trials that is necessary for the semantic interoperation of trial banks; and (2) a

trial bank with a web-based interface, with which I demonstrate concrete end-user benefits. In the next chapter, I describe the background of conceptual modeling. Then, in Chapters 5 and 6, I discuss my two dissertation products in turn.

Chapter 4

Conceptual Modeling

The most desired technical feature of a trial-bank system is the seamless interoperation of the constituent trial banks. A precondition to such interoperation is that all shared clinical-trial concepts must have the same meaning to all trial banks. A **conceptual model** is a computer-understandable encoding of the common understanding of a domain. A conceptual model defines the domain concepts and constrains how those concepts can relate to one another. For example, a clinical-trial conceptual model may define the concept TRIAL and the concept PRIMARY-OUTCOME, and constrain all TRIALS to having exactly one PRIMARY-OUTCOME. This chapter discusses how a shared conceptual model would be used for trial-bank interoperation, by presenting a technical background to the specification, encoding, and evaluation of conceptual models.

4.1 Interoperation of Information Sources

A set of electronic information sources is said to be interoperating if the sources appear to the end-user as a single system. The more heterogeneous the information sources, the more difficult interoperation is to achieve. Although the basic steps in the interoperation of heterogeneous information sources are understood, no general approach yet exists. In the following scenario, these steps are tackled with a combination of technologies from the

database and the knowledge-engineering communities. Regardless of how the trial-bank interoperation is actually implemented, it is likely that solutions from both these communities will be brought to bear on this difficult problem.[1]

4.1.1 Illustrative Scenario and Overview of Interoperation

Let us consider an illustrative scenario. Anna Lyst is using her meta-analyst workstation to synthesize evidence from trials on the effectiveness of zinc lozenges in reducing symptoms of the common cold. Her workstation has identified two randomized, placebo-controlled trials of zinc gluconate in college students who have had fewer than 24 hours of upper-respiratory–infection symptoms. One trial is stored in Malaysia in a relational database on a Windows system that is owned by a for-profit medical journal. The other trial is stored in the United Kingdom's national registry of randomized trials. The UK registry is an object-oriented knowledge base on a Unix system. The meta-analyst workstation tells Anna the name of the trials and their authors, but does not tell her where the trial information reside.

Knowing that subjects may be able to guess that they were assigned to zinc gluconate because of zinc's bitter taste, Anna is concerned that the blinding of treatment assignment may have been subverted in these trials, and that the results may therefore be biased in zinc's favor. To assess the possibility of this bias, Anna asks for data on the blinding efficacy of the trials: the percentage of subjects in each group that guessed correctly what treatment they had been given. Anna's workstation retrieves this information for her at the click of a button. She still does not know where or how this information on the two trials is stored. She does not need to know, and she does not want to know. To Anna, the meta-analyst workstation is an interface to a single information source.

---

1. In this dissertation, I focus on the sharing of data among heterogeneous information systems, and do not consider the sharing of procedural code. Extension of the trial-bank system to distributed computing is an area for future work.

Figure 4.1 illustrates the components of Anna's fictitious trial-bank world. In the following explanation of the steps in Anna's interaction with the trial-bank system, what happens at

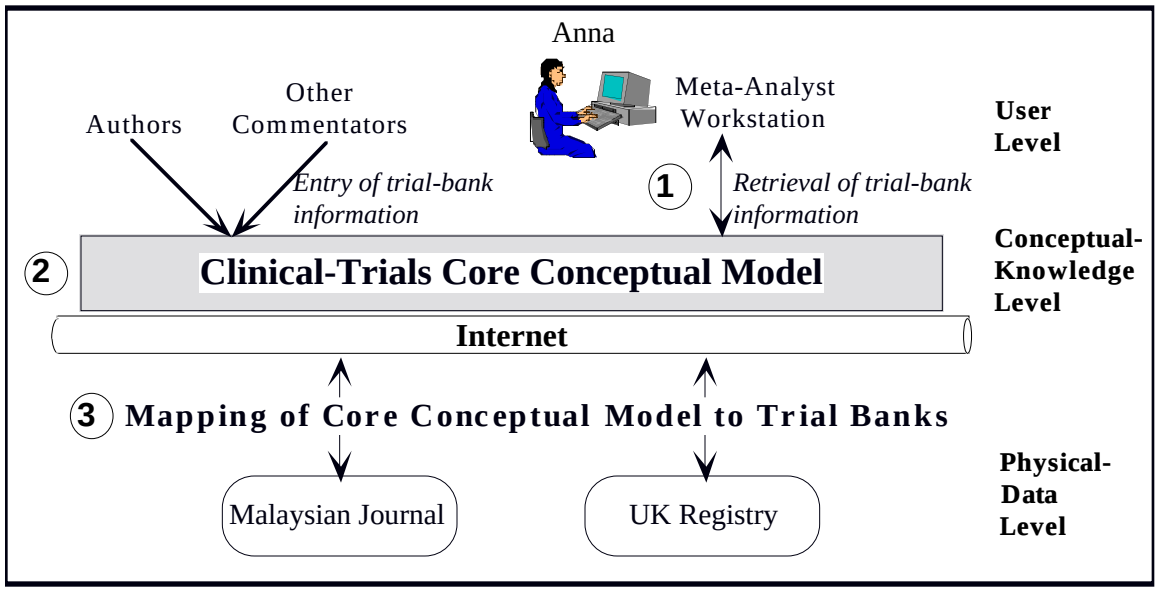

Figure 4.1. Schematic of Anna's trial-bank world. Anna is a user of the trial-bank system. Her meta-analyst workstation uses the clinical-trials core conceptual model to query two trial banks over the Internet. The text explains what happens at steps 1 to 3.

each step is of more importance than what particular technologies are used, because the technologies are in rapid flux.

1. *Step 1 (packaging the generic query)* — In the first step, the meta-analyst workstation finds on the Internet a community-defined **core conceptual model** of clinical trials and looks up the standard representation of the concept BLINDING-EFFICACY. The standardized query is then wrapped in CORBA — a set of technologies for supporting distributed computing — and the CORBA-wrapped query is then sent through the Internet.[2]

2. A full discussion of CORBA and the object-oriented distributed computing model that it espouses is given by Orfali (Orfali, 1996).

2. *Step 2 (communicating through the Internet)* —The Internet and its routers constitute the physical network over which CORBA-wrapped packages of information are shuttled correctly from one place to another. In this case, one CORBA-wrapped request is sent to Malaysia, and one to the United Kingdom. The content of the request is irrelevant to the technologies at this step.
3. *Step 3 (mapping to heterogenous trial banks)* — The receiving trial banks unpackage the request according to standard CORBA routines. They then use the core conceptual model of clinical trials — the same one used by the meta-analyst workstation — to decode the content of the information request. Now, each trial bank has to map this standardized BLINDING-EFFICACY concept to one that is stored in its own database. The mapping may or may not be successful. Thus, there are three possible outcomes to the BLINDING-EFFICACY query: (1) information on blinding efficacy is available for the specified trial, (2) information on blinding efficacy is not available for that trial, or (3) the queried trial bank does not contain any information that corresponds to the meaning of BLINDING-EFFICACY as defined in the shared core conceptual model of clinical trials. At the conclusion of this step, the two trial banks use CORBA and the same core conceptual model to return the results of the query to the meta-analyst workstation.

An analogy will highlight the fundamental requirements for requesting and receiving information from heterogeneous information sources over the Internet. Imagine the Internet as a network of highways, and information as the contents of trucks. In trucking, we distinguish among the highways themselves, the rules for driving on the highways, and what the trucks carry. Similarly, in interoperating information sources, we distinguish among the physical cables of the Internet, the protocols for sending information over the cables, and the meaningful content of the information. The rules for using highways include driving on the right, stopping at red lights, and signalling before changing lanes; these rules are analogous to transport protocols in CORBA, a syntax for sending information over the Internet. The contents of the trucks are analogous to the meaning — the semantics — of the information that is being sent across the Internet. The conventions for

how to describe a shipment (e.g., that the quantity of lumber is expressed in board-feet) are analogous to the shared conceptual model that is the convention for describing randomized-trial information.[3] Sharing of both syntax and semantics is required for trial-bank interoperation. Having syntactic interoperation (e.g., CORBA) without semantic interoperation (e.g., a core conceptual model of clinical trials) is like having a highway system via which trucks could get across the country safely, but having no common language for describing the trucks' contents. I discuss both types of interoperation further in the next two sections. Syntactic standards tend to be domain independent and are not the subject of my dissertation. On the other hand, semantic standards are of necessity domain specific; the central objective of this dissertation is the development of a core conceptual model of clinical trials to serve as the standard semantics of randomized-trial information in the trial-bank system.

4.1.2 Syntactic Interoperation

The database community is in the forefront of defining syntactic standards for the Internet. The dominant contenders for a standard syntax are Microsoft's COM/OLE, and the Object Management Group's CORBA. The knowledge-engineering community's closest candidate for a domain-independent syntactic interoperation standard is the **Generic Frame Protocol** (GFP) (Karp, 1995), which appears to have supplanted the Knowledge Query and Manipulation Language (KQML) as the upcoming standard. GFP is now the main knowledge-sharing technology of the High-Performance Knowledge Base initiative of the Defense Advanced Research Programs Agency (DARPA).

The establishment of syntactic standards is the subject of a hotly contested battle, and the database community's approaches are clearly overshadowing the knowledge-engineering

---

3. The distinction between syntax and semantics is not always clear-cut. The convention for how a bill of lading is written up could be considered a highway-navigation rule — for example, if a bill of lading is necessary for crossing state lines— or it could be considered a loose standard for describing a truck's contents. Similarly, CORBA has conventions that are neither purely syntactic nor semantic. Nevertheless, it is useful conceptually to distinguish between syntactic and semantic standards for trial-bank interoperation.

approaches. Indeed, GFP is scheduled to be CORBA-compliant in the near future. This uncertainty in what the future syntactic interoperation standard will be — or even whether there will be a single standard — reinforces the importance of a trial-bank–system design that is independent of any particular current technology (point 8 on page 75).

4.1.3 Semantic Interoperation

The importance of a shared, core conceptual model of clinical trials for interoperating trial banks cannot be overstated. In Figure 4.1 on page 79, the core conceptual model of clinical trials is depicted as playing a critical role in describing the meaning — the semantics — of the information that is shared among the trial banks. This section describes the role of core conceptual models in interoperation. The specification, encoding, and evaluation of these models is discussed later in the context of their use in interoperation.

Information sources are said to be **heterogeneous** if they represent concepts from the same domain in different ways. For example, two trial banks would be heterogeneous if, in one, the term TRIAL-TITLE refers to the title of the funded grant application, whereas in the other trial bank, TRIAL-TITLE refers to the title of the published final report of the trial. The clinical-trials core conceptual model in Figure 4.1 plays the role of a dictionary that allows strangers — or, in this case, heterogenous information sources — to talk about abstract concepts in a shared, meaningful way. At present, unfortunately, there are precious few examples of shared conceptual models being used successfully to interoperate heterogenous information sources. Why are there so few successful examples?

Shared conceptual models in the database community — In the database community, shared conceptual models are called **common data schemas** or **global schemas**. The most successful use of global data schemas for interoperation occurs in distributed database systems, where system designers specify a global schema, and then design the constituent databases to comply with that schema. The result is a constrained and manageable heterogeneity that is compatible with interoperation. In contrast, if the constituent databases are already heterogeneous in their design, system designers will have a difficult

time integrating these heterogeneous designs into a single global schema for interoperation. We will discuss the reasons for this difficulty after we discuss some knowledge-engineering approaches to interoperation. In any case, the lesson for trial-bank system builders is clear: If we wish to have an interoperating trial-bank system, it will be far easier to start with a global schema and to build compatible databases based on that schema than it will be to integrate heterogenous databases with a global schema after the fact.

Shared conceptual models in the knowledge-engineering community — The use of shared conceptual models — or of shared **ontologies**[4] — to integrate disparate knowledge bases has been a strong theme of knowledge-engineering research (Fikes, 1991; Neches, 1991; Musen, 1992). Unlike the situation with databases, there are few extant knowledge bases that need to be integrated with a post-facto shared ontology.[5] The predominant knowledge-sharing approach is rather to define shared ontologies before the knowledge bases are built, in the hopes that other knowledge engineers will use these shared ontologies to build compatible knowledge bases.

The Ontolingua library (Ontolingua, 1996) provides a cautionary tale for knowledge sharing. This library has 34 ontologies, covering domains as disparate as Bibliographic-Data and 3D-Tensor-Quantities. However, there has not been a clearly documented example of reuse of any Ontolingua ontology either for building a new knowledge base or for interoperating existing ones. There are two major reasons for this absence of reuse.

---

4. An **ontology** is a catalog of the types of things that are assumed to exist in some domain of interest, as seen from the perspective of a person whose purpose is to communicate with other people about that domain. The ontology is expressed using a natural or artificial language. I use the term *ontology* interchangeably with the terms *conceptual model* and *data schema*.

5. The distinction between a **knowledge base** and a **database** is one of degree. Knowledge bases generally contain fewer instances of more complicated entities compared to databases. For example, the EcoCyc knowledge base contains about 100 metabolic pathways of the bacterium *E. coli* (Karp, 1996). A typical database might contain 10,000 instances of *Employee* with only four attributes each: Name, Address, Employee Number, and Salary.

1. *Ontological scope must be clearly specified* — Ontologies are only approximations of the world, and thus their scope is by definition limited. As discussed by the KADS group (Wielinga, 1991) and by other researchers, an ontology's **scope** can be described with three parameters: (1) the domain, or the part of reality that is being modeled; (2) the tasks, or the goals that the ontology is to support; and (3) the methods, or the specific actions that will be undertaken to accomplish those tasks. Ideally, an ontology's scope should be the result of careful analysis and justification, and should be clearly documented for people who may wish to reuse the ontology. Although some investigators believe that an ontology can be specified with no acknowledgment, even implicitly, of the tasks and methods that it is intended to support (Lenat, 1986), I do not share this view.

   The reuse of an ontology — for building a new knowledge base, or for interoperating existing knowledge bases — is predicated on a clear specification of that ontology's scope. The scopes of the Ontolingua ontologies are not clearly documented, perhaps because their scopes were never explicit in the first place.

2. *Extensibility is required* — The reuse of an ontology probably will not be for exactly the same domains, tasks, and methods that the ontology was originally designed to support. It should therefore be less trouble to extend an ontology for sharing than to build a new one from scratch (Gruber, 1993; Gruninger, 1995).

The paucity of demonstrable ontology reuse shows that it is not a trivial task to interoperate heterogenous databases of complex information with shared ontologies or conceptual models. Ontologies are difficult to use for interoperation in large part because their scope is often poorly conceived and poorly documented. If system builders cannot tell what domain an ontology covers, and cannot tell what tasks the ontology is guaranteed to support, then it should come as no surprise that the ontology is not reused by anyone outside of the original design team. Ontologies must be more precisely designed, specified, and documented if reuse is to become routine.

Even with more precise ontology design, specification, and documentation, however, there will be other serious impediments to interoperation with ontologies, shared conceptual models, or global data schemas. How to properly index the domains, tasks, and methods of a conceptual model is an open research question. How, for example, would the meta-analyst workstation in Figure 4.1 decide between the shared clinical-trials core conceptual model and the shared model for critical appraisal? What if a query can only be represented by combining two shared models? Despite these and other challenges to implementation, shared conceptual models are necessary, although not sufficient, for the interoperation of heterogeneous information sources.

4.2 Specification of Conceptual Models

The majority of extant ontologies is poorly specified (Noy, 1997), although several specification methodologies exist. The KADS methodology (Wielinga, 1991) is geared toward the design and construction of an entire knowledge-based system, which is beyond the goals of my work. For the specification of conceptual models themselves, the task-decomposition approach of Chandrasekeran and colleagues (Chandrasekaran, 1993) and the competency-questions approach of Gruninger and colleagues (Gruninger, 1995) are more applicable. We can use these methods to state a design specification for a conceptual model that is yet to be built, and we can also use them to document the scope of a completed model.

4.2.1 Task-Decomposition Approach

The objective of the task-decomposition approach is to analyze how knowledge is related to that knowledge's use (Chandrasekaran, 1993). The approach starts from the realization that many different methods can be used to accomplish the same task, and that each method may have associated subtasks that may themselves have to be decomposed. For example, we can rate a trial's quality using the methods of Chalmers (Chalmers, 1981), Detsky (Detsky, 1992), or of many others. If we choose to use the Chalmers scale, rather

than the less extensive Detsky scale, we will have to complete a larger set of subtasks, each of which we can complete using several different methods. A task-decomposition analysis therefore generates a recursive task–method–subtask characterization of what domain knowledge is used for accomplishing what particular task with what particular method. Such a characterization is useful for the design and specification of conceptual models. The task-decomposition approach has not been used for evaluating conceptual models.

4.2.2 Competency-Question Approach

Gruninger and colleagues propound a methodology for conceptual-model (or ontology) design that is driven by **competency questions** (Gruninger, 1995). An ontology's competency questions are those questions that the ontology is guaranteed to be able to answer. For example, "What is the sample size of the trial?" may be a competency question in the design specification, or in the documentation, of a clinical-trials core conceptual model. If this competency question is in the design specification, then anyone building a clinical-trials conceptual model will know to include the concept of SAMPLE-SIZE in the modeling. If this competency question is in the documentation of a completed clinical-trials conceptual model, then we are guaranteed that the SAMPLE-SIZE concept is encoded in the model such that we can determine the sample size of a trial. To satisfy this competency, a model can encode the SAMPLE-SIZE concept either directly, or as the sum of all the patients in the control and experimental arms. Competency questions thus serve only as constraints on what the ontology must do, rather than determining how the ontology should be implemented. A well structured set of competency questions is stratified such that higher-level questions are phrased in terms of lower-level questions. Such a hierarchical organization improves the clarity of the decomposition.

4.2.3 Competency-Decomposition Specification

The differences between the task-decomposition and the competency-questions approaches are two. First, the competency questions approach does not incorporate the notion that competencies may be achieved through more than one method. Either a competency is satisfied or it is not. Second, the competency-questions approach can be used to specify and to evaluate conceptual models, whereas the task-decomposition approach is intended only for analyzing the use of knowledge in these models.

I have combined these two approaches into the **competency-decomposition** approach for specifying and evaluating conceptual models. In this approach, Chandrasekaran's tasks are analogous to Gruninger's competencies. A design specification using this approach consists of a competency decomposition, and a catalog of required domain concepts. The highest-level objectives — or competencies— are decomposed into subcompetencies, and into methods when appropriate. Justifications for the decomposition are documented. The catalog portion of a design specification then specifies the necessary and sufficient domain concepts for fulfilling each competency using the stated methods. Table 4.1 is a competency decomposition of the quantitative-computation competency.In this example, the design specification is for a conceptual model that support the tasks of quantitative meta-analysis using the Mantel–Haenszel method to combine odds ratios. The required trial

Table 4.1 Quantitative synthesis competency-decomposition.

Competency	Method	Method-Associated Subcompetency	Data Requirement of Clinical-Trials Model
I. Calculate summary statistic, for pairwise comparisons	A. OR[a]	1. Calculate OR	a. Complete 2 X 2 contingency table
II. Quantitative meta-analysis	A. Mantel–Haenszel, using OR	1. Calculate OR for each trial	a. Same as I.A.1-2.a
		2. Calculate meta-analytic summary	a. ORs for all the trials

a. odds ratio

concept is a complete 2 X 2 contingency table. This design specification states that any conceptual model intending to support Mantel–Haenszel meta-analysis with odds ratios must contain the concept of a 2 X 2 contingency table for the outcome to be combined. As is the case for competency decompositions in general, this design specification does not impose any further requirement on how the contingency table concept is encoded. The design specification for the clinical-trials core conceptual model is given in Appendix A and is discussed in Chapter 5.

4.3 Encoding of Conceptual Models

The meaning of and the relationships among the concepts in a domain must be encoded using a consistent notation — a language. Systems of notations for encoding conceptual models are called **knowledge-representation languages**. The language that we choose will constrain the kinds of abstract concepts that we can express, much as choosing German to say “Ich sehe” will constrain us from communicating the distinctions among “I see,” “I am seeing,” and “I do see.” However, English is not more correct than German in any absolute sense. Similarly, there is no one correct knowledge-representation language for encoding the semantics of the concepts that will be shared in a trial-bank system. The appropriate language is the one that is most suited to the intended purpose of the interoperation. Picking a representation language involves many tradeoffs and considerations.

1. *Formality versus human understandability* — Computers are highly intolerant of ambiguity. The more formal the language used to encode a conceptual model, the more easily that model can be shared by computers. **Formal languages** are languages with extensive and unyielding rules that leave little, or no, room for ambiguity. The tradeoff against formality is that rigidly formal languages are difficult for most humans to understand, and this difficulty can impede the model’s adoption for interoperation.

2. *Expressivity*— The knowledge-representation language in which we encode a conceptual model may force us to express aspects of the world that are not important for our needs (e.g., necessary and sufficient conditions for being a person), or it may be unable to express aspects of the world that we wish to model (e.g., that death is a permanent state). An appropriate knowledge-representation language for a conceptual model is a language that provides just the sufficient expressivity needed to accomplish that model's competencies.
3. *Standardization* — Conceptual models that are intended to support interoperation should be expressed in a standardized knowledge-representation language. However, several of the main classes of knowledge-representation languages have many dialects each, none of which serves as a standard.
4. *Conciseness and ease of maintenance* — A human-factors consideration for implementing an interoperating trial-bank system successfully is that the clinical-trials core conceptual model should be concise and well documented, and therefore easy for humans to understand and to maintain. Some knowledge-representation languages are particularly cumbersome, and others are particularly concise.
5. *Inferential power* — Depending on its intended purpose, a conceptual model may require a knowledge-representation language that is capable of inference. For example, a terminology may benefit from being encoded in a language with automatic classification capabilities.

The balancing of these considerations must be guided by the purpose of the modeling. The ideal representation language for encoding shared conceptual models would be fast, expressive, and easily understandable by humans as well as by computers. Sections 4.3.1 to 4.3.4 present several common knowledge-representation languages in the context of these considerations.

4.3.1 Natural language

Natural language is the class of knowledge-representation language with which we are most familiar. It includes all the languages of the world's peoples, and it is the most expressive of the knowledge-representation–language classes. Although there is no standard natural language, standards such as English exist for enormous communities and can therefore be widely understood.

Natural language is also the most informal of the knowledge-representation languages. It leaves much to the reader's interpretation and is notoriously difficult for computers to comprehend. Therefore, although natural language is expressive, sufficiently standardized, and easy for humans to understand, it is not sufficiently formal for encoding shared conceptual models for interoperating trial banks.

4.3.2 First-Order Logic

At the other end of the spectrum of formality from natural language is **first-order logic**, or predicate calculus.[6] Logic is the classic knowledge-representation language for computer systems, and many ontologies are expressed in variants of first-order logic (e.g., the Knowledge Interchange Format (KIF) ontologies). Logic is also one of the most concise and expressive languages, and has the capability of inference, through modus ponens and other mechanisms. However, conceptual models encoded in first-order logic are not likely to be used widely outside of a research setting, for four reasons. First, the very formality that enables computer-based interoperation often makes the language difficult for most humans to understand. Second, there does not exist a standard first-order logic.[7] Third, at least one of other major classes of computer-based knowledge representation languages —

---

6. **Logic** follows the strictures of formal mathematics. When used as a knowledge-representation language, it makes ground assumptions — called **axioms** — about the world, and posits an internally consistent set of statements about the world, called a **theory.**

7. The Knowledge Interchange Format (KIF) (Genesereth, 1990) was originally proposed as a standard. There are also no standard versions of the more advanced logics that are able to model time (temporal logics), generalizations over functions and relations (second-order logic), or uncertainty (fuzzy logics, nonmonotonic logics).

the object data-definition languages — is less difficult to maintain and is sufficiently expressive for most types of tasks for which conceptual models are intended. Fourth, most information systems do not require the type of inference that logic provides. Most systems do well with one language for encoding static knowledge (e.g., a relational data model for a database) and another language for encoding procedural knowledge (e.g., C++ for sorting database entries). In summary, first-order logic is not an ideal language for encoding a shared conceptual model for interoperation.

4.3.3 Relational Data-Definition Language

The **relational data-definition language (DDL)** and the **object data-definition languages** are computer-based knowledge-representation languages that are intermediate in formality between logic and natural language. The relational DDL represents data entities and the relationships among those entities as a collection of tables, in which each entity's properties are stored in columns with unique names (Table 4.2). The data schema, or conceptual model, of a relational database is the description of the database's tables and their associated column names. Database designers often design relational data schemas using rules derived from mathematical set theory that ensure optimal efficiency for computation. Unfortunately, these **normalization rules** often result in related concepts being decomposed into multiple tables, such that the underlying relationships among the data become obscured. The limited expressivity of the relational DDL — which I will discuss in tandem with the expressivity of object DDLs in Section 4.3.4 — and the opaqueness of relational

Patient-ID	Patient-Name	Sex	Age	Provider-Name	Clinic Location
1	Bill Lee	M	27	Owens	Palo Alto, CA
2	Janice Jones	F	82	Rennels	Santa Clara, CA

Table 4.2 Relational table example. Each patient occupies one row. Attributes of patients are entered into uniquely named columns. An example relational DDL encoding for this patient table is *Patient–scheme=(Patient-ID, Patient-Name, Sex, Age, Provider-Name, Clinic-Location).*

data schemas, makes this knowledge-representation language less than ideal for encoding a conceptual model of a domain as complex as clinical trials.

There is a role, however, for the relational data model in the trial-bank system. Most commercial databases are relational, and their security, scalability, and robustness are highly valuable for mission-critical use. Relational databases can also be queried with a standard language: SQL (Structured Query Language). It is likely, therefore, that a large proportion of trial banks in the near future will be relational databases, and that the clinical-trials shared conceptual model will often have to be mapped to a relational data model in step 3 of Figure 4.1. This mapping will entail a loss of expressivity, as described in Section 4.3.4. Several object-relational databases are now on the market, and these hybrid databases may provide both the superior functionality of relational databases and the higher expressivity and conciseness of object databases.

4.3.4 Object Data-Definition Languages

The object DDLs represent data as a collection of objects and their attributes. Attributes may themselves be described by other objects. Objects may correspond to concrete entities (e.g., *patients*) or to abstract concepts (e.g., *blinding efficacy*). Objects whose attributes do not reference other objects are **atomic**, whereas objects whose attributes are defined in terms of other objects are **compound**. For example, if a DRUG object has an attribute MANUFACTURER that is described by the COMPANY object, then DRUG is a compound object. The generic description of an object (e.g., that a drug has an attribute MANUFACTURER) is called a **class**, and the particular members of a class (e.g., amiodarone, manufactured by Wyeth-Ayerst) are called **instances**. The contents of an object database are therefore described abstractly by a collection of classes. This collection of classes is called an **object data schema** in the database world, and a **class definition**, **class hierarchy**, or a **knowledge-base ontology** in the knowledge-engineering world. These terms are used synonymously with the term **conceptual model** in this dissertation.

Unlike the relational DDL, the object DDLs are not grounded in a formal mathematical theory, and many variations of object DDLs exist. Considering only those properties that

are common to all the object DDL variants, we can still conclude that this class of knowledge-representation languages is suitable for encoding shared conceptual models for interoperating trial banks.

Expressivity — The expressivity of the main knowledge-representation language classes can be ordered from highest to lowest as follows: natural language, first-order logic, the object DDLs, and finally the relational DDL. Natural language can express any concept. First-order logic is unable to express **nonmonotonicity** (e.g., that a person whom we thought was dead is actually still alive) and probabilistic **uncertainty** (e.g., that we are not sure whether or not a person is dead). In addition to nonmonotonicity and uncertainty, the object DDLs are also unable to express **negation** (e.g., that the negation of life is death) and **disjunction** (e.g., that a person is either alive or dead, but cannot be both). The relational DDL is the least expressive of these knowledge-representation languages. It is geared toward representing particulars about the world — for example, that 462 patients were enrolled in trial A. In contrast, the object DDLs can represent both particulars and generalizations about the world — for example, that 462 patients were enrolled in trial A and that, in general, all completed clinical trials have at least one enrolled patient. Other examples of generalizations that the object DDLs can express include defaults (e.g., that people are alive unless otherwise specified), and taxonomic hierarchies (e.g., that a patient is a kind of person). Generalizations about the world constitute knowledge about the world; because the relational DDL is unable to express generalizations about the world, its expressivity is generally insufficient for encoding shared conceptual models for interoperation. The determination of whether object DDLs or first-order logic is the more appropriate language for encoding a shared conceptual model will turn on whether or not the expression of negation and disjunction is required for that model's intended competencies.

Inheritance — Many object DDLs have an inherent form of inference called **inheritance**. The basic idea behind inheritance is that, if a group of concepts share similar attributes, then those shared attributes may be abstracted into a higher-level concept. For example, in Figure 4.2, the concepts INTERVENTION and DRUG share the attributes NAME and

INTERVENTION

NAME

Datatype: string

INTENTION

Datatype: (Prevention or Treatment)

DRUG [inherits from INTERVENTION]

RECOMMENDED-DOSE

Datatype: number

Figure 4.2. The DRUG class inherits from the INTERVENTION class. DRUG inherits the NAME and INTENTION attributes of INTERVENTION, including the specification that NAME must be of the *string* datatype, and INTENTION must be either Prevention or Treatment. In addition, DRUG has its own attribute RECOMMENDED-DOSE, which must be of the *number* datatype. In this conceptual model, interventions that do and do not have recommended doses (e.g., SURGERY and DRUG, respectively) can share the specifications of the NAME and INTENTION attributes of INTERVENTION.

INTENTION. However, DRUG also has the attribute RECOMMENDED-DOSE. If we define DRUG to be a **child concept** of INTERVENTION, then DRUG inherits all the attributes of its parent concept INTERVENTION. Inheritance is a form of inferencing, because all the rules and constraints that apply to the parent apply to the child as well, unless otherwise specified. For example, the restriction of INTERVENTION indications — to only prevention or treatment — also applies to the indications for DRUG. Through inheritance, rules about datatypes and default values can be propagated systematically via explicit relationships among concepts. Different object DDLs have different approaches for when a child concept inherits conflicting attributes from two parents — a situation called **multiple inheritance**.

Inheritance can also generate new truths that may be logically implied by the information already specified. For example, if INTERVENTION is a child of TRIAL-CONCEPT, then we can conclude that DRUG, too, is a TRIAL-CONCEPT. Thus, when we are modeling complex entities (e.g., a clinical trial), inheritance is a powerful tool for managing, maintaining, and encoding a conceptual model compactly. In addition to its expressivity and explicitness,

these advantages make the object DDL class of knowledge-representation language a good choice for encoding shared conceptual models for interoperating trial banks.

4.4 Evaluation of Conceptual Models

Conceptual models are finite approximations of an infinite reality. In designing a conceptual model, a modeler chooses which aspects of reality to include in the model. There are, however, no absolute criteria for determining the appropriateness of these choices; the appropriateness can be stated only with respect to the modeler's goals. Thus, the evaluation of conceptual models is inherently tautologous, and can be unsatisfying to people accustomed to the traditional notion of evaluation as a comparison against a gold standard.

A profusion of approaches exists for conceptual-model evaluation. These approaches typically are so informal that firm conclusions cannot be drawn, or are so formal that routine use is impractical.

4.4.1 Review of Evaluation Approaches

Researchers have tried to adapt the notions of verification and validation from software engineering to the evaluation of conceptual models, but the adaptations are often more confusing than helpful. For example, Levy uses the standard notion of **verification** for evaluating knowledge bases: whether, "for any correct set of inputs (i.e., problem instance), the knowledge base entails correct outputs" (Levy, 1996). However, this evaluation approach is intended for information *systems*, and not for *conceptual models;* models do not have inputs or outputs. It is incorrect to equate verification of an information system with verification of that systems' underlying conceptual model.

Another use of the term verification is exhibited by Gomez-Perez, for which verification "refers to the technical process that guarantees the correctness of an ontology" (Gomez-Perez, 1995). This undefined process involves verifying the "correctness of definitions and axioms," determining what is and is not modeled, and determining what can and cannot be

inferred from the axioms. The correctness of the definitions, axioms, and inference is judged by whether they "satisfy [the model's] requirements, its competence questions or perform correctly in the real world." Confusion arises because Gomez-Perez defines **validation** similarly: validation is an undefined process for ensuring the "coherence, completeness, consistency and conciseness of the definitions," and "whether the ontology definitions are necessary and sufficient to represent the tasks and their solutions for different uses." Because no explicit methods are described for achieving either verification or validation, it is difficult to make material distinctions between these two approaches.

Gruninger and colleagues describe the use of competency questions (Section 4.2.2) for evaluating a conceptual model (Gruninger, 1995). The basic idea is first to encode in first-order logic the concepts in the conceptual model (or ontology), as well as its intended competencies. Then, the evaluation consists of determining whether the model yields the competencies intended, and whether it yields undesired competencies. Because the knowledge-representation language is first-order logic, the competency questions can be framed as "an entailment or consistency problem with respect to the axioms in the ontology" (Gruninger, 1995, p. 32). Then, if $T_{ontology}$ is the set of axioms in the proposed ontology, and T_{ground} is the set of ground literals (instances), and Q is a first-order sentence using only predicates in the language of $T_{ontology}$, the evaluation can be stated as follows:

1. Determine $T_{ontology} \cup T_{ground} \models Q$
2. Determine whether $T_{ontology} \cup T_{ground} \not\models \neg Q$

The advantage of this competency-questions approach is that the evaluation is performance based. Verification and validation are concerned with whether or not an ontology is "correct"; this approach is concerned with what an ontology claims to do, and with whether or not it succeeds. This competency-questions approach formally describes the tasks that are supported by an ontology, and this information is useful for deciding whether this ontology can be reused appropriately for building or interoperating other knowledge bases. The downside to this approach is that information about the ontology's

domain and methods — the other two parameters that describe the scope of an ontology — are not also formally described. In addition, the competency-questions approach requires that a conceptual model and the associated competencies be encoded in first-order logic. This requirement forms a practical barrier to widespread use of this approach.

4.4.2 Competency-Decomposition Approach

My evaluation approach adapts Gruninger's competency-questions approach to be applicable for conceptual models that are specified using competency decomposition (Section 4.2.3), and that are encoded in knowledge-representation languages other than first-order logic. Compared to Gruninger's approach, this relaxation of the first-order–logic requirement comes at the expense of being able to perform a closed-form proof of competence. However, this shortcoming is balanced by this approach being systematic, rigorous, and complete with respect to the model's claimed competencies, and yet not being too onerous to perform.

The basic idea behind this approach is to use a model's competency decomposition as the model's own gold standard for competency and for conceptual coverage of the domain. This approach quantifies a model's competency (i.e., the extent to which the model supports its claimed competencies), and its conceptual coverage (i.e., the extent to which the model meets all of the data requirements for its claimed competencies). The application of the competency-decomposition approach to the evaluation of this thesis work is detailed in Chapter 7.

4.5 Summary

Although a shared conceptual model is needed for trial-bank interoperation, there are no absolute criteria for determining the correctness of such a model. The correctness of a model can be assessed only with respect to the tasks that the model or its associated databases are designed to support. Based on this task-based perspective, I presented a new approach, the competency-decomposition approach, to the specification and evaluation of

conceptual models. In Chapter 5, I use this approach to describe the design specification for a shared, clinical-trials core conceptual model for interoperating trial banks.

Chapter 5

The Core Conceptual Model

The centerpiece of my thesis work is the design, implementation, and evaluation of a core conceptual model of clinical trials for the semantic interoperation of a trial-bank system. This core conceptual model will define the concepts that can be shared among disparate trial banks. I had four technical objectives for the construction of this conceptual model:

1. Specify the design of the core conceptual model such that the model's scope (i.e., its tasks, methods, and domain) is clear.
2. Encode the core conceptual model in a knowledge-representation language that is concise, understandable, and sufficiently expressive for evidence synthesis.
3. Ensure that the model is adaptable to foreseeable technical standards for computer-based interoperation.
4. Evaluate the core conceptual model for its competencies and conceptual coverage.

In this chapter, I present and justify the **design specification** — a stand-alone blueprint — for a clinical-trials core conceptual model that satisfies my first technical objective. I also describe my implementation of a conceptual model, called **Ocelot-CCM**, that satisfies the second and third objectives. The work for the fourth objective is discussed in Chapter 7.

5.1 Design Specification

The design specification of the clinical-trials core conceptual model is a competency decomposition (Section 4.2.3) that details the trial information that a clinical-trials core conceptual model should be able to express if it is to interoperate a trial-bank system. The design specification was the result of an iterative modeling process. I first modeled the concepts most central to evidence synthesis (i.e., from the middle out), as suggested by Uschold (Uschold, 1996), rather than starting by modeling the most general concepts (i.e., top-down), or the most specific concepts (i.e., bottom-up). Initially, I encoded the model in natural language (English). Then, I encoded the model as FilemakerPro 2.0[1] trial-abstraction forms that have an informal, but regular, structure. The meta-analysis team of the Cardiac Arrhythmia and Risk of Death Patient Outcomes Research Team (CARD PORT) used these forms to capture trial information in English, and we iteratively modified these forms — and hence the underlying conceptual model — as we abstracted 31 randomized trials for two meta-analyses (Sim, 1995; Sim, 1997).

5.1.1 Clinical-Trials Modeling Space

Modeling an abstract entity such as a clinical trial can be an amorphous task. We can visualize the task by imagining a modeling space of clinical trials with dimensions that correspond to the three parameters that characterize a conceptual model's scope (Section 4.2): the tasks, methods, and domain (trial features). Figure 5.1 (page 101) shows the figurative boundaries of the modeling space that my clinical-trials core conceptual model captures. Figure 5.1 also illustrates an additional sense in which the design specification is core: Beyond supporting the core tasks of evidence synthesis, the specification is core because it can be extended to support new tasks, methods, and domains (Section 5.1.3). The design specification does not guarantee, however, that conceptual models that adhere to the specification will themselves be extensible. Depending on how a conceptual model is encoded (Section 4.3), significant remodeling may be necessary to support new tasks, methods, and

1. FilemakerPro is a flat-file database, meaning that data are stored in named variables that are themselves stored without any higher-order organization.

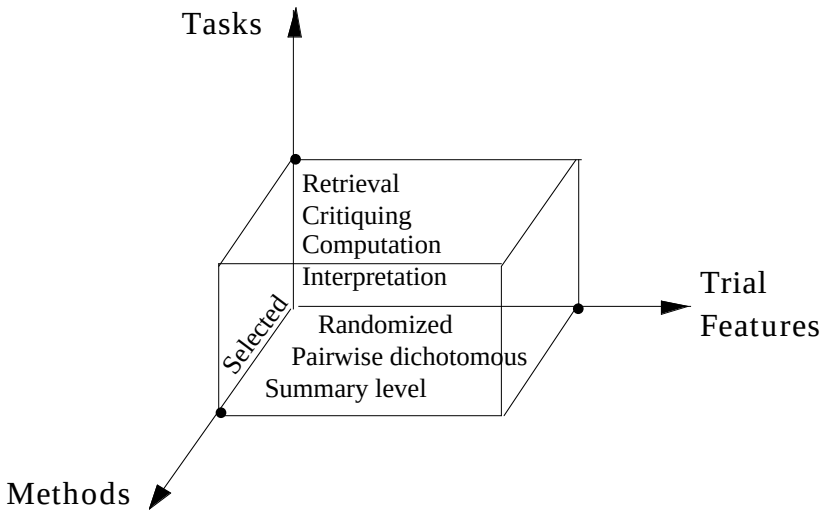

Figure 5.1. Clinical-trials modeling space. Modeling the core conceptual model of clinical trials involved choosing which trial features to model to support which tasks and which methods. No ordering of the choices along the axes is implied by this schematic.

domains. There are two additional axes not shown in Figure 5.1: the user axis and the purpose axis. For the design specification of the clinical-trials core conceptual model, the target user is an evidence synthesizer, and the overall purpose is evidence-based medicine. Ideally, both design specifications and conceptual models should also be extensible along these axes, but for some domains, new users or new purposes will require significantly different conceptualizations of the core concepts. For example, what we would consider the core concepts of the food domain would differ based on whether our purpose is nutritional analysis of foods or the representation of a gourmand's travelogue.

The remainder of this section describes the tasks, methods, and trial-features axes. Section 5.1.2 details the rationale for the modeling choices that I made along each of these axes.

5.1.1.1 Tasks

In the competency-decomposition approach that underlies the design specification, the task objectives of the specification are called **competencies**. Evidence synthesis is the highest-level competency of the trial-bank system; Figure 5.2 shows the decomposition of evidence synthesis into lower-level competencies. The second-level competencies correspond to the four core tasks of evidence synthesis: trial retrieval; trial critiquing;

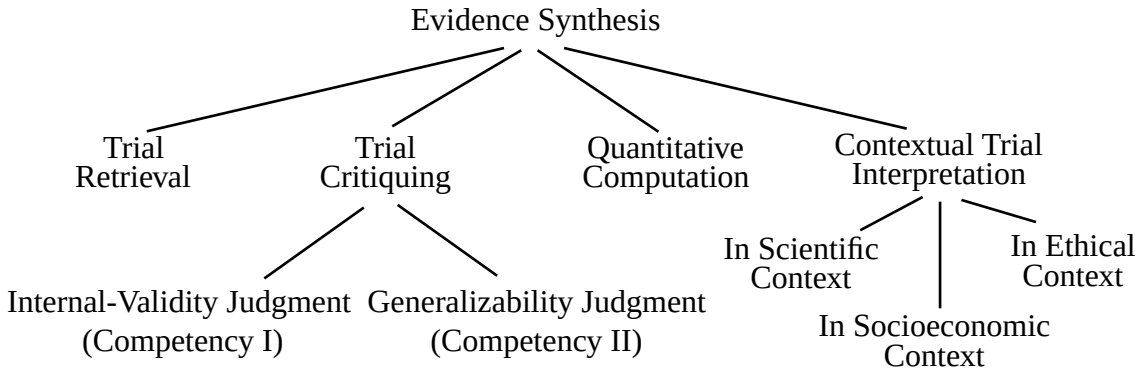

Figure 5.2. Hierarchy of target competencies of the trial-bank system. The core conceptual model is designed such that the trial-bank system can support these tasks.

quantitative computation of trial results; and interpretation of a trial in its scientific, socioeconomic, and ethical context. The competency decompositions for these tasks (Sections 5.1.2.1 to 5.1.2.4) are based on the clinical-trials interpretation literature, and on my experience as a meta-analyst. Because the task of trial critiquing is particularly large and complex, I discuss the competency decompositions for the judgment of internal validity and generalizability separately, as Competency I and II respectively.

5.1.1.2 Methods

Methods are sequences of action that lead to the accomplishment of tasks. There are two classes of methods supported by the design specification: minimal methods, and maximal methods. For the tasks of trial retrieval and quantitative synthesis, the design specification supports only the minimal methods for accomplishing those tasks. For the remaining two core tasks — trial critiquing and contextual interpretation — it supports a maximal method. The choice of methods to be supported by the design specification determines entirely the trial features that will be required by the design specification.

Minimal methods — The **minimal methods** for a task are those methods that require the fewest number of domain concepts to accomplish that task. For example, for the task of quantitative meta-analysis of dichotomous outcomes, all methods require at least a complete contingency table of the outcome of interest[2] from each trial. Several minimal

methods require no more trial information than that; examples are the Mantel-Haenszel method using odds ratios (Rothman, 1986), and the Peto method using odds ratios (Peto, 1977). In contrast, the hierarchical-Bayes method requires a distribution of prior belief in the effect being meta-analyzed, in addition to the contingency table (DuMouchel, 1983). The hierarchical-Bayes method is therefore not a minimal method for this task. Since a contingency table provides the minimal necessary and sufficient trial information for accomplishing quantitative meta-analysis of dichotomous outcomes, the contingency table — but not the prior distribution — is a data requirement in the design specification. The intent behind supporting only minimal methods is to specify models that make only minimal, but sufficient, ontological commitment. That is, conceptual models that support minimal tasks will include only the necessary and sufficient domain concepts for users to complete the target tasks; these models will not be overburdened with optional domain concepts.

Maximal methods — Despite moderate agreement in the clinical-trials community on which trial features are important to critique, no agreement exists on how to critique a randomized trial (Detsky, 1992; Greenland, 1994). In competency-decomposition terms, agreement exists on the subcompetencies for trial critiquing, but not on the methods. Because the agreement on the subcompetencies is only moderate, the design specification errs on the side of inclusiveness and supports all reasonable subcompetencies for the task of trial critiquing — the **maximal method**. The design specification also supports the maximal method for the core task of contextual trial interpretation, for analogous reasons: No standard methods exist for interpreting trials in their scientific, socioeconomic, and ethical context.

5.1.1.3 Trial Features

I restricted the domain of clinical-trials modeling to randomized trials because randomized trials yield the most internally valid evidence, and because their regular structure lends

---

2. A contingency table is also called a 2 X 2 table, for pairwise comparisons of dichotomous outcomes.

them to modeling. The design specification requires those randomized-trial features that we need to accomplish the target competencies using the designated methods. Examples of trial features required include SAMPLE-SIZE and INCLUSION-CRITERIA.

5.1.2 Competency Decomposition of the Four Core Tasks

The design specification comprises a hierarchical decomposition of the four core competencies, some justifications for the decomposition, and a catalog of the trial information required to accomplish the competencies (see Appendix A). Figure 5.3 is a schematic of the design specification. It shows that the four high-level competencies are decomposed into 30 subcompetencies and 44 subsubcompetencies. Not shown are the 162 data requirements that are derived from this decomposition.

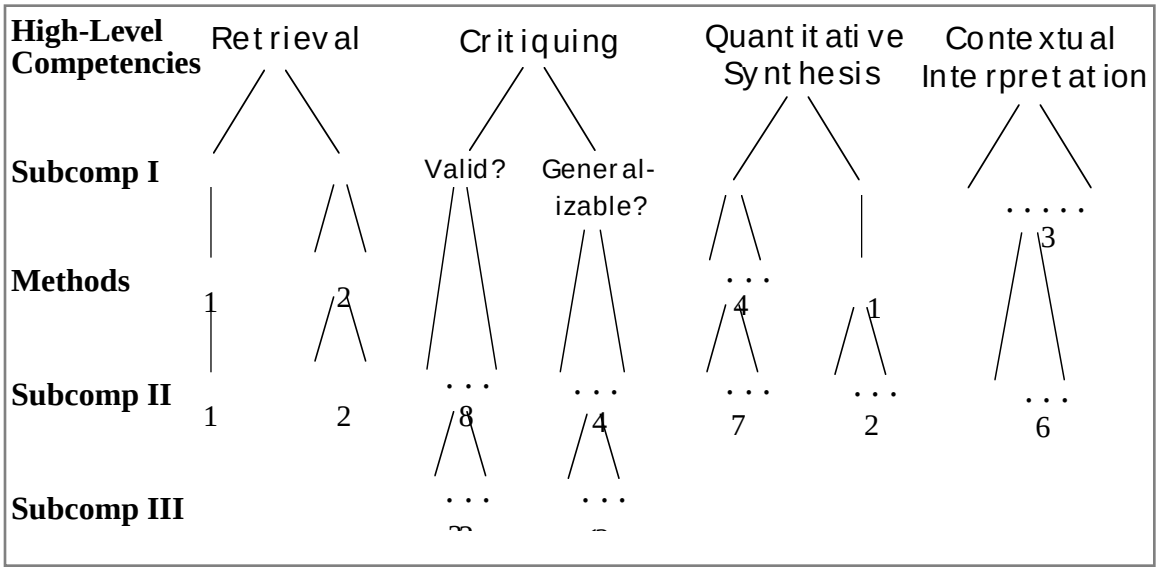

Figure 5.3. Schematic of the design specification. The four core tasks of evidence synthesis are the high-level competencies of the design specification. As described in Chapter 4, a competency is a task that a conceptual model is intended, or claims, to support. Each competency is decomposed into lower-level competencies. Where appropriate, the decomposition specifies the methods by which the competency will be achieved; the methods supported by the design specification determines entirely the design specification's method-associated subcompetencies and its required trial features.

5.1.2.1 Information Retrieval

The competency decomposition of the task of information retrieval (Table A.1 on page 182) is preliminary because information retrieval is not a central concern of my thesis work. However, accurate information retrieval is a core evidence-synthesis task, and we can use the competency-decomposition approach to sketch an initial analysis of the trial information needed for accomplishing this task.

The first step in trial retrieval is to capture the search query accurately (*I. Query Capture* in Table A.1). Because this competency is fulfilled by a trial-bank system, rather than by a clinical-trials core conceptual model, the design specification does not specify any trial-information requirements for this competency. The second step in trial retrieval is *query matching*. The design specification supports the minimal method of string matching for accomplishing the competency of query matching. For string matching, each of the instance terms in the core conceptual model should come from a controlled medical vocabulary — for example, the Unified Medical Language System (UMLS) (Lindberg, 1993).

For illustrative purposes only, Table A.1 shows how the competency-decomposition framework can associate with the competencies of a conceptual model not just data requirements, but also procedural requirements. I will not discuss procedural requirements further in this dissertation.

5.1.2.2 Judgment of Internal Validity

An internally valid trial is one whose findings reflect the true value of the outcomes of interest, rather than being systematically biased estimates of the outcomes. To determine internal validity, an evidence synthesizer must assess many details of a trial's design and execution. Over one-half of the data requirements of the design specification are associated with this competency (Competency I, Table A.2 on page 183).

Competency I.A. Was treatment assignment valid? — The main benefit of randomized trials derives from the elimination of treatment-selection bias by random assignment of

subjects to treatments. For a treatment assignment to be valid, the allocation sequence must be generated randomly, and the executors of the allocation must be blinded to the allocation schedule: An investigator should not know, for example, that the next patient will be given the placebo treatment, because the investigator may then subvert the randomization by controlling which patient is enrolled next. There are many ways to subvert randomization (Schulz, 1995), and the result of unblinded allocation is a bias toward exaggerated outcomes in favor of the experimental treatment (Schulz, 1995). Full descriptions of sequence generation and of the methods used to conceal allocation are necessary for determining the validity of a trial's treatment assignment.

Competency I.B. Was the treatment administration valid? — Treatment administration is valid if all subjects took their assigned treatment as intended, in exactly the same fashion, and completely. If only some of the subjects took the experimental treatment, then the observed treatment effect may underestimate the true treatment effect. If the subjects did not take the treatment as intended, or if the treatment differed across subjects, then the observed effects will not bear directly on the original trial hypothesis. Therefore, the design specification requires the details of the treatment administered — including any prespecified and ad hoc protocol deviations — and the details on treatment compliance.

Treatment blinding is another important aspect of treatment administration. Even though placebo pills may look and taste exactly the same as the experimental pill, subjects and other trial participants may still be able to guess which treatment had been assigned, and may change their behavior accordingly. For example, if a subject knows that he is taking a zinc lozenge rather than a placebo in a trial testing the efficacy of zinc in shortening upper-respiratory–infection symptoms, he may perceive his cold symptoms as less severe than he otherwise would if he believes zinc is efficacious for this purpose. The design specification therefore requires information on the method used to blind the subjects, providers, study nurses, and investigators to the subjects' assigned treatment, if such blinding is feasible. It also requires that a clinical-trials core conceptual model be able to capture information on the efficacy of treatment blinding (i.e., whether or not trial participants correctly guessed

the treatment assignment). An analogous argument justifies the data requirements for the method and efficacy of blinding trial participants to the trial's interim results.

Competency I.C. Were there any confounding cointerventions? — There are two prerequisites for a trial's observed results to reflect only the effect of the experimental versus the control intervention: (1) treatment assignment must be truly randomized, and (2) post-randomization treatment must be identical except for the assigned treatment. Competency I.A. addressed the first prerequisite; competency I.B. partially addressed the second. This competency addresses the other treatments that subjects in a randomized trial may receive in addition to their assigned treatment, treatments that may affect the trial's primary outcome. Information about these cointerventions includes the types and dosages of other drugs, for example, and the nature and frequency of follow-up visits. Therefore, the design specification requires details on these matters, and on the proportion of subjects in each treatment group that had each cointervention.

Competency I.D. Were the outcome definitions valid? — The outcomes of the trial should be defined clearly, especially with regards to whether they are primary or secondary, and a priori or post hoc. Outcome definitions should not change during the execution of the trial. A trial's primary outcome should also be closely related to that trial's primary hypothesis: If a drug is being tested for its ability to reduce heart attacks, then the primary outcome should be heart attacks, rather than an intermediate outcome such as the incidence of chest pain.

Competency I.E. Were the outcomes assessed in a valid manner? — The method by which outcomes are assessed can introduce bias in several circumstances. When outcome assessors are aware of the treatment status of the subjects, or when they are aware of the interim results of the trial, they may knowingly or unknowingly skew their observation of results. The training of the assessors may also be important for accurate outcomes assessment in some trials (e.g., neuroradiologists assessed the cranial CT scans for stroke in the SPINAF trial). Bias can also be introduced if the assessment method is not valid or is not

reproducible. For example, assessing whether a patient has had a heart attack by asking about that patient's rating of chest-pain intensity is neither a valid nor a reproducible heart-attack–assessment method.

Competency I.F. Are the outcome results valid? — One of the greatest threats to a trial's internal validity is incomplete followup of the subjects. Subjects who are lost to followup may be systematically different from those who remained in the study, such that the outcomes observed in the remaining subjects fail to reflect the true effect of the intervention. The larger the number of subjects that are lost to followup, the larger might be this bias. Other threats to the validity of the reported results include inappropriate transformations or parameterizations, and inappropriate use of statistical tests and approaches (e.g., censoring). All results must be clearly and completely described, including denominators for all percentage outcomes, precision estimates for summary descriptors, and exact *p* values for statistical tests.

Competency I.G. Was the trial design and conduct valid? — Elements of good trial design and execution include the specification of a primary hypothesis and outcome, the performance of a power and sample-size calculation for the primary hypothesis, the a priori specification of subgroup analyses, appropriate and unbiased trial monitoring and termination, and careful documentation of any unanticipated protocol changes. Other elements of good trial design, such as complete followup of subjects, are discussed under other trial-critiquing competencies.

Competency I.H. Was there an outside source of bias? — The execution and reporting of randomized trials are subject to many external sources of bias, financial and otherwise. Examples of these sources of bias include the interests of the funders, the medical specialties of the trial investigators, and the interests of the publishers. Unbiased, statistically sound trial monitoring is a defense against these external biases in trial execution.

5.1.2.3 Judgment of Generalizability

The competency of judgments about the generalizability of a trial's results (Competency II, Table A.3 on page 192) is decomposed into four subcompetencies. The overall objective is to determine whether or not the study's results can be extended properly to a particular patient in a particular situation outside of the study.

Competency II.A. Were the patients similar to the target population? — A subtle problem in judging the generalizability of a trial's results is to determine the extent of selection bias in the enrolled subjects. If a trial disproportionately enrolled subjects that are unrepresentative of the norm, the generalizability of its results may be threatened.

How is selection bias to be assessed? This competency decomposition requires that a clinical-trials core conceptual model be able to capture information on the method used to recruit patients; the inclusion and exclusion criteria; the number of patients screened, eligible, enrolled, randomized, and analyzed; and the baseline rate of the target condition in the enrolled subjects. It also requires that the model be able to capture the number of patients excluded on the basis of each exclusion criterion. This information is sometimes misleading, however: Because subjects are often excluded on the basis of the first exclusion rule they satisfy, the number excluded for a particular reason depends strongly on the order in which the rules are applied. Furthermore, if the rules are not applied in a fixed order, the number excluded for any particular reason will be even less reflective of the characteristics of the excluded subjects. Nevertheless, many users of the trial-bank system will seek out this trial information, and this information is therefore included in the design specification.

Competency II.B. Is the setting comparable? — To determine the generalizability of a trial's results, we need information on when and where that trial was conducted. For example, trials of myocardial-infarction treatment from before the thrombolytic era may not now be applicable. Results of trials conducted in inner-city, county hospitals may not generalize to wealthy, suburban, capitated populations, because of differences in the patients,

in ancillary treatments such as those delivered by paraprofessional staff, and in socioeconomic factors.

Competency II.C. Is the intervention available locally? — A trial's results are relevant to a local practice only if the tested intervention is available locally. Thus, the trial's interventions must be described fully, including the training of the operators if the intervention was a procedure, the cointerventions taken by the subjects, the frequency and nature of follow-up care, and the subjects' compliance with assigned treatment. If compliance with the intervention is higher in the trial than can be expected locally, for example, then the observed result will not be completely generalizable.

Competency II.D. Are the study outcomes of local interest? — A trial's results are relevant to practices outside of a study only if the outcomes measured are relevant to those practices. For example, the marginal efficacy of drugs to prevent maternal transmission of HIV relative to zidovudine is not a relevant outcome for countries that cannot afford zidovudine for any of its women.

5.1.2.4 Quantitative Synthesis

The quantitative-synthesis competency comprises two subcompetencies: (1) calculation of a summary statistic for pairwise comparisons of dichotomous outcomes, and (2) performance of quantitative meta-analysis of the summary statistics from several trials. For both of these subcompetencies, the only trial information required is a complete 2 X 2 contingency table (Table A.4 on page 195). The design specification for this competency also specifies the procedural requirements, but, as they are in the competency decomposition for information retrieval (Section 5.1.2.1), the procedural requirements are listed for illustrative purposes only.

5.1.2.5 Contextual Interpretation

Just as for the task of trial critiquing, there are no standard methods for the task of interpreting a trial in its proper scientific, socioeconomic, and ethical context. Thus, the design specification supports the maximal method for contextual interpretation, by requiring that

the clinical-trials core conceptual model store references to the most important types of contextual information (Table A.5 on page 196). Examples of the requested types of information include a clinical background in the trial topic, references to letters to the editor, references to decision models that incorporate evidence from the trial, information on ongoing related trials (e.g., pointers to trial registry entries), and human-subjects approval. This decomposition is preliminary.

5.1.2.6 Classes of Trial Features not Required

In the competency decompositions of the four core evidence-synthesis tasks, there were several classes of trial features that I chose not to require in the design specification. These trial features were excluded based on tradeoffs between the achievement of greater conceptual coverage and the construction of finite conceptual models for everyday use. None of these excluded trial features are necessary for accomplishing any of the competencies of the design specification. Representative examples of each of the classes of excluded trial features are given below.

1. *Trial feature not commonly required* — The design specification does not require details about the recall of medical devices, because devices are recalled too infrequently to justify including this information in a clinical-trials core conceptual model.

2. *Trial feature not uniformly defined* — The meaning of the terms *dropout* and *withdrawal* are not clear in common usage. Instead of codifying any particular definition of these terms, I chose instead to reformulate the underlying trial concepts into the following trial features: (1) those subjects who did and did not complete their assigned treatment; and (2) those subjects who did and did not have their outcomes assessed. The design specification does not include the terms *dropout* or *withdrawal*.

   Two other examples of common terms with unclear common usage are *primary hypothesis* and *primary outcome*. In Ocelot-CCM, these terms are used, but they have a restricted meaning (Section 5.2.2.2).

3. *Information is of unclear use or benefit* — In trials that trigger a statistical stopping rule, some subjects stop their assigned treatment prematurely, whereas other subjects stop according to the protocol. It is unclear how information on the numbers of subjects who stopped prematurely and by protocol could be used in data analysis, so these trial features are not required by the design specification.
4. *Trial feature reflects incorrect conceptualization of randomized trials* — As discussed in Chapter 2, randomized trials are not always designed, conducted, reported, or analyzed as well as they could be. The design specification does not perpetuate the poor practices of not designating hypotheses and data analyses as either a priori or post hoc, for example.

Should these judgement calls on what to include in the design specification prove inappropriate, the design specification can be extended easily to correct them.

5.1.3 Extensibility of the Design Specification

The design specification's modularity facilitates incremental extensions. We can add new competencies, methods, and method-associated subcompetencies without having to change previous competency decompositions. For example, many regulatory agencies worldwide have adopted the International Conference on Harmonization randomized-trial–reporting standards (ICH, 1995). Suppose that we extend the design specification's competencies to include the reporting of a trial to a regulatory agency. We can add this competency, its lower-level competencies, and its associated trial-features requirements to the design specification without having to change any of the current competency decompositions. Since many of the trial features required by the ICH reporting standard are already in the design specification, many of the new data requirements will overlap with previously specified requirements, and the design specification will have been extended easily to support the task of regulatory-agency reporting.

Other directions for extending the design specification include new methods for trial retrieval and quantitative computation, users other than evidence synthesizers, and purposes other than evidence-based medicine.

5.2 Implementation

I implemented a conceptual model called **Ocelot-CCM** according to the design specification presented in Section 5.1. **Ocelot-CCM** is encoded in an object data-definition language (DDL) for reasons explained in Section 4.3. I seriously considered three object-based (or frame-based) knowledge-representation systems: the Ontolingua, Protégé, and Ocelot systems. I chose the Ocelot system because it was the most user-friendly, and because its models were the easiest to understand.

Ontolingua — The Ontolingua language (Ontolingua, 1996) is designed expressly for encoding shared ontologies such as the clinical-trials core conceptual model; unfortunately, however, the Ontolingua Editor is extremely cumbersome to use. It is difficult both to encode a model and to browse a completed one using the Ontolingua Editor, especially for a user who is not conversant with first-order logic.

Protégé — The Protégé knowledge-engineering system (Musen, 1993) includes an ontology editor (Maître) coupled with a graphical viewer (Gennari, 1993). The strength of the Protégé system lies in its ability to generate a data-entry interface[3] from a frame-based ontology automatically. The 1993 version of Maître was, however, ill-suited to the early phase of conceptual modeling, because the model under construction could not be edited in the graphical viewer directly, and because the system was slow.

Ocelot — The GKB Editor (Karp, 1995) is an intuitive graphical interface for editing conceptual models in the Ocelot knowledge-representation system. A conceptual-model

3. A data-entry interface is also known as a **knowledge-acquisition interface**, for acquiring instances of an ontology.

designer can easily make changes directly in the graphical viewer, which can display the model in multiple perspectives that highlight the relationships between and among frames (or object) and slots (or attributes). In addition, the GKB Editor communicates with the Ocelot system via the Generic Frame Protocol (GFP), which is a syntax for sharing ontologies among frame-based knowledge representation systems (Karp, 1995).The translation of Ocelot-CCM into several other knowledge-representation systems (e.g., LOOM, Ontolingua, and THEO) is therefore easy using GFP.[4] The combined GKB Editor and Ocelot system offered a user-friendly and versatile system for rapid modeling of a complex domain.

The data-representation model in Ocelot is essentially identical to the data model specified in GFP. A general description of the object data-model was presented in Section 4.3.4.

1. Objects are called **frames.** Attributes of frames are called **slots**. Attributes of slots are called **facets.** Slot values can be restricted to defined character-strings, Boolean values, numbers, or Lisp S-expressions, and can be annotated with character strings. Frames and slots are both stored as frame data structures, all of which have special slots for documentation. Slots of frames can be instantiated with other frames, resulting in **compound frames**.
2. Child frames inherit the properties of their parent frames. Multiple inheritance is supported (see GFP documentation (Karp, 1995) for details).

Ocelot's expressivity characteristics are typical of object data-definition languages (Section 4.3.4). Ocelot cannot represent negation, disjunction, or uncertainty. Despite these limitations in expressivity, Ocelot's frame-based knowledge-representation language was sufficiently expressive for encoding the clinical-trials core conceptual model (see Section 7.2 for the evaluation), yet the language is reasonably compact and understandable. In addition, a conceptual model encoded in this language can be re-encoded in other object

4. Since the time I chose Ocelot, both the Ontolingua and Protégé systems have also implemented a GFP communications interface.

DDLs without extreme difficulty. This adaptability is important, because it is likely that a commercial-grade core conceptual model will be written in an object-based language such as C++ or Java, rather than in Ocelot.

Other limitations of expressivity that are specific to Ocelot-CCM include its shortcomings with temporal, functional, and procedural representation.

Temporal representation — Ocelot-CCM's representation of time is rudimentary. Ocelot-CCM can represent sequences of actions over time. For example, the model can represent a drug intervention as a loading dose for 1 week followed by maintenance therapy for a year. It can also represent externally defined (e.g., 8/15/96) or internally defined (e.g., since randomization) timepoints, durations, and schedules (e.g., every 3 months for 1 year). However, Ocelot-CCM does not incorporate advanced temporal modeling (e.g., overlapping intervals, concurrent actions), and this limitation leads to two major implications: (1) Ocelot-CCM cannot support temporal reasoning about trial execution; and (2) it cannot elegantly represent crossover trials, because doing so requires temporal sequencing of large-granularity actions (i.e., treatment assignment, administering an intervention to a defined population, measuring outcomes in a defined population). Upgrading of Ocelot-CCM's temporal representation is highly desirable but will be difficult.

Representation of mathematical functions — Ocelot-CCM cannot capture mathematical functions, such as arbitrary nonparametric distributions. A consequence of this limitation is that Bayesian interpretation of randomized trials cannot be supported by Ocelot-CCM: Neither prior nor posterior distributions can be captured, and Bayesian design methods can be described only in unstructured, free-text entries.

Procedural representation — Ocelot-CCM does not model the execution of a trial declaratively. For instance, there is no explicit statement that subjects receive a treatment only after they have been assigned to one. Therefore, Ocelot-CCM cannot be used to simulate a trial. Much procedural knowledge must be added to Ocelot-CCM before the model

can simulate a hypothetical group of subjects following a hypothetical trial protocol from trial inception to outcomes assessment.

5.2.1 Overview of the Structure and Content

The root frame for Ocelot-CCM is the frame TRIAL, which has slots DEFINITION, DOCUMENTATION, and SYNONYMS. All the remaining 127 Ocelot-CCM frames are children of TRIAL, and inherit the three documentation slots. The complete class hierarchy of Ocelot-CCM is given in Appendix B. For discussion purposes, we can partition the trial features into the following groups:

1. *Administration* — details of a trial's administration (e.g., funding, investigators), and its publications
2. *Design* — a trial's hypotheses, sample-size calculations, analytic methods, and protocol (the sequence of actions that constitute a trial)
3. *Subjects and recruitment* — the identification, recruitment, and enrollment of subjects, and their clinical characteristics and followup
4. *Treatment assignment* — details of the randomization and the allocation-concealment process used by the investigators to assign treatments to subjects
5. *Interventions* — the drugs, procedures, or devices that are administered to subjects in a controlled fashion during the course of a trial
6. *Followup* — details of the followup of enrolled patients from the start of the trial to its end and beyond, including the follow-up methods
7. *Outcomes and measurements* — definition of the events or variables (e.g., kidney failure) postulated to be causally related to the administration of an intervention, and details of its measurement
8. *Results* — the observed effects of the intervention on the outcome variables in the subjects

Ocelot-CCM is a conceptual model of clinical trials, rather than one of clinical medicine. As such, it does not model the domain of medicine; no frames relate to the different forms of heart disease, for example, or to the types of drugs for treating chronic pain. If a trial-bank system is to be interoperable, however, it is imperative that all trial banks share not only the semantics of clinical trials, but also the semantics of clinical medicine. For example, the clinical term *arrhythmia* must be standardized across trial banks for complete interoperation. Unfortunately, none of the many controlled clinical vocabularies that currently exist can be considered a worldwide standard (Cimino, 1996). Thus, if trial banks use different controlled vocabularies, the trial-bank system as a whole will lack a shared semantics of clinical medicine.

The UMLS Metathesaurus® provides one approach to integrating multiple, controlled, clinical vocabularies. The Metathesaurus integrates over 30 biomedical vocabularies and classifications (e.g., MeSH, SNOMED, ICD-9) by cross-linking their terms with UMLS terms and with one another. Therefore, if UMLS is the controlled vocabulary for the clinical-trials core conceptual model, then a UMLS term such as *arrhythmia* could be shared semantically with trial banks that use any of the Metathesaurus vocabularies. This sharing would extend even to trial banks that use the 1996 French version of the NLM's MeSH vocabulary, for which *arrhythmia* is synonymous with *trouble rythme cardiaque*. The incorporation of UMLS into Ocelot-CCM is a high priority for future work.

5.2.2 Trial-Feature Groups

Sections 5.2.2.1 to 5.2.2.8 highlight the trial features modeled in each trial-feature group. The examples used throughout this section come from the SPINAF trial (Ezekowitz, 1992)[5] and from the CHF-STAT trial (Singh, 1995).[6] The rationale for why Ocelot-CCM

5. The SPINAF trial was a randomized, placebo-controlled trial that examined the efficacy of warfarin (an anticoagulant) for preventing stroke in patients with nonrheumatic, nonvalvular, atrial fibrillation.

6. The CHF-STAT trial was a randomized, placebo-controlled trial that examined the efficacy of amiodarone (an antiarrhythmic drug) for preventing death in patients with congestive heart failure.

includes these trial features is detailed in the discussion of the design specification (Section 5.1.2).

5.2.2.1 Administration

The administrative trial features modeled by Ocelot-CCM include investigators; investigator groups (e.g., The SPINAF Investigators); names and members of trial committees; names and characteristics of study sites; and details about the ethics approval, funding, and publications of the trial. For example, the source and type of trial funding (e.g., government or industry) is captured, as well as the funder's right, if any, of veto over the trial's reporting. The model includes complete citations (e.g., journal name, year, and pages) to official trial publications, and to related publications, such as editorials, letters to the editor, and systematic reviews that referred to this trial. For official trial publications, the high-level structure of the text articles (e.g., abstract, background, discussion) is included in Ocelot-CCM as well.

5.2.2.2 Design

Ocelot-CCM captures extensive information on a trial's design. The modeling of several of these design features is discussed under other trial-feature groups; here, we discuss the modeling of a trial's protocol, hypotheses, and power and sample-size calculations.

Protocol — A protocol is a sequence of actions over time. Ocelot-CCM can capture randomized trials of two or more arms, prospective cohort studies, and — because it can represent delays between enrollment and randomization, and between randomization and intervention — the model can also represent run-in and wash-out trial designs. Because its temporal modeling is rudimentary, however (see page 115), Ocelot-CCM cannot represent cross-over protocols, in which subjects take first one intervention and then cross over to another. The model cannot explicitly represent Zelen's randomization[7] either, because

7. In Zelen's randomization, subjects are asked to provide informed consent after they have already been randomized (Zelen, 1979). Subjects therefore know the treatment to which they have been assigned before they consent to enter the trial.

Ocelot-CCM does not explicitly model when treatment assignment occurs in relation to informed consent. Improvements in Ocelot-CCM's temporal modeling will allow more complete and explicit modeling of protocols.

Hypotheses — Ocelot-CCM requires that all trial hypotheses be designated either primary or secondary, and either a priori or post hoc. Every trial must have one and only one primary hypothesis, and the primary hypothesis must be phrased in terms of the primary outcome of the trial. For example, if the hypothesis of a trial is that amiodarone reduces sudden death in patients who have heart disease, then the primary outcome should be sudden death. These restricted meanings of *primary hypothesis* and *primary outcome* are not standard within the clinical-trials community, but are sufficiently common to codify in a core conceptual model for interoperating trial banks.

Power and sample-size calculation — A trial's primary hypothesis and outcome should be the basis for that trial's power and sample-size calculations. Ocelot-CCM captures the following attributes of a trial's power and sample-size calculation: the expected baseline rate of the primary outcome, the threshold difference between the outcomes of the control and experimental groups, the power, the alpha and its number of tails, the sample-size–calculation method, the target sample size, the required sample size, and a justification for any differences between the target and required sample sizes. For example, the SPINAF trial's target sample size was 556 subjects for detecting, with 80-percent power, a decrease in cerebral infarction from 5 percent to 2 percent after 3 years on warfarin compared to placebo, at a two-tailed alpha of 0.05. Ocelot-CCM does not represent post-hoc power calculations for nonsignificant observed effects, for reasons explained in Section 7.1.2 (page 152).

5.2.2.3 Subjects and Recruitment

In Ocelot-CCM, patient-eligibility rules are represented as logical expressions (see Appendix E.1). The inclusion rule from CHF-STAT is represented as follows:

```
(>= 10 PVCs per hour on a 24 hour Holter)
AND (prior history of
   ((rest dypsnea) OR (dypsnea with minimal exertion)
    OR (paroxysmal nocturnal dypsnea))
AND ((left ventricular internal dimension by echocardiogram >=0.55mm)
   OR (cardiothoracic ratio > 0.5 on chest X ray))
AND ((ejection fraction by radionucleide multiple gated scan <= 40%)
   OR (ejection fraction by cardiac catheterization <= 40%)))
```

If the clinical terms in this eligibility rule are from the same controlled vocabulary as that used by a computer-based patient record, then an expert system could use this rule to identify eligible patients from the patient records automatically.

Ocelot-CCM represents the recruitment, enrollment, and followup of a trial's subjects as nested subsets of groups, as shown in Figure 5.4. Each of the 11 subject groups can be described by its clinical characteristics, and by the numbers of subjects excluded or lost to

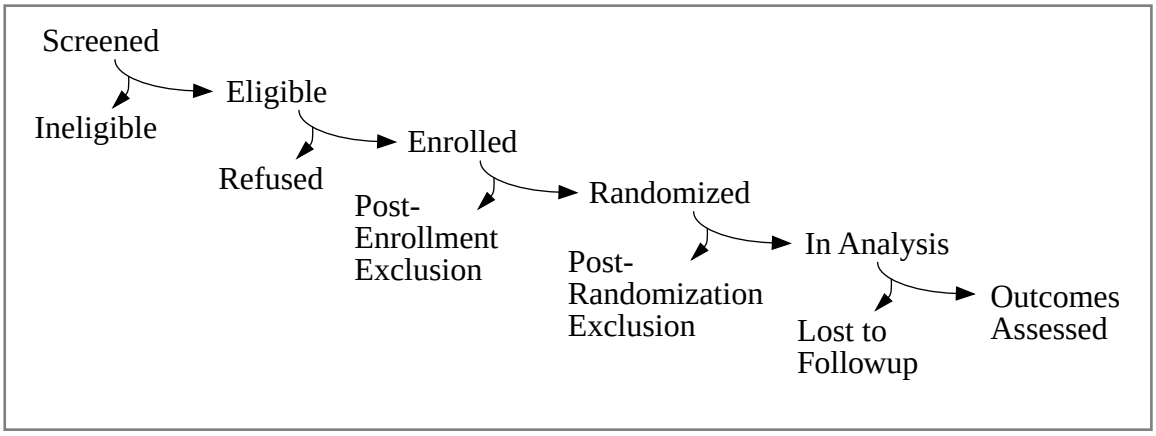

Figure 5.4. Subject groups modeled as nested subsets. A trial's enrolled population constitutes only a subset of those subjects who were screened; subjects whose outcomes are analyzed constitute an even smaller subset of the screened subjects.

followup for particular reasons. The RCT Presenter system (Chapter 6) uses this nested-subset modeling of subject groups to generate automatically a flowchart of subject recruitment and participation (Figure 6.8 on page 135) as recommended by the CONSORT group (Begg, 1996).

5.2.2.4 Treatment Assignment

The treatment-assignment process consists of two steps: (1) generating a random allocation schedule, and (2) using that schedule to allocate subjects to the interventions. Ocelot-CCM captures descriptions of the generation of the random sequence (e.g., the name of a random number generator), as well as descriptions of any blocking or stratification in treatment assignment. Ocelot-CCM also captures the method by which allocation is concealed (e.g., phoning a central laboratory for the assigned treatment). Evidence on the efficacy of randomization is captured by the clinical characteristics of the enrolled subjects; if the subjects were indeed assigned randomly, then their characteristics should be equally distributed across treatment groups.

5.2.2.5 Interventions

Drug interventions — the most common intervention tested in clinical trials — are extensively modeled in Ocelot-CCM. Drug generic and trade names, manufacturers, dosages, schedules, formulations (e.g., PO or IV), and adjustments are represented. Dosage schedules can be single-stepped or multiply stepped (e.g., in the CHF-STAT trial, amiodarone is given as a 2-week loading dose followed by 50 weeks of high-dose maintenance). Alternatively, dosages can be captured as a titration to a target goal (e.g., in the SPINAF trial, warfarin is given in sufficient doses to maintain an INR of 1.2 to 1.5 times normal). For placebo interventions, Ocelot-CCM can capture the similarity of the placebo to the experimental intervention (e.g., that the pills look the same), in addition to a justification for why a placebo rather than an active control was used. Cointerventions also are modeled. In contrast to the extensive modeling of drugs, Ocelot-CCM models surgical, procedural, medical-device, and behavioral interventions mostly as textual descriptions only.

For each intervention and cointervention, the percentage of each subgroup who received it is captured. In addition to this information on who was administered what interventions, Ocelot-CCM also captures information on how patients, providers, study nurses, and investigators were blinded to the assigned intervention; on how compliance was encouraged and checked; and on what blinding and compliance were achieved by the trial.

5.2.2.6 Followup

As discussed in point 2 on page 111, the commonly used terms *dropout* and *withdrawal* are not modeled in Ocelot-CCM. Instead, for each group and subgroup of enrolled subjects, Ocelot-CCM captures the numbers who did and did not have their outcomes assessed, and the numbers who did and did not complete their assigned treatment. The model also captures the mean length of followup, in absolute time (e.g., 1.8 years), and in person-years for each subgroup and each outcome. For example, in the SPINAF trial, the mean followup in the warfarin group was 1.8 years for the outcome of cerebral infarction, with a total of 456 person-years of followup. In the placebo group, the respective numbers were 1.7 years and 440 person-years. Statistical handling of loss to followup can be described in Ocelot-CCM.

5.2.2.7 Outcomes and Measurements

Ocelot-CCM captures information on the definition, assessment method, and statistical analysis of each outcome of a trial. The definition is a textual description. The assessment information includes how, by whom, on whom, and when the outcome was assessed, and what was the blinding of the outcome assessors to the treatment assignment and to the interim results of the trial. The statistical-analysis information includes the names of the analyses conducted, and, if applicable, the censoring approach used. Outcomes must also be designated as being primary, secondary, or ancillary, and as either a priori or post hoc.

5.2.2.8 Results

Ocelot-CCM can capture both descriptive and analytic statistics of observed trial outcomes. Descriptive results can be reported as real numbers, means and standard deviations, medians, minimum and maximum ranges, or Kaplan–Meier life tables. All measurement units must be reported, and denominators must be stated clearly for all percentage results. For analytic statistics, names of statistical tests used (e.g., chi-square) must be identified, and results can be reported with their 95-percent confidence intervals or with their standard errors. Ocelot-CCM also models regression equations — linear, logistic, and Cox proportional hazards — as variable names and their associated

coefficients, standard errors, and *p*-values as appropriate (based on Lang and Secic (Lang, 1997)). Ocelot-CCM cannot store graphical files; therefore, it cannot capture results that are in the form of figures or pictures.

5.2.3 Descriptive Statistics of Ocelot-CCM

As I iteratively re-modeled Ocelot-CCM during testing and evaluation, its size shrank. The number of frames decreased from 219 to 128, and the number of unique slots decreased from 532 to 430. This shrinkage implies that the model is more efficient at representing clinical trials than before, but I have not formally evaluated Ocelot-CCM's structural characteristics. Each frame has an average of 6.7 nondocumentation slots, of which 27 percent take other frames as instances. The maximum depth of the class hierarchy is 5, and 9 percent of the frames have multiple parents.

5.2.4 Extensibility of Ocelot-CCM

It would be most desirable if we could add new trial features to a clinical-trials core conceptual model without having to change the existing modeling. This extensibility would minimize the need to propagate modeling changes from the shared model to the trial banks. A conceptual model's extensibility depends on the properties of the language in which it is encoded, and on the structure of the model itself. Thus, not all models that adhere to the same design specification have the same extensibility characteristics. For Ocelot-CCM, several new trial features can probably be added without major changes to the existing modeling.

Extension to other outcome types — We can add new outcomes types to Ocelot-CCM by defining new child frames of the frame OUTCOME-VARIABLE-TYPE. Instances of these new frames will inherit all the generic properties and relationships that have already been modeled for other outcome types, and no existing modeling needs to be changed. Examples of new outcome types to extend include costs, preference-based quality-of-life measures (utilities), functional-status measures, and genetic-sequencing and mapping outcomes.

Extension to other clinical domains — As discussed on page 117, the ability of Ocelot-CCM to express clinical terms is determined solely by the clinical vocabulary that we use to instantiate Ocelot-CCM's frames. Ideally, we would use Ocelot-CCM in conjunction with a controlled clinical vocabulary such as the UMLS. If so, then the clinical domain of Ocelot-CCM will be the clinical domain of UMLS. We do not need to make any changes to Ocelot-CCM itself to incorporate the UMLS or any other controlled clinical vocabulary.

Extension to individual patient-level data — The set-based representation of subject groups in Ocelot-CCM makes it trivial to represent individual subjects: A subject is simply a subgroup of size 1. Thus, Ocelot-CCM can capture individual subject characteristics, followup, and outcomes using the modeling for subject groups. A core conceptual model that captures patient-level data is needed for patient-level meta-analysis of randomized trials, for reporting trials to regulatory agencies, and for simulating trials.

Extension to separate representations of design and execution — Clinical trials are not always executed in the way that they were designed, and protocol deviations can threaten a trial's internal validity. Therefore, Ocelot-CCM should model both a trial's intended and executed protocol. Ocelot-CCM can store both protocols separately; alternatively, Ocelot-CCM can store only the intended protocol and a list of protocol deviations, and then compute the executed protocol from them. The grounds for the latter approach are in Ocelot-CCM. The addition of temporal modeling to Ocelot-CCM will facilitate either approach to implementation of this extension.

Extension to other study types — Just as the design specification for the clinical-trials core conceptual model should be extensible to new tasks, methods, and domains (Section 5.1.3), so too should Ocelot-CCM ideally be extensible beyond randomized trials. At least two types of nonrandomized studies could well be captured by Ocelot-CCM without significant remodeling.

1. *Outcomes research with instrumental variables* — In instrumental-variables (IVs) analysis of large observational databases, IVs are used to pseudorandomize subjects, and the effect of an intervention is analyzed as it is in a traditional, pro-

spective, randomized trial. Therefore, Ocelot-CCM can probably represent this study type if we just add IV-based methods to the modeling of treatment assignment.

2. *Meta-analysis of randomized trials* — Meta-analyses are highly structured studies; randomized trials are their units of analysis. As has been done for randomized trials, an international group of researchers have propounded reporting standards for meta-analyses (Cook, 1995). We can define a conceptual model of meta-analysis that follow these standards, with the constituent trials of each meta-analysis described by Ocelot-CCM. The combined conceptual model can then interoperate both meta-analysis banks and trial banks (see discussion on the role of trial banks in an evidence-based informatics infrastructure, in Section 1.3).

5.3 Summary

This chapter presented the centerpiece of my thesis work: the design specification for a clinical-trials core conceptual model, and its implementation as Ocelot-CCM. I used the competency-decomposition approach to relate explicitly tasks, methods, and trial features to one another, and to the encoding of the core conceptual model. The result of this approach is a principled design specification that maps the domain knowledge of clinical-trials interpretation to the abstract requirements for a clinical-trials core conceptual model. Chapter 6 describes the application of this abstract modeling work to the construction of a concrete trial-bank–presentation system.

Chapter 6

The RCT Presenter System

The **RCT Presenter** system exemplifies two major components of a trial-bank system: (1) a trial bank; and (2) an interface for browsing a trial bank. With RCT Presenter, a user can browse the contents of a knowledge base of randomized trials over the web. In this chapter, I describe the architecture of RCT Presenter. In addition, I present the results of a pilot evaluation in which health-services researchers used RCT Presenter to critique a randomized trial. This experience with building and using RCT Presenter offers lessons on the opportunities and challenges for web-based trial-bank publishing.

6.1 Design Goals for RCT Presenter

RCT Presenter is designed to help evidence synthesizers critique a set of trials soundly and thoroughly. It does so by providing them with all the trial information that they need — in hypertext, in linear form, or as summary tables of attributes across trials. The organization of information in the interface was designed to correspond to the way that evidence synthesizers commonly think about randomized-trial information.

RCT Presenter neither supports advanced queries about the contents of its knowledge base, nor supports the entering of trials into the knowledge base. It presents information

about randomized trials in hypertext and in HTML (Hypertext Markup Language) tables only. No multimedia and no Java applets are used. The reasons for this simplicity are that a trial-bank–browsing interface was not central to this dissertation, and that a simpler browsing interface would be less likely to confound the evaluation of the system's usefulness.

6.2 Architecture of RCT Presenter

The system architecture of RCT Presenter follows the client–server model (Figure 6.1).

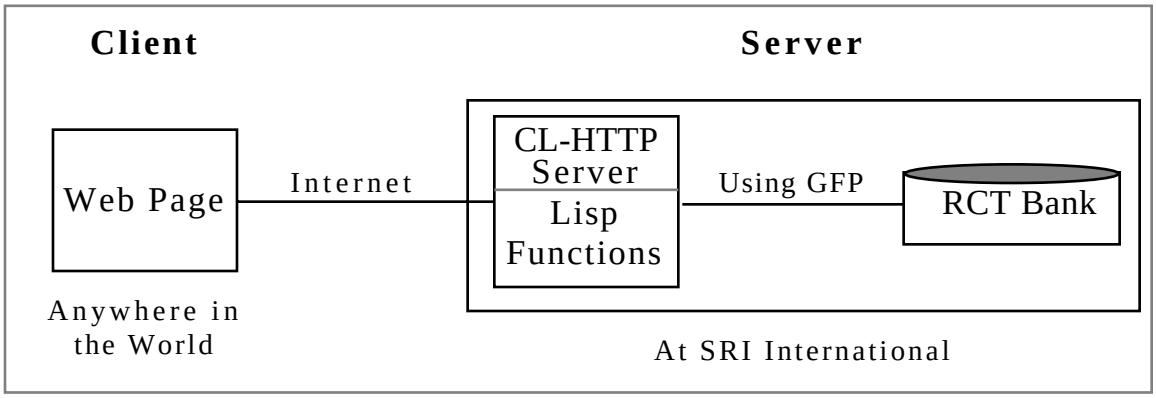

Figure 6.1. RCT Presenter system architecture. The RCT Presenter architecture follows the client–server model. The knowledge base (RCT Bank) resides on a Unix server at SRI International. To browse RCT Bank, a user accesses the RCT Presenter URL (Uniform Resource Locator) with a web browser over the Internet. The system thus can be accessed from anywhere in the world via the most popular operating systems (e.g., Macintosh, Windows, and Unix).

The client in the RCT Presenter system is an Internet web browser — for example, Netscape Navigator® or Internet Explorer®. The server is a freeware HTTP server[1] called CL-HTTP (Common Lisp Hypermedia Server) from the Massachusetts Institute of Technology (Mallery, 1997). The server runs on an UltraSparc Unix workstation at SRI International.

1. An HTTP server is a program that responds to Internet messages that adhere to the Hypertext Transfer Protocol (HTTP), the standard protocol for web-based messaging.

When the server is running, CL-HTTP listens on an internet port for HTTP requests, which are submitted as URLs to the server. For example, *http://mission.ai.sri.com:8000/tb/trial-bank.html* is a request to port 8000 of a machine named *mission.ai.sri.com* to return a web page called */tb/trial-bank.html*. Once an HTTP request is received by the server, a set of Lisp functions that incorporate CL-HTTP functions respond to the request by generating entire web pages dynamically from scratch. For example, if the server receives a request *http://mission.ai.sri.com:8000/tb/ProtocolOverview?2*, then it calls the function *write-protocol-overview* to compute an HTML page called Protocol Overview

```
;;;*** Function to export Protocol Overview window ***

(http:export-url #u"/tb/ProtocolOverview?"
     :search
     :response-function #'write-protocol-overview
     :expiration `(:interval ,(* 15. 60.))
          :public t
          :language :en
          :keywords `())

;;;*** Function to compute Protocol Overview page ***

(defmethod write-protocol-overview  ((url url:http-search) stream)
   (let ((id-num (car (url:search-keys url))))

   (http:with-conditional-get-response (stream
       :html
       :expires (url:expiration-universal-time url)
       :content-language (url:languages url)
       :additional-headers (ns2.0:client-target-window-http-headers))

   (html:with-html-document (:stream stream)
    (html:with-document-preamble (:stream stream)
     (html:declare-title "Protocol Overview Window" :stream stream)))

   (ns2.0:with-document-body (:background :white :link *link-color*
:visited-link *vlink-color* :stream stream)
      (html:with-paragraph (:stream stream)
         (display-protocol-graphics-section id-num stream))))))
```

Figure 6.2. Sample code for RCT Presenter. These two Lisp functions handle the automatic generation of the protocol overview page for a trial. When the server receives a request for the web page */tb/ProtocolOverview?* with a trial identification number appended, it calls the `write-protocol-overview` function. This function parses the *id-num* of the trial, writes the HTML headers for the page, and fills in the content by calling the `display-protocol-graphics-section` function for the trial whose *id-num* is `2`.

Window for the trial whose *id-num* is *2* (Figure 6.2). This on-the-fly HTML page is sent back to the user's web browser as HTML source, exactly as though that page had simply been cached on the server. The user's browser then loads the pages, and the user sees custom-generated web pages that appear exactly as regular HTML files.

To obtain information from RCT Bank, RCT Presenter communicates with RCT Bank using the Generic Frame Protocol (GFP). Since Ocelot, the knowledge-representation system of RCT Bank, is written in Lisp just as the server functions are, the integration of the RCT Presenter server with RCT Bank was easy.

The RCT Presenter interface is currently at the URL *http://mission.ai.sri.com:8000/tb/trial-bank.html*. Because this URL can be accessed by any operating system using any web browser, RCT Presenter is platform independent and is accessible worldwide. The client web browser must be configured to accept cookies and must be able to support the display of frames and tables.

6.2.1 The RCT Bank Knowledge Base

The class definition for RCT Bank is Ocelot-CCM, an Ocelot-based conceptual model built according to the design specification for the clinical-trials core conceptual model (Chapter 5; see Appendix A). Based on the form and content of Ocelot-CCM, RCT Bank qualifies as a trial bank under the operational definition in Section 3.2.2. The contents of RCT Bank are two complete trials — CHF-STAT (Singh, 1995) and SPINAF (Ezekowitz, 1992) — and fragments of five others. The SPINAF trial information came from the trial's design and execution records; the information on the other trials came from published trial reports. These trials are the same as those used to demonstrate the conceptual coverage of Ocelot-CCM (Section 7.2).

The clinical content of the trial-bank entries include information on the trial's design, subjects and recruitment, treatment assignment, treatment, followup, outcomes definition and measurement, results, administration, and publications.

6.2.2 Sample Session

The user interface for RCT Presenter is based on common web-site–design principles. The RCT Presenter home page consists of a banner across the top, a navigation panel on the left, and a main frame where the information is displayed (Figure 6.3). A sample session

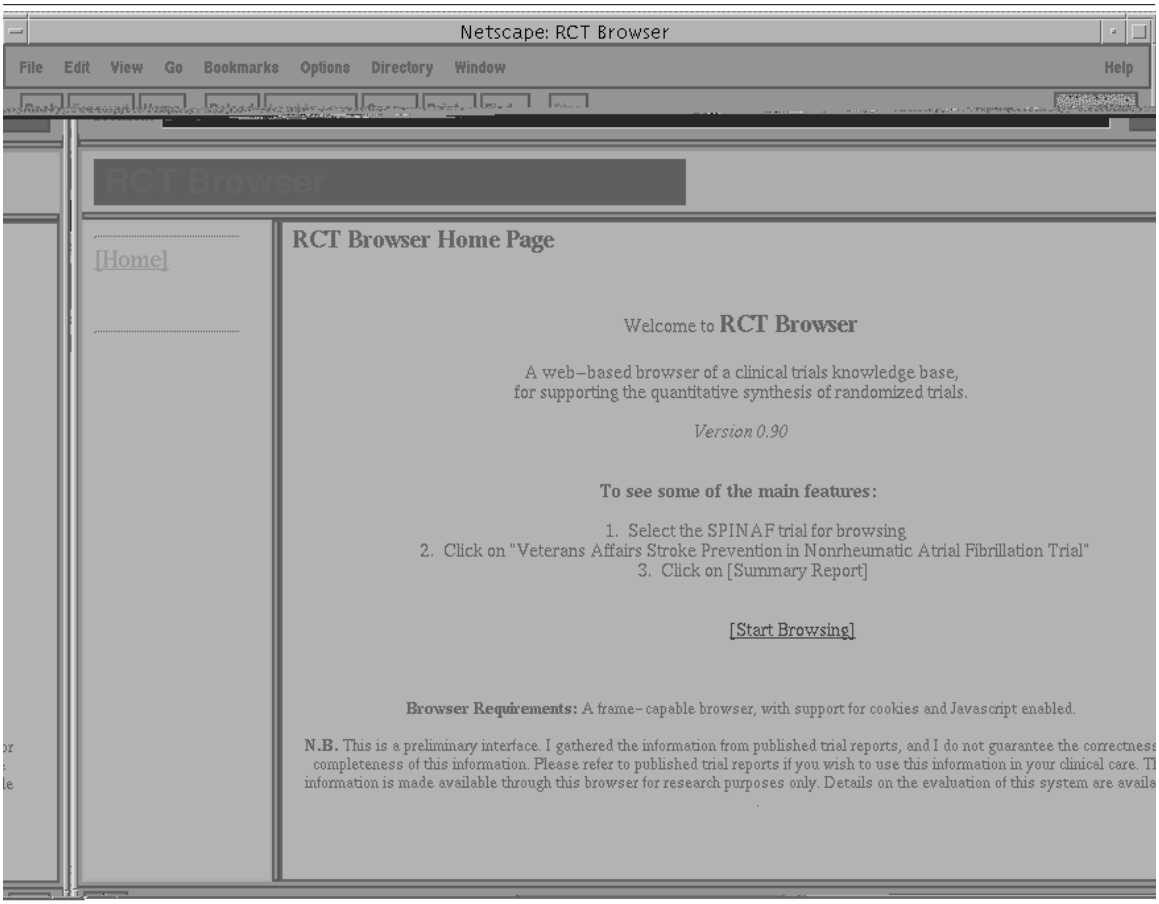

Figure 6.3. Home page of RCT Presenter. The RCT Presenter home page consists of three frames: (1) a title banner, (2) a left-sided panel that displays context-sensitive navigation buttons, and (3) a main frame where trial information is displayed.

of the following four steps will illustrate the user interface: (1) selecting trials for browsing, (2) browsing one trial in detail, (3) viewing information about multiple trials together, and (4) assessing a trial's reporting and methodological quality.

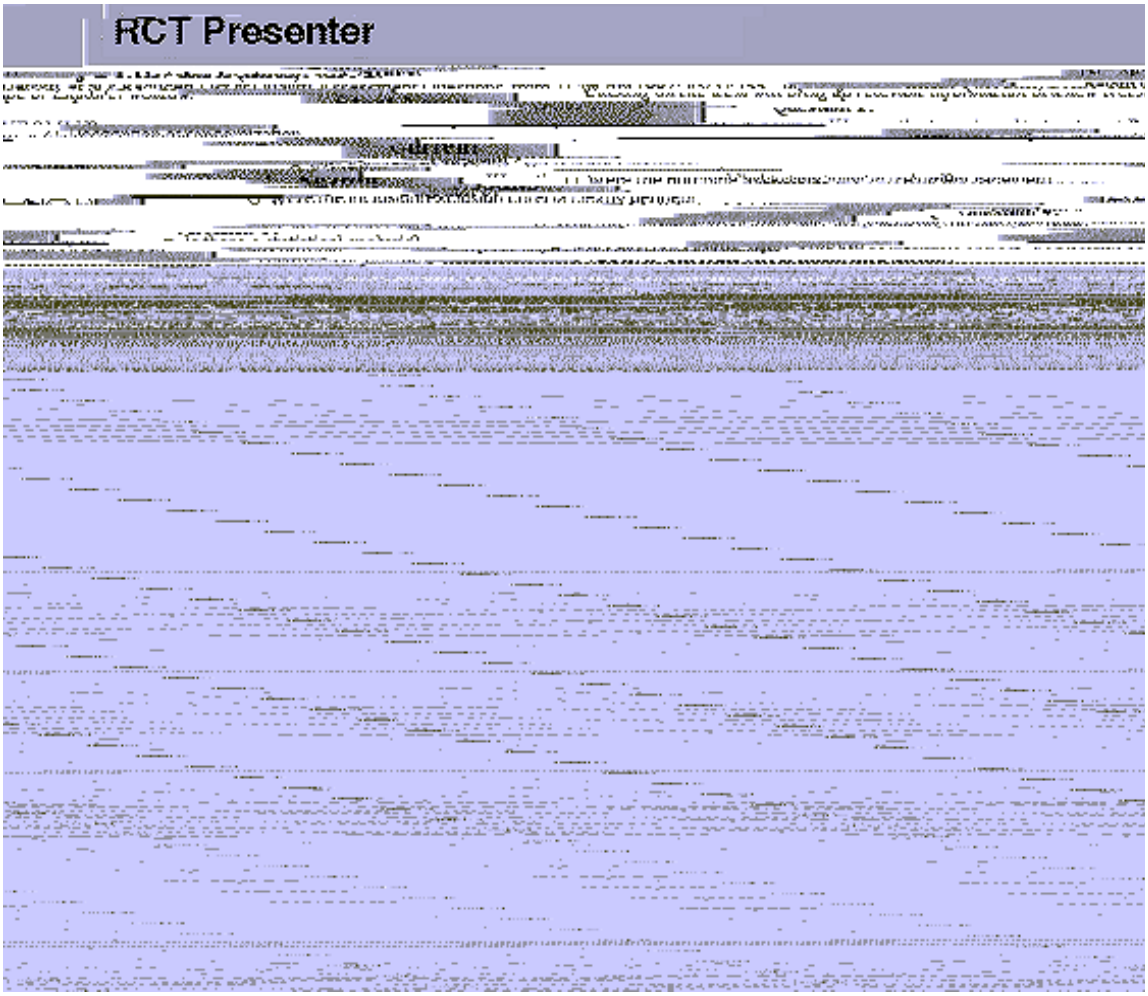

Figure 6.10. Online trial-critiquing tools. The items in this Detsky trial-critiquing questionnaire are hyperlinked automatically to the relevant information for the trial to be critiqued. In this example, clicking on the item "Description of Randomization" has brought up, in another browser window, information on the SPINAF trial's Randomization and Allocation.

hyperlinked to the relevant information for that trial automatically, such that the user needs only to click on an item to have that information brought up in a separate browser window (Figure 6.10). Users of these instruments are thus freed from having to search through a web site — or through pages of printed text in a journal — for the needed information. Such trial-critiquing and reporting-assessment aids can also be of assistance to peer reviewers and editors.

6.3 Potential Extensions

The client–server architecture of RCT Presenter allows for many extensions to the system's functionality. Several types of extensions would be particularly effective for demonstrating the benefits of a fully implemented trial-bank system.

6.3.1 Expert Systems and Tutorials

Expert systems that assist with and explain clinical-trials reasoning and reporting are an obvious extension to RCT Presenter. Added functionality can range from Java applets that perform quantitative meta-analysis, to multimedia tutorials on clinical-trials reasoning, to expert systems that detect fraudulent trial reporting.

6.3.2 Integration with a Vocabulary Server

For full implementation of a trial-bank system to be successful, a standardized medical vocabulary has to be widely adopted by the publishing community (Section 5.2.1, page 117). Because standardized medical vocabularies are large and are expensive to maintain, integration of trial banks with a vocabulary server (Oliver, 1996) would be highly desirable. RCT Presenter could be integrated with a vocabulary server as follows: When a user enters a search term (e.g., "abnormal heart rhythm"), RCT Presenter queries a vocabulary server over the Internet for the controlled term that most closely matches the meaning of "abnormal heart rhythm." The vocabulary server returns the controlled term (e.g., "arrhythmia") as well as all the conceptual descendants of that term (e.g., "atrial fibrillation," "ventricular tachycardia"). Instead of searching for "abnormal heart rhythm," RCT Presenter searches RCT Bank for "arrhythmia" and all of that term's descendants. If the information in RCT Bank is also coded in the same controlled vocabulary, then searching with this refined query will be more accurate than searching with the original "abnormal heart rhythm" keyword. In this way, RCT Presenter can improve its retrieval capabilities by exploiting the conceptual hierarchies of a controlled terminology.

6.3.3 An Open API for Distributed Computing and Database Interoperation

RCT Presenter provides web-based access to only the RCT Bank knowledge base of randomized trials. In a full-fledged trial-bank system, many other client-side applications will want to access RCT Bank. If we are to make RCT Bank accessible to other applications, we would first publish an open applications programming interface (API) for it. An API for RCT Bank will tell developers of client applications how to specify requests for RCT Bank information. This API could be in the form of Java classes, a CORBA or a COM implementation, or a set of GFP calls. If the API were a Java class hierarchy, for example, developers would know which Java classes to invoke to access what information, and they would know what contents and behavior to expect when using these Java classes. There are plans for GFP to become CORBA compliant in the near future, and Java and CORBA are becoming increasingly alike. At present, there is no clear standard technology for encoding open APIs.

6.4 Pilot Evaluation

In May 1997, I pilot tested an early version of RCT Presenter with health-services researchers. I used questionnaires and structured interviews (1) to determine whether the trial-bank description of the randomized trial CHF-STAT contained sufficient information for critiquing that trial, and (2) to evaluate the usability of the RCT Presenter interface.

6.4.1 Study Design and Results

Figure 6.11 presents a structured abstract of the pilot evaluation of RCT Presenter.

Subjects — A convenience sample of 11 health-services research and epidemiology fellows and faculty who were familiar with critical appraisal of randomized trials agreed to participate in this evaluation. None of the subjects had worked on constructing or testing the representation of CHF-STAT in Ocelot-CCM. Three of the subjects had advanced

Objective: To evaluate the information content of RCT Bank for trial critiquing, and to assess the usability of the RCT Presenter web-based interface.

Subjects: Eleven health services research fellows and faculty.

Methods: Subjects completed a 15-item trial-assessment questionnaire (Detsky, 1992), and an end-user computing-satisfaction questionnaire (adapted from Doll, 1988). In addition, subjects were interviewed to solicit open-ended comments.

Outcomes: Percent of questionnaire items completed, accuracy of the answers, satisfaction with RCT Presenter, and responses to the structured interview.

Results: On average, subjects completed the questionnaire in 13.9 minutes, and found the information for 97 percent of the trial-assessment questionnaire items. Correctness of the answers ranged from 55 to 100 percent, with 10 items answered correctly more than 80 percent of the time. On a scale of 1 to 5 where 5 is ideal, the subjects rated the system's ease of use at 4.35, the usefulness of the content at 4.45, and the format of presentation at 4.4. In open-ended interviews, subjects preferred that authors fill in the trial-bank entries themselves, and that all entries be peer reviewed.

Conclusion: The RCT Presenter system was easy to use and provided sufficient information for judging important aspects of trial quality.

Figure 6.11. A structured abstract of the pilot evaluation of RCT Presenter.

training in biostatistics or clinical-trials reasoning (the **advanced subjects**), whereas the other subjects did not (the **basic subjects**).

The sample size was determined by the accessibility of subjects, rather than by any power calculations. Human Subjects Committee approval was not sought, because there was only a minimal risk to the participants: the participants will probably not suffer harm or discomfort more than they ordinarily encounter in daily life, or during the performance of routine physical or psychological examinations or tests.[3]

3. This definition of minimal risk comes from the Administrative Panels Office, Stanford University.

Trial-critiquing and user-satisfaction questions — Subjects were asked to complete the Detsky 15-item trial-critiquing questionnaire (Detsky, 1992; see Appendix C) as best they could, or to write in "Can't find answer" if they were unable to find any information bearing on the trial-critiquing question. The subjects were given a questionnaire that was identical to the published version. I chose to use this questionnaire because it includes many of the most important attributes for judging trial quality, and because it yields results comparable to other, much longer trial-quality instruments (Detsky, 1992).[4]

Subjects were also asked to answer questions adapted from the Computing Satisfaction Questionnaire (see Anderson, 1994, p. 100; and Appendix C). This previously validated questionnaire has a Cronbach's alpha of 0.92.

Target trial for the trial-critiquing questions — Subjects completed the Detsky trial-critiquing questionnaire about the CHF-STAT trial. This two-armed, randomized trial examined the efficacy of amiodarone (an antiarrhythmic drug) for preventing death in patients who had a history of heart disease.

Intervention — On enrollment, all subjects completed three questions about their clinical and demographic background (Appendix C), and were given a quick tour of the RCT Presenter interface. The subjects were then started at the RCT Presenter home page (Figure 6.3, page 131), and were told to take as much time as they needed to answer the trial-critiquing questionnaire. Several subjects made verbal comments as they completed the questionnaire; others were silent. I observed and noted all interface commands, and recorded all comments. I answered only those questions that were about the evaluation study.

After finishing with the trial-critiquing questionnaire, the subjects completed the user-satisfaction questions and were asked to give their reactions to using RCT Presenter. As is

---

4. Trial-quality assessment results are deemed comparable when the trial-critiquing instruments yield the same rank ordering of trial quality. Moher and colleagues estimate that the Detksy questionnaire can be completed in 10 minutes with traditional, paper-based trial reports (Moher, 1995).

common in qualitative research, the interview questions that I asked changed over the course of the evaluation as I identified and pursued respondent themes. The duration of the structured interview was not limited, although I generally kept the total subject participation time to approximately 60 minutes.

Outcomes and analysis — The three outcomes of this descriptive study were (1) the percentage of subjects who answered each question correctly, (2) the user satisfaction with RCT Presenter as scored by the modified Computing Satisfaction Questionnaire, and (3) the responses to the structured interview. I followed the general outlines of the grounded-theory approach (Strauss, 1990) to identify, organize, and analyze the recurrent themes in the interview responses. For the trial-critiquing questionnaire, the unit of analysis was the question.

Results — The trial bank contained sufficient information about CHF-STAT for subjects to complete the entire trial-critiquing questionnaire. On average, the subjects completed the questionnaire in 13.9 minutes, and found the information for 97 percent of the items. Seven of the 11 subjects answered all 15 items; four subjects missed one or two items each. Questions that subjects had most difficulty finding the information for were questions about treatment-assignment bias and outcomes-assessment blinding. For the 12 items that had a definitive answer, the correctness of the subjects' responses ranged from 55 to 100 percent, with 10 items answered correctly more than 80 percent of the time. The three remaining items requested an opinion, and correctness therefore could not be assessed. On a scale of 1 to 5 where 5 is ideal, the subjects rated the system's ease of use at 4.35, the usefulness of the content at 4.45, and the format of presentation at 4.4. On a scale of 1 to 7 where 1 is strong disagreement and 7 is strong agreement, the subjects agreed that it will be worth the time and effort to learn how to use RCT Presenter (6.6 out of 7), that RCT Presenter will make meta-analysis easier (6.7 out of 7), and that RCT Presenter will make meta-analysis better (6.4 out of 7).

Free-form comments from the first subjects revealed several broad themes. Since almost all the comments from the later interviews fell into these same themes, the themes can be

said to have been saturated. This saturation implies that no significantly different themes are likely to appear if more subjects are interviewed. These were the major themes:

- *Linear versus hypertext* — Several subjects were intensely uncomfortable with the hypertext presentation of trial information. One comment was "I want the paper!" and several people yearned for "a beginning, a middle, and an end." Other subjects enjoyed the freedom to wander, and their wandering added several minutes to the time that they took to complete the trial-critiquing questionnaire.
- *Matching of graphics to users' mental models* — The domain graphic (Figure 6.7) was particularly well received by those subjects who had more biostatistical training: "People trained in epidemiology think exactly this way." Subjects with less training in clinical-trials reasoning were occasionally confused by the graphics; several wanted all available trial information listed in one place "so I won't miss anything." The tables and the flowchart of participant followup were well liked by all.
- *Mistrust.of computer-based information* — Many of the subjects wondered about the trustworthiness of the information in the trial bank. Were the entries "in the authors own words?" Were standardized terms used that might have misrepresented the truth? Several subjects said they were far more skeptical of any information "on a computer screen," because on a computer, the absence of traditional cues of quality we expect from a paper-based journal makes both the good and the bad "all look the same." Indeed, one skeptic said, "If I can see the [journal] paper, it'll be okay." Subjects frequently voiced a preference that authors themselves be responsible for their trial-bank entries, and that peer review and editorial oversight be maintained in computer-based reporting.
- *Concern for implications of a trial-bank system* — A few subjects were somewhat concerned that credulous investigators would be enticed by the ease of trial-information retrieval to perform "bad meta-analyses." People might "check their

criticism at the door" and think "since it's here, it's got to be correct." This fear that other people would not be sufficiently skeptical of computer-based reporting contrasts with the skepticism of the subjects themselves.

6.4.2 Discussion of the Pilot Evaluation

This empirical evaluation of RCT Presenter demonstrated that a clinical-trials knowledge base that is built to the design specification for a clinical-trials core conceptual model contained sufficient information for users to complete a trial-critiquing questionnaire from the literature. The subjects were generally pleased with the system's presentation of information, with the usefulness of that information, and with the system's ease of use. Because only one trial-critiquing instrument was used by only 11 subjects on only one randomized trial, the conclusions and the generalizability of this study are limited. Nevertheless, this pilot study complements the evaluation of the abstract properties of my clinical-trials core conceptual model (Chapter 7), and lays the groundwork for future evaluations of trial-bank–browsing systems.

There are several confounding factors to the findings of this study. Chief among these factors were problems with the user interface. In several cases, the poor performance of the subjects was due to poor interface design, rather than to the absence of needed information in RCT Bank. The interface was designed for users who have some advanced biostatistical training, and it was clear that the interface matched the mental model of randomized trials of the three subjects who had advanced biostatistical training, but did not match the mental models of the other eight subjects who had only basic biostatistical training. For example, several of the basic subjects erroneously thought of outcomes-assessment blinding as a trial-level concept. They therefore failed to find this information, even when they were staring right at it in the section on outcomes definition and measurement. They simply did not see what they did not expect to see. In contrast, the advanced subjects quickly and correctly navigated to the section on outcomes definition and measurement, and found the blinding information there. This observation reinforces the need for interface designers to design different interfaces for users with different levels of expertise.

An additional caution to interface designers for trial-bank browsers is that a significant proportion of clinical-trial readers may be discomfited by a purely hypertext presentation of trial information. In response to observing this discomfiture in several of the pilot-evaluation subjects, I added the *Summary Report* feature to RCT Presenter. This feature provides a linear presentation of a trial; it orients users with "a beginning, a middle, and an end" to a trial-bank entry. As people use the World Wide Web and become increasingly familiar with its hypertext format, we will develop new conventions and new design principles for how to present and to read digital publications. Interfaces to the trial-bank system can and should, like interfaces for all web-based systems, cater to and evolve with the needs of its target users.

Another serious confounding factor to the findings in this study was the ambiguity of several of the trial-critiquing questionnaire items (see Appendix C). For example, item 6b asks "Do we know how many patients were excluded from the trial (not enrolled for logistical reasons, refused consent, not eligible)?" If the trial-bank entry tells us how many patients were excluded from the trial, but not the reasons, should the answer be *Yes*, *No* or *Partial*? Item 8c asks "If trial is negative, were confidence intervals or post-hoc power calculations performed?" Does "if trial is negative" refer to trials with a point estimate showing that an experimental treatment was less effective than a control, or to trials in which the confidence interval includes the null hypothesis? The ambiguity of the Detsky questionnaire is average for the 17 other trial-critiquing questionnaires (Section 7.1). In future evaluations of trial critiquing, users of the system must have access to precise definitions for every term.

One clear finding from this pilot evaluation is that extra care must be taken in web-based publishing to assure readers that the information is trustworthy and of high quality. To this end, direct authoring of trials into trial banks is most desirable, and the role of highly respected journals will probably be enhanced — rather than diminished— in the era of trial-bank publishing.

6.5 Summary

RCT Presenter consists of a trial bank called RCT Bank and a web-based browsing interface. The system exploits the structuring of trial information in RCT Bank to generate complex, hyperlinked, web pages automatically and dynamically in response to user queries. In this pilot evaluation, users familiar with clinical-trials reasoning were able to use RCT Presenter to critique a randomized trial, and were generally pleased with the interface.

This evaluation therefore demonstrates empirically the utility of this particular trial bank and this particular interface, but we will not reap from independent systems such as RCT Presenter the myriad benefits described in Chapter 3. Rather, we will reap those benefits from only an interoperating trial-bank system. What assurances do we have that a trial-bank system will be useful? The next chapter addresses this question with results from the evaluation of the core-conceptual-model's design specification, and of Ocelot-CCM.

Chapter 7

Evaluation

Reality is infinite; conceptual models are not. How is one to judge whether a conceptual model captures the appropriate aspects of reality? On what basis can we designate an aspect of reality as appropriate for modeling? Chapter 4 presented the reasons why the modeling and the evaluation of a conceptual model should be task based: whether or not the capture of a certain aspect of reality is consequential depends solely on the intended uses of that model. The intended uses of Ocelot-CCM — my implementation of a clinical-trials core conceptual model — are detailed in its design specification. Therefore, the appropriateness of the what is modeled in Ocelot-CCM depends solely on the design specification. In addition, the task-based view of conceptual modeling suggests that the proper metric for evaluating a conceptual model is not what aspects of reality the model captures, but rather how many of its intended tasks the model supports.

The evaluation of my clinical-trials modeling work therefore had three critical goals: (1) demonstrating the reasonableness of the design specification, (2) defining the tasks that Ocelot-CCM can support, and (3) defining the types of trials for which these tasks can be accomplished with Ocelot-CCM.

7.1 Evaluation of the Design Specification

The design specification for the clinical-trials core conceptual model consists of the competency decompositions for the four core tasks of evidence synthesis (Appendix A). As described in Chapter 4, a competency is a task that a conceptual model is intended, or claims, to support. A competency decomposition decomposes a competency into lower-level subcompetencies and methods, and specifies the trial information required for these competencies to be accomplished. In this section, I compare the design specification to a selection of trial-critiquing instruments in the literature to show that the design specification's competencies and trial-features requirements are reasonable.

7.1.1 Method

Various authors and organizations have devised scales and checklists for critiquing randomized trials. These critiquing instruments are heterogeneous, because there is no consensus on what constitutes a high-quality trial (Section 2.1.2.4). The scales compile a quality score, whereas the checklists serve only to remind readers of putative quality indicators. Nevertheless, the questions posed by these trial-critiquing instruments reflect what the clinical-trials community deems to be reasonable ways to critique a trial.

In 1993, Moher and colleagues systematically searched both published and unpublished sources to obtain a complete list of these trial-critiquing scales and checklists (Moher, 1995). They identified 25 instruments, of which I included 18 in this evaluation. Of the seven excluded instruments, two were in foreign languages, three were unpublished, one was a Public Health Services publication from 1980, and one was more a description of rheumatology trials than a trial-critiquing instrument (Gotzsche, 1989).

I identified all the types of trial information requested by the 18 trial-critiquing instruments. Each **data request**[1] is at the finest granularity of information requested by the instrument. For example, *Study Design* is counted as one request, whereas *Study Design (Type, Model, Blinding)* is counted as three separate requests. Data requests also had to be relevant to the critiquing of randomized trials, and to all clinical domains. For example, I

did not count as a data request the measurement of pain in acupuncture trials (ter Riet, 1990). Figure 7.1 shows an example of identifying data requests.

> For each item either nil (not satisfied) or one point was given. Subsequently, the scores were added to form an eight-point scale of methodological strength. The items were (1) type of publication; (2) inclusion and exclusion criteria clearly described; (3) randomisation method clearly specified; (4) clinical characteristics of the study groups adequately described (ie, at least three of the following characteristics had to be mentioned: age, sex, [relevant clinical characteristics]); (5) description of bleeding complications; (6) accurate diagnosis of DVT; (7) blinded end-point assessment; (8) adequate description of patients not completing the study protocol. A study was considered to have a strong methodology if it satisfied seven or eight of the standards.

Figure 7.1. Data requests in the trial-critiquing instruments. This paragraph is the trial-critiquing instrument by Nurmohamed and colleagues (Nurmohamed, 1992). The instrument was for rating trials comparing low-molecular weight heparin to standard heparin for the prevention of post-surgical deep venous thrombosis (DVT). Each underlined phrase in the paragraph above was counted as a data request.

After identifying all the data requests in the 18 instruments, I tried to match each data request to an equivalent data requirement in the design specification. For example, the data request *accurate diagnosis* in the Nurmohamed instrument was matched to two data requirements in the design specification: *Description of outcome assessment method* (I.E.1.a in Appendix A), and *Validity of outcome assessment method* (I.E.2.a).

I also performed the reverse comparison: For each data requirement in the design specification, I evaluated whether or not any of the 18 trial-critiquing instruments also required

---

1.A *data request* is not the same as an *item* in a checklist. For example, in Brown (Brown, 1991), the item on "Specification of Illness or Condition" includes four data requests: (1) inclusion criteria, (2) exclusion criteria, (3) diagnostic criteria, and (4) description of comorbidities.

Critiquing Instrument	Data Requests	Data Requests Not in the Design Specification
Andrew, 1984	20 / 22	Where and to whom informed consent was given
Goodman, 1994	40 / 42	Adjustment for multiple comparisons; an abstract
Brown, 1991	19 / 22	Race and socioeconomic status of patients; and whether patient knew purpose of outcome assessment
Chalmers, 1990	12 / 12	
Chalmers, 1981	42 / 46	Placebo appearance and taste; power for nonsignificant findings; and adjustment for multiple comparisons
Cho, 1994	24 / 24	
Colditz, 1989	10 / 10	
Detsky, 1992	16 / 16	
Evans, 1985	27 / 29	An abstract, and references
Imperiale, 1990	8 / 8	
Kleijnen, 1991	13 / 13	
Koes, 1991	16 / 16	
Nurmohamed, 1992	12 / 12	
Onghena, 1992	14 / 14	
Reisch, 1989)	46 / 50	Race and socioeconomic status of subjects; procedures for excluding subjects after entry, and for minimizing the loss of subjects
Smith, 1992	12 / 12	
Spitzer, 1990	24 / 26	Power for nonsignificant findings, and adjustment for multiple comparisons
ter Riet, 1990	14 / 14	
TOTAL	**369 / 388 (95%)**	

Table 7.1 Comparison of critiquing-instrument and design-specification requirements. The first number in the "Data Requests" column is the number of overt requests for trial information in the trial-critiquing instrument that were successfully matched in the design specification. The second number is the total number of requests in the trial-critiquing instrument.

that piece of trial information. Data requirements of the design specification that were successfully matched with a data request of a trial-critiquing instrument have a checkmark in the right-most column of the competency-decomposition tables in Appendix A.

7.1.2 Results and Discussion

Table 7.1 shows the number of data requests in each of the 18 trial-critiquing instruments, and the number that were matched successfully to a data requirement in the design specification. Of a total of 388 data requests, 369 (95 percent) were matched in the design specification. Four of the unmatched data requests were duplicates, resulting in only 13 unique unmatched data requests. Conversely, 105 out of 162 (65 percent) of the data requirements in the design specification were matched by one or more data requests of the 18 trial-critiquing instruments.

In summary, almost all the data requests of the trial-critiquing instruments can also be found in the design specification, but the reverse is not true. In other words, a conceptual model that adheres to the data requirements of the design specification will include 95 percent of the trial information that an evidence synthesizer needs to critique a randomized trial using these 18 instruments. The conceptual model will include an additional 57 types of trial information that the biostatistical literature suggests is useful for critiquing randomized trials.

7.1.2.1 Data Requests Not in Design Specification

Table 7.2 (page 152) shows the reasons why the 13 trial-critiquing–instrument data requests are unmatched in the design specification. Two of these six reasons (*unclear use or benefit*, and *trial information not commonly required*) were discussed in Section 5.1.2.6. Rather than requiring fine-grained information on the appearance and taste of a placebo, the design specification requires only more general information on the similarity of the control and experimental treatments. Likewise, the design specification does not require fine-grained information on the race and socioeconomic status of subjects. The requirement for this information is subsumed under the requirement for information on generic characteristics of the subjects. Besides, the race and socioeconomic status of subjects are not clinical-trial concepts; information on these subject characteristics should not be specifically required by the design specification any more than should information on the diabetic status of subjects. The design specification does not require an abstract, because an

Unmatched Data Request	Explanation for Absence in Design Specification
1. Abstract	Not applicable for a trial bank
2. Adjustment for multiple comparisons	Future work
3. References	Future work
4. Appearance of placebo	Too detailed
5. Taste of placebo	Too detailed
6. Race of subjects	Not a clinical-trial concept
7. Socioeconomic status of subjects	Not a clinical-trial concept
8. Where informed consent was given	Too detailed/unclear use or benefit
9. To whom informed consent was given	Too detailed/unclear use or benefit
10. Procedure for excluding subjects after entry	Not commonly required
10. Procedures for minimizing loss of subjects	Not commonly required
12. Whether subjects knew the purpose of outcomes assessment	Not commonly required
13. Post-hoc power for nonsignificant findings	Incorrect conceptualization

Table 7.2 Reasons for unmatched data requests. The reasons that these data requests are not in the design explanation are the same as the reasons for not modeling a trial concept in Ocelot-CCM (Section 5.1.2.6).

abstract is a component not of a trial but rather of a trial report; as such, the concept of an article's abstract is also not a clinical-trial concept.

Post-hoc power calculations for nonsignificant findings are not required by the design specification, because I believe that they reflect an incorrect conceptualization of statistical inference. As described by Goodman and colleagues (Goodman, 1994), it is extremely unclear what to make of a study's power to detect a finding when the study has already observed that very finding in reality. The confidence interval around a negative finding is the proper basis for statistical inferences about the observed effects in a trial.

These modeling choices can be changed if future use of the clinical-trials core conceptual model so suggests.

7.1.2.2 Data Requests in Only the Design Specification

Only 57 (35 percent) of the data items requested by the design specification were not also requested by at least one of the trial-critiquing instruments. One reason for this mismatch is that none of the 18 instruments claimed to be comprehensive. The instruments averaged only 21.5 data requests per instrument (range 8 to 50), for a total of only 117 unique requests after the elimination of duplicates. In contrast, the design specification was designed expressly to be comprehensive so as to support the maximal method for trial critiquing (Section 5.1.1.2). It is therefore not surprising that the design specification had many more data requirements than did the trial-critiquing instruments.

The data-request–mismatch rate was highest for judgments of internal validity and for contextual interpretation. Because no standard definition of trial critiquing exists in the literature, several of the trial-critiquing instruments requested data for the task of contextual interpretation as well (e.g., whether informed consent was obtained). The bulk of the data requests that were in the design specification but that were not in the trial-critiquing instruments were for judgments of internal validity. These unmatched data requests fall into the following three broad types:

Competency Decomposition	In Design Specification	In Instruments	Mismatches	Mismatch Percent
Information retrieval	1	1	0	–
Judgment of internal validity	122	77	45	37
Judgment of generalizability	23	17	6	26
Quantitative computation	0	0	0	–
Contextual interpretation	16	10	6	37
TOTALS	162	105	57	35

Table 7.3 Matching of data requests in the design specification. The design specification requires 162 types of trial information. Of these 162 requests, 105 (65 percent) were also requested by one or more of the 18 trial-critiquing instruments, which together reflect the state of the art in trial critiquing.

1. *Otherwise common requests* — Many of the unmatched data requests are commonly considered important for trial critiquing in the biostatistical literature. An example is information on the reasons that subjects did not complete their assigned treatment. The absence of these common data requests from the 18 trial-critiquing instruments reflects the instruments' lack of comprehensiveness.
2. *Otherwise less common requests*— Several of the unmatched data requests are less commonly discussed in the biostatistical literature, but are nevertheless of sufficient importance to include in a clinical-trials core conceptual model. Example of this class of data request include the details of a trial's interim analysis method, and the stopping rules. Several of these requests are included in highly regarded recommendations for trial reporting (e.g., stopping-rule descriptions in the CONSORT trial-reporting requirements (Begg, 1996)).
3. *Data requests of increasing importance*— Placebo-controlled, randomized trials have long been considered — by regulatory agencies and by the clinical community — to be the gold standard in clinical experiments. Recent controversies have highlighted the possible error of this view. A widely publicized editorial in the New England Journal of Medicine criticized the use of placebo controls in trials on preventing maternal–fetal transmission of HIV in the developing world (Angell, 1997), and was itself then criticized for misunderstanding the science and ethics of randomized trials. Although rarely mentioned until now as an important aspect of trial design, the justification for the type of control used in a trial is important for trial critiquing (Rothman, 1994).

The remaining 54 trial information types that were requested by the design specification, but not by the trial-critiquing instruments, are similar to these examples.

In summary, all the correct, relevant, and reasonable data requests of the 18 trial-critiquing instruments were also data requests of the design specification. Because the design specification is designed to be comprehensive, many of its data requests were not also

requested by the trial-critiquing instruments, but these unmatched data requests are supported by the rest of the biostatistical literature as being important for trial critiquing.

7.2 Demonstration of Conceptual Coverage and Competence

If we can express in Ocelot-CCM every concept there is to express about randomized trials, then we can say that Ocelot-CCM's conceptual coverage is complete. We cannot, however, define what constitutes every concept about randomized trials without reference to what we are trying to do. Therefore, a more meaningful definition of Ocelot-CCM's **conceptual coverage** is the extent to which Ocelot-CCM can capture all the trial features that are stipulated by the design specification (Appendix A). The yardstick by which Ocelot-CCM's evaluation is quantified — the design specification — was shown in Section 7.1 to be a reasonable specification.

7.2.1 Method

The design specification of Ocelot-CCM specifies every trial feature that an evidence synthesizer requires to accomplish each target competency. The design specification is therefore an organizing framework for the trial features that Ocelot-CCM should be able to capture. For example, suppose that the target competency is to determine the validity of a trial's treatment administration (competency I.B in Appendix A, and in Table 7.4). To accomplish this competency, we require information about the trial's experimental intervention (data requirement I.B.1.a). Ocelot-CCM therefore should be able to capture the relevant information about all types of interventions. For example, can Ocelot-CCM capture the description of a medical-device intervention? To determine the answer to this question, I first identified an example of a medical device from the literature: the Medtronic PCD Model 7217B implantable cardioverter-defibrillator device (ICD) allowed in the AVID trial (The AVID Investigators, 1995). This ICD example is a **criterion instance** for testing the coverage of Ocelot-CCM. I then tried to enter into Ocelot-CCM as

Competency Decomposition		Data Requirements	Criterion Instance
Competency	**Subcompetency**		
B. Was the treatment administration valid?	**1. Is the intended treatment clearly described?**	**a.** description of intervention (type, schedule, method, duration, setting)	Loading and maintenance dosages of amiodarone
			Titration of warfarin to prothrombin time
			Implantable Cardioverter-Defibrillator
			Percutaneous Coronary Angioplasty
			Alcohol abstinence counselling

Table 7.4 Organization of criterion instances using the design specification. To accomplish the target task of determining the validity of treatment administration, we require a description of a trial's intervention. Here, five criterion instances reflecting a range of intervention types are associated with this competency. A complete list of criterion instances and their instance-test results is given in Appendix D.

much information as I could about the Medtronic ICD criterion instance. Each attempt to express a criterion instance in Ocelot-CCM is called an **instance test**.

Table 7.5 lists the potential outcomes of an instance test. The instance may be captured by an existing frame (or object, or class) in the conceptual model, or a frame may have to be added or modified. The instance test fails if the model cannot capture a criterion instance without a fundamental change to its structure, defined arbitrarily as changes in the root frame, or in frames in the top two of Ocelot-CCM's five levels (Section 5.2.3). An instance-test outcome is *by argument* if Ocelot-CCM captures this criterion instance exactly as it captures another criterion instance. For example, Ocelot-CCM can capture the mean age of enrolled subjects, and, *by argument*, it can also capture the mean age of subjects lost to followup because, in Ocelot-CCM, mean age can be captured for all subject groups. An instance-test outcome is *cross-reference* if the criterion instance is identical to that of another competency. For example, the instance test of Ocelot-CCM that captures outcome definitions for competency I.D.2.a. is the same as, and is thus cross-referenced to, the instance test that captures outcome definitions for competency I.D.1.a. The difference between the *by argument* and *cross-reference* outcomes is that the former reflects the structure of the conceptual model, whereas the latter reflects the structure of the design specification.

Instance-Test Outcome	Description
OK	Already in the model
Add	Had to add a frame, or to modify an existing one
Failed	Could not capture without fundamentally changing the model
Cross-reference	Instance-test outcome same as that for another competency
By argument	Argued by conceptual similarity to another instance test
Deferred	Modeling planned for future work

Table 7.5 Potential outcomes of an instance test. These are the potential outcomes when I attempted to enter an instance of a concept, called a criterion instance, into Ocelot-CCM.

To enhance their verisimilitude, I gathered criterion instances from published trial reports, and from the original design and execution records of a Veteran's Affairs Cooperative Studies Program trial, the Stroke Prevention in Nonrheumatic Atrial Fibrillation (SPINAF) trial. The SPINAF records are extensive and high quality, and they provided criterion instances that are not commonly found in the published literature. Examples of these uncommon criterion instances include details on the postrandomization exclusion of patients from the trial, details on the use of temporary exclusion rules that allowed initially ineligible patients to became eligible on later rescreening, and data on the subjects' guesses about which intervention — placebo or anticoagulant — they had been taking.

7.2.2 Results and Discussion

Table 7.6 summarizes the results of the 152 instance tests. The complete list of the 152 criterion instances and their instance-test results is tabulated in Appendix D.Overall, Ocelot-CCM captured 142 of 152 (93 percent) of the criterion instances, although, in 33 percent of the instance tests, the model had to be added to or modified. None of the instance tests failed. The 10 instance tests that were classified as *Deferred* are

1. Differences between planned and actual treatment (data requirement I.B.3.).

2–9. The method and efficacy of blinding patients, providers, study nurses, and investigators to interim results of the trial (data requirement I.B.5.).

Instance-Test Outcome	Internal Validity	Generalizability	Contextual Interpretation	Total (%)
Already in model	54	5	4	63 (42)
Had to add or modify	31	9	10	50 (33)
Failed	0	0	0	–
By argument	14	4	0	18 (12)
Cross-reference	5	6	0	11 (7)
Deferred	10	0	0	10 (7)
TOTALS	114	24	14	152
% in final model	91	100	100	93

Table 7.6 Outcomes of instance tests. The bulk of the instance tests were associated with the competency of judging internal validity. Overall, 93 percent of the criterion instances tested were successfully captured by Ocelot-CCM. The other high-level competencies of Ocelot-CCM — information retrieval and quantitative computation — were not tested with criterion instances for reasons discussed in the text.

10. Justification for parameterization or transformation of a result (data requirement I.F.2.c.).

The implications of these deferred instance tests for Ocelot-CCM's competencies are discussed in Section 7.2.2.1.

7.2.2.1 Competency for Core Evidence-Synthesis Tasks

If Ocelot-CCM can capture all the criterion instances associated with the competency decomposition of a task, then Ocelot-CCM is competent for that task. On the basis of the instance-test results by competency (Appendix D and Table 7.6), the competency of Ocelot-CCM for the four core evidence-synthesis tasks is as follows.

Information retrieval — Without a controlled clinical vocabulary, Ocelot-CCM fails the information retrieval data requirement II.A.1.a (Table A.1 on page 182). Thus, Ocelot-CCM cannot, in its current state, support information retrieval using string matching of keywords. Ocelot-CCM is designed to be used in conjunction with a controlled vocabulary, however (Section 5.2.4).

Trial critiquing — Because 10 instance tests associated with this task had a *deferred* outcome, Ocelot-CCM is competent for only 41 of the 44 subcompetencies for the core task of trial critiquing. For the three unsatisfied competencies, however, the criterion instances that could not be captured are not critical in the vast majority of cases.

1. *What treatment was received (I.B.3.)?* — Although Ocelot-CCM should be able to capture any differences between planned and actual treatment, most trials can be critiqued without this information. These differences are often not reported in the current literature.
2. *Were the trial participants blinded to interim trial results (I.B.5.)?* — Details on the method and efficacy of blinding patients, providers, study nurses, and investigators to interim results of the trial accounted for eight of the 10 deferred instance-test results, but this information is rarely requested by trial-critiquing instruments or by the trials-interpretation literature. Therefore, in the majority of cases, Ocelot-CCM is capable of supporting the competency that is one level higher than I.B.5.: *Was the trial administration valid (I.B.)?*.
3. *Were the statistical methods valid (I.F.2.c.)?* — Ocelot-CCM does not yet model the justification for parameterization or transformation of a result. Such justification is useful for assessing the validity of a trial's results and interpretation, but most trials do not include parameterized or transformed results.

Therefore, for a large proportion of trials, and in many approaches to trial critiquing, Ocelot-CCM is competent for all 44 of its trial-critiquing subcompetencies.

Quantitative computation — The only data requirement for the competency of quantitative computation is a complete 2 X 2 contingency table, which Ocelot-CCM captured successfully in instance testing. Therefore, Ocelot-CCM is competent for calculating summary statistics for pairwise comparisons (i.e., odds ratio, relative and absolute relative risk, and the number needed to treat), and is competent for doing quantitative meta-analysis of pairwise dichotomous outcomes using minimal methods.

Contextual interpretation — The competency decomposition for the core task of contextual interpretation is not extensive. Given this decomposition, however, Ocelot-CCM is competent for this task.

7.2.2.2 Conceptual Coverage

In contrast to Table 7.6, which presented instance-test results by their associated competencies, Table 7.8 summarizes the results in terms of the trial features that Ocelot-CCM was able to capture.

How does this conceptual coverage translate into the proportion of published randomized trials that Ocelot-CCM can capture? In a review of 113 trials, Meinert and colleagues found that 62 percent had two treatment arms, and the remaining 38 percent had three or more treatment arms (Meinert, 1984)**.** In this sample of trials, 66 percent involved a crossover design, a design that Ocelot-CCM cannot capture. This proportion of trials with the crossover design is abnormally high. Of the 8836 human-subject trials indexed as a *randomized controlled trial* in the 1996 Medline, only 1160 (13 percent) had the terms *cross, cross-over,* or *crossover* in any part of their Medline records. Therefore, although Ocelot-CCM's inability to capture crossover designs is a drawback, Ocelot-CCM can probably capture the designs of a large majority of trials.

In Meinert's sample of 113 trials, 91 percent had a drug intervention, 5 percent a procedural, 2 percent a behavior-modification, and 1 percent a device intervention. Thus, at most 1 percent of the trials had interventions that could not be captured by Ocelot-CCM. Mortality was an outcome in 12 percent of these 113 trials; most of the other outcomes were laboratory measurements. In a study of sample-size reporting, 52 of 70 trials (74 percent) had dichotomous outcomes. The other 18 trials had continuous outcomes (Moher, 1994).Ocelot-CCM can capture all of these endpoint and result types..

Ocelot-CCM can therefore capture a large proportion of the types of trials in the published literature. One caveat to this analysis is that, since Meinert's study in 1984, more randomized trials include cost, functional-status, and quality-of-life outcomes. Because Ocelot-

Feature Dimension and Range	Comments
Design	
More than two treatment arms	E.g., trials with three interventions
Nested randomization	E.g., subjects randomized to the experimental group are again randomized to Intervention A or B
Run-in/Washout	Delay between enrollment and randomization, or between randomization and intervention
Factorial	Subjects are assigned to one treatment each from a set of randomizations
Prospective Cohort	Subjects with or without an intervention are followed over time
Not Supported	Crossover designs (subjects take one intervention followed by the other)
Subjects	
Patient, MD, etc.	Nature of subjects, and the unity of randomization, is limited only by the controlled vocabulary used
Intervention	
Drugs	Drugs administered either as fixed doses, stepped doses, or by titration to effect
Procedure	E.g., surgical or radiological procedure
Device	E.g., hearing aid
Behavior change	E.g., counseling, or computer-based reminders
Endpoint Type	
Clinical	E.g., lab results, disease state, currently in free text
Death	As either total or cause-specific mortality, or as survival
Not Supported	Costs, functional status, quality of life, genomics
Data-Aggregation Level	
Summary	Only summary descriptors of groups of subjects
Individual	Individual, subject-level data
Result Type	
Dichotomous	Result can only be one of two possible outcomes, e.g., dead or alive

Table 7.7 Clinical summary of conceptual coverage of Ocelot-CCM. This table states the trial features that Ocelot-CCM can and cannot capture. The format is partially adapted from Bailar ((Bailar, 1992), page 48).

Feature Dimension and Range	Comments
Continuous	A real number, e.g., age
Ordinal	E.g., NYHA angina scale
Categorical	E.g., ABO blood types
Proportion	E.g., 62 percent of subjects were over age 65
Parametric summaries	Mean, standard deviation, standard error
Non-parametric summaries	Median, quintiles
Comparative Statistics	Relative risk, odds ratio, absolute risk, number-needed-to-treat
Kaplan-Meier	Survival curves, reported as life tables
Regression	Linear, logistic, Cox
Not Supported	ROC curves (test characteristics of test)
Statistical Method	
Contingency Table and t-tests	Chi-square, Fischers, McNemar's; t-test; non-parametric (Wilcoxon, Mann-Whitney, Sign test)
Not Supported	Multiway tables, Pearson or non-parametric correlation, analysis of variance, sensitivity analysis, transformations

Table 7.7 Clinical summary of conceptual coverage of Ocelot-CCM. This table states the trial features that Ocelot-CCM can and cannot capture. The format is partially adapted from Bailar ((Bailar, 1992), page 48).

CCM cannot yet capture these endpoint types, this analysis slightly overstates the proportion of publishable trials that this model can capture.

7.2.3 Summary of Conceptual Coverage and Competence

Ocelot-CCM implements the design specification with sufficient conceptual coverage to be competent for three of the four core tasks of evidence synthesis: trial critiquing, quantitative computation, and contextual interpretation. To be competent for the core task of information retrieval, Ocelot-CCM must be augmented with a controlled clinical vocabulary. The model's current competencies hold across a broad range of randomized trials, spanning trials with multiple treatment arms, to trials using behavioral interventions, to trials with regression-equation results. Trials with crossover designs, and trials with cost, functional-status, and quality-of-life outcomes are outside of Ocelot-CCM's conceptual

coverage at the moment. Extending Ocelot-CCM to capture crossover trials will be difficult to achieve elegantly, but easy to achieve otherwise (see "Temporal Representation" on page 115); extending Ocelot-CCM to capture cost, functional-status, and quality-of-life outcomes will be easy (see Section 5.2.4).

7.2.3.1 Comparative Competency Analysis

Just as we used the design specification as the yardstick for evaluating Ocelot-CCM's competency, we can use the specification to evaluate the competency of other approaches to supporting trial critiquing. In particular, what competencies are possible if we follow the trial-registry approach, or the CONSORT trial-reporting recommendations?

Trial Registries — Table 7.8 presents the standard trial-registry data-inclusion list (Easterbrook, 1992). A fully completed trial-registry entry would provide information relevant to four of the eight competencies for judging internal validity (competencies I.B., I.D., I.G., I.H.), and to two of the four competencies for judging generalizability (competencies II.A-B.). A trial registry will achieve more competencies if it contains more

Trial-Registry Item	Data-Requirement Match
Protocol title	I.H.4.b
Protocol reference number	None
Name and telephone number of contact person	I.H.2.c
Accrual status (active, closed, or completed)	I.G.1.d
Trial location and number of treatment sites	II.B.1.a
Test and control treatments	I.B.1.a-b and I.B.2.a
Drug information	I.B.1.a-b
Eligibility criteria	II.A.2.a-b
Design (controls, randomization, blinding, placebo)	I.B.2.a-c; I.A.1.a, I.A.2.a; I.B.5.a
Target sample size	I.G.1.b
Principal outcomes or endpoints	I.D.1.a-b
Start and study completion dates	II.B.4.a
Funding source	I.H.1.a

Table 7.8 Competency analysis of standard trial-registry contents. Matched data requirements are from the design specification given in Appendix A.

CONSORT Item	Data-Requirement Match
Title	I.H.4.b
Structured abstract	None
Introduction (e.g., prospectively defined hypothesis, clinical objectives, planned subgroup or covariate analyses)	I.G.1.a; I.D.2.c; I.D.2.b.
Planned study population (inclusion and exclusion criteria)	II.A.2.a-b
Planned interventions and their timing	I.B.1.a; I.B.2.a
Primary and secondary outcome measures	I.D.1.a-b
Rationale and methods for statistical analyses	I.F.4.a-c
Detail main comparative analysis[a]	
Whether analysis was intention-to-treat	I.F.4.a
Prospectively defined stopping rules	I.G.4.a
Unit of randomization	I.A.1.a
Method for generating allocation schedule	I.A.2.a
Method of allocation concealment, timing of assignment	I.A.3.a
Method to separate assignment generator and executor	subsumed by I.A.3.b
Mechanism of masking	I.B.5.a
Similarity of treatment characteristics	I.B.2.c
Allocation schedule control	subsumed by I.A.3.a
Evidence of blinding efficacy (in subjects, investigator, outcomes assessor analyst)	I.B.5.a; I.B.5.d; I.E.3.a
Trial profile	II.A.1.b
Estimate and precision of observed effect on primary and secondary measures	I.F.2.a-c
Present summary data and appropriate descriptive/inferential statistics to permit alternative analyses and replication	I.F.5.; I.F.2.a-b; I.F.3.a-b
Describe prognostic variables by treatment group, and any attempt to adjust for them	II.A.2.d; and subsumed under I.G.1.c; I.F.4.c
Describe protocol deviations and reasons	I.G.1.e-f
Comments on internal validity, generalizability, and interpretation in context of all available evidence	subsumed by I.G.3.a,c,e

Table 7.9 Competency analysis of the CONSORT recommendations. Each of the items that CONSORT recommends to be reported is matched, if possible, to a data requirement in the design specification.

a. The trial information for “Detail main comparative analysis” is not defined, and therefore is not matched to any data requirement in the design specification.

trial information than is specified in this standard list; as it attains increasing competency, it will increasingly meet the content definition of a trial bank (Section 3.2.2.2).

CONSORT — The CONSORT group recommends that trial reports contain all the trial information listed in Table 7.9 on page 164 (Begg, 1996). If a trial report adheres completely to this reporting recommendation, it would provide information relevant to all eight of the competencies for judging internal validity (competencies I.A-H.) and to one of the four competencies for judging generalizability (competency II.A.). However, it guarantees sufficient information for the completion of only four of the 44 trial-critiquing subcompetencies (competencies I and II combined).

If faithfully adhered to, both the trial registry and the CONSORT approaches provide trial information relevant to — but not of sufficient conceptual coverage to complete — most of the competencies for trial critiquing. In contrast, if Ocelot-CCM is completed faithfully for a randomized trial, then sufficient information will be available to complete all but three of the 44 trial-critiquing subcompetencies. Table 7.10 summarizes this comparison of competencies if all the approaches were strictly adhered to.

Approach	**Competency I Judgment of Internal Validity**		**Competency II Judgment of Generalizability**	
	Competency N = 8	Subcompetency N = 32	Competency N = 4	Subcompetency N = 12
Ocelot-CCM	8	32 (29 complete)	4	12 (12 complete)
CONSORT	7	16 (4 complete)	1	2
Trial registry	5	10 (3 complete)	2	3

Table 7.10 Comparative competencies of approaches to trial-information management. This table shows the number of trial-critiquing competencies or subcompetencies for which an approach provides relevant information. In the parentheses are the numbers of subcompetencies for which the approach, if faithfully adhered to, will provide sufficient trial information to complete that subcompetency.

In reality, of course, no approach is always strictly followed. The effective competencies of the trial registry, CONSORT, and Ocelot-CCM trial-bank approaches will be less than those shown in Table 7.10, and will depend on the adherence of trial investigators to the

recommendations of these approaches. Few investigators register their trials with trial registries, because there is no mechanism for doing so efficiently. The CONSORT recommendations are new and untested, and they have not been adopted by all medical journals. Direct entry of trials into trial banks by investigators is completely new, and the acceptability of such direct trial-bank authoring will influence the effective competency of a trial-bank system.

7.3 Trial-Bank Authoring

As described in Chapter 3, academic medical journals will request that prospective trial-report authors submit for editorial review both a prose manuscript and a trial-bank entry describing their trial. For prospective authors to comply, the time and work required to enter a trial directly into a trial bank must be reasonable, and authors must be assured that their trial-bank entries will be fair and accurate. Section 7.3.1 presents preliminary information on how arduous direct trial-bank authoring may be, based on my experience in entering the SPINAF trial directly from the trial's design and execution records. Section 7.3.2 discusses the nature and implementation of quality-control constraints on trial-bank authoring.

7.3.1 Direct Entry of the SPINAF Trial

I used the GKB Editor to enter the SPINAF trial into my Ocelot-based trial bank (RCT Bank, Section 6.2.1). The GKB Editor is a tool for designing and instantiating generic, frame-based knowledge bases. Entering trials into RCT Bank using the GKB Editor requires a familiarity with frame-based knowledge representation, and with the structure of Ocelot-CCM. Using this less-than-ideal direct-authoring interface (Figure 7.2), I took 10 hours to enter the SPINAF trial into RCT Bank. With a more user-friendly direct-authoring interface, trial investigators may spend less time on trial authoring than I did, but they will have to respond to peer-review comments on their trial-bank entries. Thus, I estimate that direct trial-bank authoring will require more than 1 hour, and less than 100

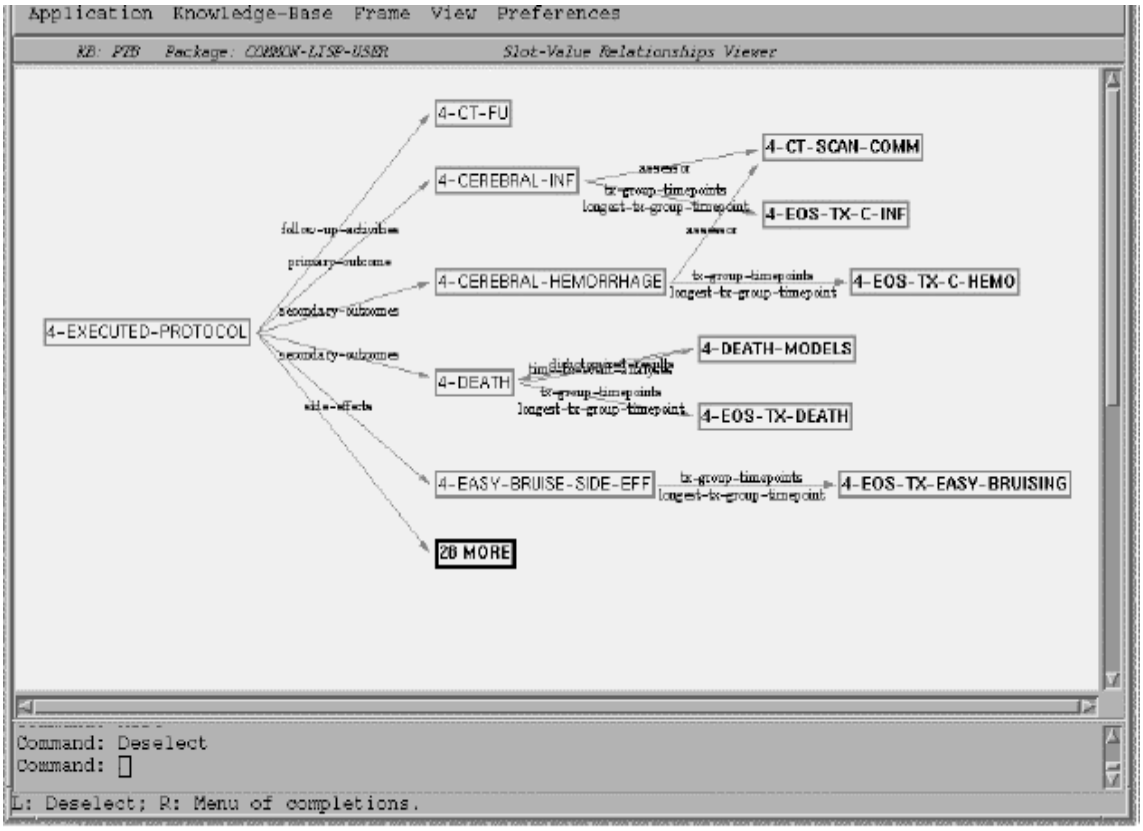

Figure 7.2. Direct trial authoring using GKB Editor. I used this interface to enter trial instances into my RCT Presenter knowledge base. In the view shown here, one can browse and edit the frames that describe the executed protocol of the SPINAF trial. This is not an interface that trial investigators should have to use for trial authoring, because the interface is designed for knowledge engineers rather than for trial investigators.

hours, and probably between 10 and 30 hours. Compared to the time required for preparing a traditional prose manuscript, this is neither a prohibitive nor a trivial amount of time. However, for trial-bank publishing as I envision it, the time required for trial-bank authoring will be in addition to the time required to prepare the traditional manuscript.

In addition to describing their trial in a trial bank, prospective trial authors will have to submit prose manuscripts to the journals. How should trial-report articles best complement trial-bank entries? Initially, readers probably will resist any change to the prose article. With time, however, what trial information should be reported in only the prose article, or in only the trial-bank entry? If articles are freed from having to recite numeric data and

statistics, will they elaborate on why the trial was conducted, what it adds to what we know, and what the implications of the results are? These research questions deserve investigation and experimentation.

7.3.2 Constraints on Trial Authoring

Not all randomized trials can be entered into Ocelot-CCM, even if they fall well within the conceptual coverage of Ocelot-CCM (Section 7.2). Trials with egregiously incorrect designs, or trials without certain critical information, cannot be expressed faithfully by Ocelot-CCM. A trial by Biswas and colleagues (Biswas, 1996) is an example of such a trial.

The Biswas trial was a randomized, placebo-controlled trial to assess the efficacy of low-dose amiodarone (an antiarrhythmic drug) for improving several outcomes in patients who had severe congestive heart failure. Using just the information from the published trial report in the Indian Heart Journal, I could not express this trial in Ocelot-CCM. The trial report gave internally inconsistent data for total survival, and was unclear about the treatment groups from which the withdrawals came. The Biswas trial also violated a critical assumption of Ocelot-CCM: that all trials have a primary hypothesis and a primary outcome.[2] No hypothesis or outcome were designated as primary in the Biswas trial report. Since no power calculations were reported either, the primary hypothesis and outcome could not be inferred. It is unclear whether a primary hypothesis and outcome were simply not reported, or whether no primary designations were ever made in the original trial design. The consequence of the ambiguities, missing information, and the possibly suboptimal design was that the Biswas trial could not be expressed in Ocelot-CCM. Unfortunately, this trial is by no means a rare example of poor design or reporting. Many trials reported to the clinical community in many journals have shortcomings similar to the shortcomings in this case.

---

2. The statistical analyses of the primary hypothesis and outcome are considered confirmatory, rather than exploratory. Other analyses are explorations of multiple comparisons, and must therefore be interpreted in light of with the higher risk of false-positives.

These problems suggest that many of the trials in the current literature cannot be entered retrospectively into a trial-bank system that is based on Ocelot-CCM. These problems also suggest that trial-authoring interfaces must enforce quality constraints on trial-bank authoring — constraints that should be explained by online help and tutorials. Examples of the trial-authoring constraints for Ocelot-CCM include that a primary hypothesis and outcome must be designated, that all results (including denominators for percentage results) must be described clearly and fully, that all statistical tests must be named and *p*-values given, and that participant recruitment and followup must be described fully. Owners of individual trial banks may impose additional constraints on trial authoring — for example, that all funding sources must be revealed.

In summary, all trial-bank entries should be guaranteed to meet a minimum standard of trial design and reporting. These guarantees derive from three sources. First, the design specification for a trial-bank system's core conceptual model specifies the kinds of trial information that the system should guarantee to be sharable among its trial banks. Second, the clinical-trials core conceptual model itself, depending on how it is implemented, may impose constraints on acceptable trial design and authoring for the entire system. Third, trial-bank authoring software may impose additional constraints for individual trial banks. Ideally, any new constraint should be justified with respect to how that constraint will help a user to accomplish a task. The effective minimum-quality standard of trial-bank entries will thus derive from many, interrelated sources, and detailed studies of trial-bank authoring must await a large-scale deployment of the trial-bank system.

7.4 Summary

In this chapter, I presented the evaluation of two major artefacts of my thesis work: the design specification for the clinical-trials core conceptual model, and the implementation of a conceptual model — Ocelot-CCM — according to this design specification. By comparing the design specification to trial-critiquing instruments in the literature, I showed that the design specification is reasonable. By testing criterion instances from published randomized trials, I showed that Ocelot-CCM is able to capture a broad range of trials. By

doing a comparative competency analysis, I showed that a trial-bank system has the potential to support more of the four core tasks of evidence synthesis than do either trial registries or CONSORT trial reports. The actual competency of a trial-bank system will depend on the acceptability of trial-bank authoring; preliminary experience suggests that such authoring will not be prohibitively arduous.

The findings in this chapter complement the empiric findings presented in Chapter 6 — that health-services researchers used RCT Presenter to critique a randomized trial successfully. In total, the evaluations of the design specification, of Ocelot-CCM, and of RCT Presenter show the principled foundations and the utility of my thesis work.

Chapter 8

Summary

8.1 An Informatics Foundation for Evidence-Based Medicine

For all the emphasis we place on the value of randomized-trial evidence for the scientific practice of medicine, we have no coordinated plan for bringing that evidence from the literature to the point of care. We report randomized trials into paper-based journals that are commonly shelved far from the point of care. We rely on individual practitioners to search out new evidence actively, and we expect them to find this evidence using inaccurate search systems. We publish trial reports in a single format, and yet we hope to satisfy the diverse information needs of practitioners, researchers, patients, and methodologists. We provide practitioners scant assistance with that task of synthesizing appropriately the results of multiple trials. Most fundamentally, we act as though randomized-trial evidence by itself is sufficient to change clinical behavior: We provide no systems for placing randomized-trial evidence into the context of local guidelines, resource constraints, and patient records. It is no wonder that randomized-trial evidence languishes on printed pages. It *is* a wonder that we tolerate this state of affairs.

Experience tells us that the several-thousand-word, paper-based, prose article is not the best vehicle for the dissemination of randomized-trial evidence. In the upcoming and inevitable transition to digital publishing, we have a chance to reinvent randomized-trial publishing such that we correct its current shortcomings.

In this dissertation, I propose and present a foundation for a new approach to publishing randomized trials. This new approach, called **trial-bank publishing**, is predicated on academic medical journals requiring prospective randomized-trial authors to submit for editorial review both a prose manuscript, and a description of the trial in a randomized-trial knowledge base — a **trial bank**. Journals will provide authors with web-based trial-bank–authoring tools so that authors can submit their trials accurately, completely, and in a standardized fashion. Manuscripts that are accepted and published will reference their corresponding trial-bank entries; a manuscripts that is published electronically will have a hyperlink that takes the reader directly to that manuscript's trial-bank entry.

This dual publishing of scientific information as both prose and as entries into a structured knowledge base is not new. It has been implemented successfully for the reporting of genomic sequences. The entire *E. coli* genome was recently published by *Science* (Blattner, 1997). The *Science* article described the sequencing work and the implications of that work; the sequence data were published via GenBank, a genomic-sequence database.

How will trial-bank publishing help to get randomized-trial evidence to the point of care? A key postulate of my thesis work is that the evidence from a single randomized trial is not ready to be used until it has been synthesized, or meta-analyzed, with the evidence from all related trials. Thus, evidence synthesis of randomized trials is a necessary and labor-intensive step in transferring randomized-trial evidence to the clinic. Trial-bank publishing is a key to facilitating this critical task. With trial-bank publishing, trials will be acquired directly into trial banks that are designed specifically to support evidence synthesis. These trial banks will guarantee a basic standard of trial reporting and trial quality, and all trial banks worldwide will be integrated, such that users will be able to access all trial banks as though they were a single large trial bank. With an integrated trial-bank system, we can build accurate search engines for randomized-trial reports, we can assist with evidence

synthesis, we can link randomized-trial entries to guidelines and to computer-based patient records, and we can build expert systems that help us to reason about randomized trials stored in trial banks anywhere in the world. In short, we will have an informatics foundation for evidence-based medicine.

My thesis work concerned the design of this trial-bank system. Drawing on database, network, and knowledge engineering, I specified, implemented, and evaluated an abstract model of randomized trials that supports the four core tasks of evidence synthesis: trial retrieval; trial critiquing; quantitative synthesis of trial results; and interpretation of the trial in its scientific, socioeconomic, and ethical context. This abstract model — the clinical-trials **core conceptual model** — is crucial for integrating the trial-bank system. I showed that **Ocelot-CCM**, my implementation of this model, is based on a reasonable design specification, is able to support most of the core tasks of evidence synthesis, and supports these tasks for a broad range of randomized trials.

I also built a trial bank that contains several randomized trials. I entered one of these trials into my trial bank directly from its design and execution records from the Veterans Affairs Cooperative Studies Program. This experience yielded a preliminary estimate that reporting a randomized trial directly into a trial bank will take between 10 and 30 hours.

I used a freeware web-server program (CL-HTTP) from the Massachusetts Institute of Technology to place my trial bank's contents on the web as a web site called RCT Presenter. In a pilot evaluation of this trial-bank–browsing system, health-services researchers were able to use the information in the trial bank to complete a trial-critiquing questionnaire that I culled from the literature. The subjects in this study were generally pleased with the browsing interface. They expressed a strong desire for quality assurance of the trial-bank information, preferably through peer review and through sanctioning by trusted editorial processes.

8.2 Future Work

The work presented in this dissertation is intended to lay a foundation for an extensive informatics infrastructure for evidence-based medicine. There are therefore many areas for future work. As is the case for much work in medical informatics, work on the trial-bank system would be greatly facilitated by the adoption of a standard clinical vocabulary, and by widespread implementation of computer-based patient records.

8.2.1 Trial-Bank–Authoring Software

Trial-bank–authoring software is necessary for any deployment of a trial-bank system. Trial investigators should themselves enter trials into trial banks for several reasons. First, they are the ones who must comply with trial-bank–reporting standards, and it would be inefficient if, for example, abstractors submitted trial-bank entries based on manuscripts that have incomplete information. Second, as shown by my pilot evaluation of RCT Presenter, readers probably will prefer that authors themselves, rather than abstractors, be responsible for the contents of a trial-bank entry. Third, given the numbers of randomized trials conducted worldwide, employing abstractors for trial-bank entry is not a scalable solution. Trial-bank–authoring software should be designed to work in concert with software for other types of electronic publication. A separate line of future research is to characterize the work required for direct trial-bank authoring, and to explore the effect of standardized trial-bank authoring on the design and reporting of future trials.

8.2.2 Incorporation of a Controlled Clinical Vocabulary

As mentioned repeatedly in this dissertation, a trial-bank system must use a controlled clinical vocabulary to describe the clinical content of the trials. Because no global clinical vocabulary standard exists, and because individual trial banks probably will be using different controlled vocabularies (e.g, the Read code in the United Kingdom, and SNOMED terms in the United States), the trial-bank system probably should incorporate a highly cross-referenced system of vocabularies, such as the Unified Medical Language System's Meta-Thesaurus®. This is an extensive area for future research.

8.2.3 Trial-Bank Publishing

Once we have a trial-bank–authoring interface in hand, we can recruit trial investigators to author their trials directly into trial banks, either in conjunction with publishing in a traditional journal, or as part of a program for registering completed trials. Another approach is to work with funding agencies to recruit grant applicants to enter their proposed trial protocols into a restricted-access trial bank, and, should their trial proposals be funded, to complete their trial-bank entries with the executed protocols and the trial results. Yet another approach is to work with regulatory agencies, or with clinical research organizations, to experiment with trial-bank authoring.

Each of these cases of trial-bank authoring will present its own quirks. For example, journal-associated trial-bank publishing may include software to help peer reviewers evaluate trial-bank entries. If so, then peer-review research can be run in parallel with trial-bank–authoring research. A cultural-anthropological study of the adoption of trial-bank publishing would be interesting, as would studies on how trial-bank publishing affects the traditional prose article. The general research question is what effects trial-bank publishing will have on the ease of evidence synthesis, and on the transfer of randomized-trial evidence to the point of care.

8.2.4 Demonstration of Trial-Bank Interoperation

A large area for technical research is to interoperate a networked collection of trial banks using the Ocelot-CCM core conceptual model. Because the knowledge-representation system (Ocelot) that Ocelot-CCM is built in is compliant with the Generic Frame Protocol (GFP), Ocelot-CCM can be mapped easily into Ontolingua, LOOM, and Theo knowledge bases. With GFP as the common syntax, we can demonstrate knowledge sharing among these knowledge bases using Ocelot-CCM as the shared ontology. Alternatively, we can implement the conceptual modeling in Ocelot-CCM as an open applications programming interface (API) in a common object-based language such as C++ or Java, or we can propose it as a sanctioned Object Modeling Group (OMG) health-care standard for storing

randomized-trial information. These latter programs of research are 5- to 10-year plans for demonstrating trial-bank interoperation.

8.2.5 Extensions to the Modeling

Ocelot-CCM can be improved in many ways. The addition of temporal modeling is desirable, but would involve tackling fundamental problems in the knowledge representation of time. If Bayesian design and interpretation of randomized trials becomes more widespread, it will be worthwhile to extend Ocelot-CCM to support this and other probabilistic statistical approaches. As stated in Chapter 5 on page 124, Ocelot-CCM can also be extended to capture instrumental-variables analysis, and to meta-analysis. The addition of procedural knowledge to Ocelot-CCM or to the RCT Presenter system opens up vast possibilities, including semiautomated meta-analysis, data mining, and patient-eligibility determination in conjunction with a computer-based patient record.

8.2.6 Integrated Evidence-Delivery Systems

At present, front-line practitioners must turn to myriad sources for information pertaining to a clinical decision: the medical record, textbooks, the clinical literature, local practice-guideline memoranda, rules on reimbursement, and the World Wide Web. It will soon be common for all these sources to be accessible from a single computer terminal, but the practitioner will not be helped much if all these sources are just in plain text or hypertext. For instance, we do not gain much functionality by storing medical records as electronic text files, rather than as paper-based charts; leading-edge computer-based medical records are structured databases with controlled terms and embedded intelligence. Similarly, we gain comparatively little if we provide practitioners with randomized-trial reports only as electronic text; leading-edge systems for delivering randomized-trial evidence to practitioners will incorporate trial banks, as well as structured databases for systematic reviews, decision and cost-effectiveness analyses, and practice guidelines. These evidence-delivery systems will include embedded intelligence to guide practitioners through the evidence, and to help practitioners make optimal, evidence-based, clinical decisions.

Such an integrated clinical workstation for evidence-based medicine can be built now, but it would require tremendous resources to develop and maintain the randomized-trial and other structured databases. These workstations will be far less expensive to build if we had interoperating, worldwide, self-sustaining networks of these databases, such as the trial-bank system. Developing and evaluating prototypes of these workstations will teach us how best to extend our informatics foundation for evidence-based medicine.

8.3 Contributions

The contributions of this work derive from the application of my expertise in meta-analysis and evidence-based medicine, and of my knowledge of informatics and knowledge engineering, to the problem of transferring scientific evidence from the literature to the clinic.

8.3.1 To Evidence-Based Medicine

My main contribution to the practice and study of evidence-based medicine is a task and information-needs analysis of a critical step in evidence-based medicine: the systematic review of randomized trials. It is widely known in the evidence-based–medicine community that the reporting of randomized trials is suboptimal for the performance of systematic reviews. However, recommendations for improving randomized-trial reporting have not been based on a thorough analysis of the information needs of systematic reviewers. Reporting recommendations are often incomplete, and mix recommendations on the *content* of reporting with those on the *format* of reporting. They also mix requirements of generic, randomized-trial information (e.g., what the sample-size calculation is) with requirements for clinical information (e.g., whether care was delivered in a coronary-care unit). Furthermore, they sometimes do not justify why a trial attribute should be reported.

My design specification for a clinical-trials core conceptual model is a task and information-needs analysis of systematic evidence synthesis. Its format as a competency decomposition states and justifies the tasks, methods, and information needs for evidence

synthesis. The analysis is generic to all clinical domains, and does not confuse what the information-content needs of an evidence synthesizer are with how that synthesizer would like to see that information presented. This comprehensive analysis of the evidence-synthesis task can serve as a starting point for a consensus on randomized-trial–reporting requirements.

8.3.2 To Medical Informatics

It is an important medical-informatics problem that the clinical literature is so inefficient and ineffective at improving clinical care. With a broad perspective drawn from evidence-based medicine, information retrieval, the design and evaluation of informatics systems, and medical journalism, I have concluded that a comprehensive solution to the problem of transferring evidence from the literature to the clinic must start by freeing us from the tyranny of the article: we must stop equating journal articles with evidence. For randomized-trial evidence, the unit of information is the trial, rather than the trial report. Based on this realization, I propose — and provide a principled foundation for — a trial-bank system that can be the seed for an extensive informatics infrastructure for managing the evidence that supports the scientific practice of medicine.

8.3.3 To Knowledge Engineering

The primary contribution of this work to knowledge engineering is in the area of conceptual modeling. Ocelot-CCM is a rich ontology that incorporates significant domain knowledge. It is designed to support the sharing of complex knowledge bases for real-world users who are performing real-world tasks. Ocelot-CCM may provide insight for other researchers on the challenges of modeling domain ontologies. In addition, the trial-bank system will be a potential testbed for knowledge-sharing technologies.

In addition to providing an example of a rich domain ontology, I contribute the competency-decomposition approach to specifying and evaluating ontologies. This approach offers a structured, easily understandable framework for describing the tasks, methods, and domain coverage of an ontology. The approach does not require that ontologies be in

first-order logic; we can thus use this approach to analyze and compare ontologies that claim to have similar competencies, but that are in a variety of knowledge-representation languages.

8.3.4 To Medical Library Science

This work's main contribution to medical library science is to emphasize that effective use of the clinical literature involves far more than just the accurate retrieval of relevant articles. The clinical literature is suboptimal for our clinical decision-making needs; finding suboptimal articles more accurately is only a temporizing solution. Rather, we must switch our research emphasis away from the accurate retrieval of articles to the accurate retrieval of information. We must also pay more attention to what the quality of the information is, and to whether that information is presented to the end user such that optimal decision making is supported.

8.4 Concluding Remarks

This dissertation lays the foundation for a new approach to publishing and disseminating randomized-trial evidence. This new approach is exemplified by trial-bank publishing into an interoperating trial-bank system. It is a comprehensive approach to designing our publication systems such that they support and enable our future clinical decision-support systems. The practice of evidence-based medicine demands no less.

Appendix A

Design Specification

This Appendix contains the competency decomposition of the four core tasks of evidence synthesis: information retrieval (Table A.1), trial critiquing (Table A.2 and Table A.3), quantitative computation (Table A.4), and contextual interpretation (Table A.5). Together, these tables constitute the design specification for a clinical-trials core conceptual model for interoperating trial banks. These tables are discussed in detail in Chapter 5.

For the tasks of information retrieval and quantitative computation, the tables list the target competencies (labeled with Roman numerals, e.g., *I*), the methods by which they are to be accomplished (labeled with capital letters, e.g., *A*), and any method-associated subcompetencies (labeled with Arabic numerals, e.g., *1*). The last two columns show the data and the procedural information that the core conceptual model must include for trial banks to support the accomplishment of each subcompetency, and hence of each competency.

The tables for the tasks of trial critiquing and contextual interpretation do not decompose the competencies into methods (see Section 5.1.1.2). The *Data Requirements* column lists the clinical-trial information needed to accomplish each subcompetency. The last column is checked if the data requirement was also requested by at least one of the 18 trial-critiquing instruments used in the evaluation of the design specification (Section 7.1).

Table A.1 Competency of information retrieval

Competency Decomposition			Requirement of Clinical-Trials Model	
Competency	**Method**	**Method-Associated Subcompetency.**	**Data**	**Procedural**
I. Query capture	**A.** Keyword capture	**1.** Capture Boolean expression	None	None
II. Query matching	**A.** String matching	**1.** Match Boolean combinations	**a.** controlled medical vocabulary for all instance terms **b.** title of trial	**i.** string matching procedure **ii.** logical operators
	B. Matching to numeric values	**1.** Match Boolean combinations		**i.** relational operators (e.g., >, <, =)

Table A.2 Trial-critiquing competency I: Judgment of internal validity

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
A. Was treatment assignment valid?	**1. What was the unit of randomization?**	Definition of unit of randomization necessary to judge appropriateness of statistics, and to identify potential sources of bias	**a.** unit of randomization	✔
	2. What was the randomization method?	Randomization minimizes selection bias by equally distributing unknown confounders between the two treatment groups	**a.** sequence generation method	✔
		Variables that are explicitly controlled for are not randomly distributed in the treatment groups	**b.** stratification variables	✔
		Smaller blocking sizes interfere with randomization	**c.** blocking size	
	3. Was the allocation concealed?	Subjects have to be allocated to a treatment based on some application of the randomization schedule	**a.** method of treatment allocation	✔
		Unconcealed allocation is associated with exaggerated outcomes (Schulz, 1995)	**b.** method of allocation concealment	✔
	4. How effective was the randomization?	If baseline characteristics are equally distributed statistically between the randomized groups, unknown characteristics are also likely to be equally distributed.	**a.** baseline characteristics, as in II.A.2.e, and statistical differences	✔
B. Was the treatment administration valid?	**1. Is the intended treatment described clearly?**	The intended treatment is what the trial was designed to test	**a.** description of intervention (type, schedule, method, duration, setting)	✔
		Intended treatment may include modifications for specific patient circumstances	**b.** patient-specific adjustments allowed	✔
		Intended treatment efficacy can vary depending on skill of execution of treatment	**c.** training and/or skill level of provider of treatment	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	2. Is the control intervention described clearly?	Since the treatment effect is specified as a comparison to the control, we must know what the control treatment was	**a.** description of control (type, schedule, method, duration)	✔
		Rationale for a placebo control should be explicitly discussed	**b.** justification for type of control	
		Explicit description of similarity of interventions yields information on probability of success in masking treatment	**c.** similarity of control and experimental intervention	✔
	3. What treatment was received?	Treatment effect can only be ascertained if it was clear who got what treatment	**a.** which groups and subgroups got which treatment	
		Performance bias if treatment received differed substantially from what was intended	**b.** differences between planned and actual treatment	
		If treatment not given at "start" of trial, outcomes may be falsely attributed to treatment: a performance bias	**c.** time from randomization until treatment	
	4. Did subjects complete their assigned treatment?	Patients who cross-over dilute the treatment effect	**a.** number who crossed over to other intervention	
		Patients who do not complete their assigned intervention dilute the treatment effect	**b.** number who did not complete assigned intervention	✔
		Presence of systematically different reasons for the treatment groups to discontinue assigned treatment indicates a hidden bias	**c.** reasons for not completing assigned treatment	
		Patients who complete their assigned intervention but do so with less than 100% compliance dilute the treatment effect	**e.** compliance in each treatment group and each subgroup	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	5. Was treatment blinded?	Unblinding of patients may lead to performance bias	**a.** method, and efficacy, of blinding of patients to treatment	✔
		Unblinding of care providers may lead to performance bias	**b.** method, and efficacy, of blinding of provider(s) to treatment	✔
		Unblinding of study nurses may lead to performance bias	**c.** method, and efficacy, of blinding of study nurse(s) to treatment	
		Unblinding of investigators may lead to performance bias	**d.** method, and efficacy, of blinding of investigator(s) to treatment	✔
	6. Were trial participants blinded to interim trial results?	Unblinding of patients to results may lead to performance bias	**a.** method, and efficacy, of blinding of patients to results	✔
		Unblinding of care providers to results may lead to performance bias	**b.** method, and efficacy, of blinding of provider(s) to results	
		Unblinding of study nurses to results may lead to performance bias	**c.** method, and efficacy, of blinding of study nurse(s) to results	✔
		Unblinding of investigators to results may lead to performance bias	**d.** method, and efficacy, of blinding of investigator(s) to results	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
C. Were there any confounding cointerventions?	**1. What were the cointerventions?**	Effects that are in fact due to cointerventions may be falsely attributed to the treatment	**a.** description of cointerventions (type, schedule, method, duration)	✔
		If cointerventions were disproportionately taken by one group, then the observed treatment effect cannot so easily be ascribed only to the tested treatment	**b.** proportion of each treatment group taking each cointervention	✔
		Frequent clinic visits during trial followup may lead to improved outcomes that are not generalizable to the non-experimental setting	**c.** frequency and nature of follow-up clinic visits	✔
	2. Was there a washout period?	A prior intervention may still be a confounder if the effect is still present	**a.** wait between enrollment and randomization and/or treatment	
D. Were the outcome definitions valid?	**1. Were the outcome definitions clear?**	Well-defined outcomes (e.g. death) are less subject to error in measurement than poorly defined ones	**a.** outcome definitions (when assessed, by whom, on which subjects)	✔
		Primary outcome is the one used in the a priori power calculation for the trial	**b.** designation of primary and secondary outcomes	
	2. Are the outcomes intermediate or final?	Intermediate outcomes may give only weak support to the study's hypothesis	**a.** outcome definitions (as in I.D.1.a)	
		Need the study hypotheses to determine if the outcomes are intermediate or not	**b.** primary and secondary hypotheses	
		Need the objective of the study to determine if the outcomes are intermediate or not	**c.** objective of the study	✔
	3. Were the side effect definitions clear?	Side effects important for establishing the clinical context of the treatment effect	**a.** definitions of side effects	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	4. Did any outcome definitions change between design and execution?	Trial may not be as valid if trial actually measured something other than originally intended	**a.** any changes in outcome definition	
E. Were outcomes assessed in a valid manner?	**1. Was the assessment method described clearly?**	Full description of assessment method is needed to assess presence or absence of detection bias	**a.** description of assessment method	✔
		Untrained or improperly trained assessors can introduce detection bias	**b.** training of assessor	✔
	2. How accurate was the assessment method?	Unreliable or poorly validated measurement may cause detection bias	**a.** validity and reproducibility of assessment method	✔
	3. Were the outcome assessors blinded?	Lack of assessor blinding can lead to detection bias	**a.** blinding of assessor(s) to treatment received	✔
		Lack of assessor blinding can lead to detection bias	**b.** blinding of assessor(s) to interim and final results	✔
F. Are the outcome results valid?	**1. Were the measurements complete?**	Missing data can lead to exclusion bias	**a.**% of patients yielding usable data at each timepoint, in each treatment group, and in each subgroup	✔
		Exclusion bias can result if certain patients are systematically more likely not to complete assigned treatment.	**b.** characteristics of those who did not complete treatment as assigned and why	✔
		Exclusion bias can result if certain patients are systematically more likely to be lost to followup.	**c.** characteristics of those lost to followup and why	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	2. Were the raw results described clearly?	Raw results must be clear, eg must have a denominator	**a.** raw results of outcomes	✔
		Both the estimate of the effect and its precision (standard deviation or error) are needed	**b.** summary descriptors, with precision	✔
		Parameterized summary descriptors can be misleading if data is not normally distributed	**c.** justification for parameterization, or transformation	
		Total person-years of followup gives best idea of amount of followup for detecting the occurrence of designated outcomes	**d.** total person-years of followup per treatment group	
		Time from randomization till assessment of outcome important	**e.** follow-up time per datapoint	✔
	3. Were the statistical results described clearly?	Need to know which statistical method was used	**a.** name of statistical test	✔
		Actual value of test statistic more useful than a declaration of significance	**b.** actual result of test statistic	✔
	4. Were the statistical methods valid?	Intention-to-treat analysis less biased than efficacy analysis	**a.** intention to treat and/or efficacy analysis?	✔
		Software errors may invalidate results	**b.** name of computer program used	✔
		Inappropriate methods can yield misleading results	**c.** justification for use of statistical methods	
		Inappropriate censoring can lead to misleading results	**d.** handling of losses to followup	✔
	5. Are the results robust to alternative analyses and inferential statistics?		**a.** raw results (as in I.F.2.a-e) and follow-up time and completeness (as in I.F.1.a-c)	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
G. Was the trial design and conduct valid?	**1. Was the design and execution valid?**	Findings for post-hoc hypotheses less persuasive than for *a priori* hypotheses	**a.** primary and secondary hypotheses (as in I.D.2.b), and *a priori* and post-hoc hypotheses	✔
		A negative trial with low power to detect a clinically significant effect lends weaker support to absence of effect	**b.** power calculation (alpha level, tails, power, target effect size), and required sample size	✔
		Findings for post-hoc subgroup analyses less persuasive than for *a priori* ones	**c.** specification of *a priori* and post-hoc subgroup analyses	✔
		Trial critiquing based on stage of trial	**d.** current stage of trial	
		If the protocol changed from design to execution, the trial may no longer be a valid test of the trial hypotheses	**e.** changes between intended and executed protocols	
		Knowing when protocol changed gives idea of how many subjects affected by change	**f.** reasons for protocol changes	
	2. How was any interim analysis conducted?	Knowing the methods of interim analysis, who performed them and how the results affected execution of the trial is helpful for determining presence of any bias	**a.** interim analysis method, schedule, by whom, adjustment for multiple looks, reporting procedure	
	3. Were the trial's conclusions supported by the data?	Need the authors' interpretation of the trial	**a.** authors' conclusion of the trial	✔
		Authors identification and discussion of study limitations helps judging proper strength of conclusion	**b.** authors' statement of study limitations	✔
		Need the authors' recommendation for clinical action, if any	**c.** authors' statement of clinical application	✔
		Actual sample size is needed to judge power of the study for any given effect size of interest	**d.** actual sample size	✔

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
		Conclusions are supported to the extent that the trial is internally valid	**e.** all the other data requirements for I.A-H.	
	4. Why was the trial stopped?	Premature termination of trial may exaggerate finding	**a.** stopping rule	
		Details of interim analysis methods needed to assess whether stopping rule applied with more or less bias	**b.** interim analysis, as in I.G.2.d	
H. Was there an outside source of bias?	**1. Did the funders of the trial influence the results?**	Commercial or other interests may influence a study's outcome	**a.** funding source (who, what type)	✔
		The reporting may be biased if biased sponsors reviewed the manuscript	**b.** funder's right to review or approve the manuscript	
	2. Were the investigators reputable?	Some investigators conduct good trials; some do not	**a.** investigators	
		Area of specialization may bias design and/or results	**b.** area of specialization of each investigator	✔
		Open help and clarification from investigators helps to support faith in results	**c.** name and contact information for contact person	
	3. Was the trial monitoring appropriate?	Information on monitoring committees needed	**a.** name and composition of data monitoring committees	
		A data monitoring committee member who was also an author may not be independent.	**b.** whether any committee members were also authors	
		If no committee members were trained in statistics, they may miss errors.	**c.** background and training of committee members	
		Area of specialization of committee members may bias oversight	**d.** area of specialization of committee members	

Table A.2 Trial-critiquing competency I: Judgment of internal validity (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	4. Was the publication of the study influenced by outside events?	Outside motives for releasing data, e.g. stock price, may introduce bias	**a.** why study published when it was	
		Different publishing sources espouse different peer review standards and may promote particular biases (e.g. commercial or specialty biases)	**b.** details of trial publications (title, journal, volume/no, pages, year, peer-review status)	✔
		Events around time of publication may have influenced reporting	**c.** submission and publication date	✔

Table A.3 Trial-critiquing competency II: Judgment of generalizability

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
A. Were the patients similar to the target population?	**1. How highly selected were the patients?**	How subjects were initially identified can be a main source of selection bias	**a.** method of sampling for potential subjects	✔
		Highly selected patient populations limit generalizability	**b.** recruitment flowchart	✔
	2. What were the patients' clinical characteristics?	Clinical characteristics of enrolled patients should be similar to those of the target population	**a.** inclusion criteria	✔
		Application of exclusion rules yield more highly selected, and less generalizable, trial populations	**b.** exclusion criteria	✔
		In conjunction with Generalizability A.2.c-e., tells one how selected the enrolled population is	**c.** number who got excluded for each exclusion criterion	✔
		Large differences in the clinical characteristics of the included and excluded groups suggests that the included group is less representative of those sampled	**d.** baseline characteristics (including age and sex), of included and excluded groups	✔
		Subjects who complete a run-in/washout period are more highly selected	**e.** run-in or washout period?	
	3. What was the baseline rate?	Effects may not generalize to populations with higher or lower baseline rates	**a.** outcome result in control group	
B. Is the setting comparable?	**1. Where was the trial conducted?**	Site(s) of study may be associated with unobserved variables that affect outcome	**a.** final study sites and enrollment	✔
	2. What was the referral level of the study sites?	Patients from tertiary referral centers are generally sicker than those from primary referral centers	**a.** referral level of each study site	✔

Table A.3 Trial-critiquing competency II: Judgment of generalizability (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	3. What was the health care setting?	Unobserved variables associated with place of treatment may affect outcomes	**a.** inpatient or outpatient treatment	✔
		Unobserved variables associated with payment structure may affect outcomes	**b.** payment method of each site	
	4. When was the study conducted?	Technologies and cointerventions may have changed since the time of the study	**a.** start and end enrollment dates	✔
C. Is the intervention reproducible locally?	**1. What was the objective of the intervention?**	Intention should be equivalent to the target objective, e.g. primary or secondary prevention, or acute or chronic treatment	**a.** treatment objective	
	2. Is the intervention described clearly enough for local duplication?	Customization of itnervention to local constraints may reduce applicability of trial results?	**a.** intervention description, as in I.B.1-2.	✔
	3. How often was treatment taken as assigned?	Overall compliance with assigned treatment should be comparable to local expectations	**a.** completion of assigned treatment, as in I.B.4.	✔
		Degree of compliance may not be achievable in the field if compliance enhancing method of trial very intensive	**b.** method of increasing compliance	✔
		Analysis of the study should appropriately adjust for degree of compliance	**c.** actual compliance, as in I.B.4.e18	✔

Table A.3 Trial-critiquing competency II: Judgment of generalizability (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	4. What were the associated cointerventions?	Treatment effects may be due to cointerventions that may not be generalizable	**a.** cointerventions, as in I.C.1-2.	✔
		Frequent clinic visits during trial follow-up may lead to improved outcomes that are not generalizable to the non-experimental setting	**b.** frequency and nature of follow-up clinic visits, as in I.C.3.	✔
D. Are the study outcomes of local interest?	**1. What was the outcome?**	Measured outcome may or may not be of interest to target population	**a.** outcome definitions, as in I.D.1-3.	

Table A.4 Competency for quantitative synthesis

Competency Decomposition			Requirement of Clinical-Trials Model	
Competency	**Method**	**Method-Associated Subcompetency**	**Data**	**Procedural**
I. Calculate summary statistic, for pairwise comparisons	**A.** Odds Ratio (OR)	**1.** Calculate OR **2.** Calculate 95% confidence interval (ci) for OR	**a.** complete 2 X 2 contingency table	**i.** OR = a*d/b*c **ii.** 95% ci formulas **iii.** deduce 2*2 from necessary, sufficient data
	B. Relative Risk Reduction (RRR)	**1.** Calculate RRR **2.** Calculate 95% confidence interval (ci) for RRR	**a.** same as I.A.1-2.a	**i.** RRR = a/(a+b) c/(c+d) **ii.** 95% ci formula
	C. Absolute Risk Reduction (ARR)	**1.** Calculate ARR 2. Calculate 95% confidence interval (ci) for ARR	**a.** same as I.A.1-2.a	**i.** ARR = a/(a+b) - c/(c+d) **ii.** 95% ci formula
	D. Number Needed to Treat (NNT)	**1.** Calculate NNT	**a.** ARR	**i.** NNT= 1/ARR
II. Quantitative meta-analysis	**A.** Mantel–Haenszel, using odds ratio	**1.** Calculate OR for each trial	**a.** same as I.A.1-2.a	**i.** same as I.A.1-2.a.i
		1. Calculate meta-analytic summary	**a.** ORs for all the trials	**i.** Mantel–Haenszel formula

Table A.5 Competency for contextual interpretation of a trial

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
A. Interpret the trial in its scientific context	**1. What is the biological and clinical background to the trial?**	Pathophysiologic context for interpretation of mechanism and efficacy	**a.** basic science background	
		Epidemiological context	**b.** clinical background	✔
	2. What commentary is there on this trial?	Reflects opinions of leading investigators	**a.** editorials	✔
		Reflects selected opinions of readers	**b.** letters to the editor	✔
		Reflects non-systematic, and therefore possibly biased, commentary on related subjects	**c.** non-systematic review that cite this trial	✔
		Reflects opinions of others in addition to above	**d.** bulletin boards, discussion groups that cite this trial	
	3. What related work is exists?	Other relevant completed studies are part of the scientific context	**a.** other primary research in the literature	✔
		Trials that are ongoing may soon resolve questions	**b.** ongoing related trials	
		Reflects upcoming trends, research questions, etc.	**c.** planned trials	
	4. What work has formally placed this trial's results in the context of others?	Best approach to finding the scientific context of the trial	**a.** systematic reviews that include this trial	✔
		Places a trial into a decision-analytic framework	**b.** decision analyses/cost-effectiveness analyses that include this trial	✔
B. Interpret the trial in its ethical context	**1. If applicable, was human subjects committee clearance obtained?**	Human Subjects Committees (i.e., Institutional Review Boards) are a committee of peers charged with ensuring human rights compliance of study design	**a.** whether Human Subjects Committee approval sought and granted	

Table A.5 Competency for contextual interpretation of a trial (Continued)

Competency Decomposition			Data Requirements of Clinical-Trials Model	✔
Competency	**Subcompetency**	**Justification**		
	2. If applicable, was informed consent obtained?	By Geneva human-rights convention, subjects must be informed of risks of study, and must give informed consent	**a.** whether informed consent obtained from all subjects	✔
		Method may reveal undue pressuring of subjects to give informed consent	**b.** method for obtaining informed consent	✔
C. Interpret the trial in its socioeconomic context		Formal frameworks for incorporating costs, societal tradeoffs, and patient preferences into interpretation of trial results	**a.** decision analyses/cost-effectiveness analyses that include this trial	
		Synthesizes information from trial into action for clinicians	**b.** guidelines that cite this trial	

Appendix B

Ocelot-CCM Class Hierarchy

This Appendix presents selected features of Ocelot-CCM. The Ocelot-CCM class hierarchy comprises 128 class frames with 430 slots. It has a maximum depth of 5. Figure B.1 on page 200 shows the top three levels of the Ocelot-CCM class hierarchy. The second-level classes (e.g., PROTOCOL-CONCEPT) are abstract classes whose purpose is solely organizational.

The class TRIAL-ROOT has three slots: *Definition*, *Documentation*, and *Synonym*. These three documentation slots are inherited by all the other 127 class frames, but all three slots are instantiated for only a few of the classes.

Section E.1 shows how I modeled rules in Ocelot-CCM. This example is intended to illustrate the mechanics of modeling abstract concepts in a frame-based, or object-based, data model.

All class definitions are presented in the following format:

CLASS-NAME
Slot-Name
allowed values: (i.e., the value types that this slot can be instantiated with)

Slots have a single cardinality unless otherwise specified.

RULE
Clauses
allowed values: RULE, a string
cardinality: multiple
Connector
allowed values: AND or OR
Rule-name
allowed values: a string
Temporary
allowed values: YES or No

As an example, the frames that capture the following portion of the inclusion rule for the CHF-STAT trial:

```
(>= 10 PVCs per hour on a 24 hour Holter)
AND (prior history of
   ((rest dypsnea) OR (dypsnea with minimal exertion)
    OR (paroxysmal nocturnal dypsnea))
```

In Ocelot-CCM, this CHF-STAT rule is captured as the following six instance frames:

CHF-STAT-INCLUSION-RULE
Clauses PVCS-RULE, PRIOR-HISTORY-RULE
Connector AND
Rule-name "Partial inclusion rule for the CHF-STAT trial"
Temporary NO

PVCS-RULE
Clauses "`>= 10 PVCs per hour on a 24 hour Holter`"
Connector
Rule-name "PVCs on Holter clause"
Temporary NO

PRIOR-HISTORY-RULE
Clauses REST-DYPSNEA-RULE, DOE-RULE, PND-RULE
Connector OR
Rule-name "Prior history of dypsnea clause"
Temporary NO

1. Subjects excluded by temporary rules can become eligible for the same trial on later rescreening.

REST-DYSPNEA-RULE
Clauses "rest dypsnea"
Connector
Rule-name "Dypsnea at rest clause"
Temporary NO

DOE-RULE
Clauses "dypsnea on minimal exertion"
Connector
Rule-name "Dypsnea on minimal exertion clause"
Temporary NO

PND-RULE
Clauses "paroxysmal nocturnal dypsnea"
Connector
Rule-name "Paroxysmal nocturnal dypsnea clause"
Temporary NO

As discussed in Section 5.2.2.3, the clauses should be instantiated with terms from a controlled clinical vocabulary.

Appendix C

RCT Presenter Questionnaire

This Appendix contains the questionnaire used by the subjects of the RCT Presenter pilot evaluation (Section 6.4). The questionnaire comprises two major parts: (1) a trial-critiquing questionnaire (items 4 to 8) by Detsky and colleagues (Detsky, 1992) that the subjects completed for the CHF-STAT trial; and (2) questions (items 9 to 16) adapted from the End-User Computing Satisfaction questionnaire that was developed and validated by Doll and Torkzadeh ([Doll, 1988).

The results of this questionnaire are presented on page 142.

Appendix D

Instance Tests

This Appendix presents the instance tests that demonstrate the conceptual coverage of Ocelot-CCM. As described in Section 7.2, a criterion instance was selected for each data requirement of the design specification (Appendix A). I then attempted to express the criterion instances in Ocelot-CCM, with the potential outcomes listed in Table D.1. The results are discussed in Section 7.2.2.

Instance Test Outcome	Description
OK	Already in the model
Add	Had to add a frame, or modify an existing one
Failed	Could not capture without fundamentally changing the model
Cross-reference	Instance test outcome same as for another competency
By argument	Argued by conceptual similarity to another instance test
Deferred	Modeling planned for future work

Table D.1 Potential outcomes of an instance test. When attempting to enter an instance of a concept into a conceptual model, these are the potential outcomes.

Table D.2 Instance tests for competency I: Judgment of internal validity

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
A. Was treatment assignment valid?	**1. What was the unit of randomization?**	**a.** unit of randomization	Patient	Ok
			Physician, clinic, hospital	By Arg
	2. What was the randomization method?	**a.** sequence generation method	"Randomization was carried out with a computer allocation schedule"	Add
		b. stratification variables	By hospital and by presence of non-sustained VT on 24 hour Holter	Ok
		c. blocking size	Blocking size of 10	Ok
	3. Was the allocation concealed?	**a.** method of treatment allocation	Description of allocation by central site	Ok
		b. method of allocation concealment	Description of blinding	Ok
			Zelen's method	By Arg
	4. How effective was the randomization?	**a.** baseline characteristics, as in II.A.2.e, and statistical significance	Selected baseline characteristics of the randomized groups	Ok
B. Was the treatment administration valid?	**1. Is the intended treatment described clearly?**	**a.** description of intervention (type, schedule, method, duration, setting)	Loading and maintenance dosages of amiodarone	Mod
			Titration of warfarin to prothrombin time	Add
			Implantable Cardioverter-Defibrillator	Mod
			Percutaneous Coronary Angioplasty (PTCA)	Mod
			Counselling of problem alcohol drinkers by physicians	Mod
		b. patient-specific adjustments allowed	Adjustments for liver and renal dysfunction	Ok
		c. training and/or skill level of provider of treatment	Cardiologist	Add
	2. Is the control intervention described clearly?	**a.** description of control (type, schedule, method, duration)	Placebo	Ok
			Drugs, Surgery, Behavior Modification, or Device	By Arg

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
		b. justification for type of control	To "allow some basis for comparison of the consequences of treatment and non-treatment"	Add
		c. similarity of control and experimental intervention	Placebo of same appearance, made by same manufacturer	Add
	3. What treatment was received?	**a.** which groups and subgroups got which treatment	Ischemic heart disease subgroup and what they got	Ok
		b. differences between planned and actual treatment		Deferred
		c. time from randomization until treatment	19 days for PTCA arm	Add
	4. Did subjects complete their assigned treatment?	**a.** number who crossed over to other intervention	No crossovers	Ok
		b. numbers who did not complete assigned intervention	40.5% in the amiodarone group, 32.5% in the placebo group	Ok
			Numbers discontinuing assigned treatment for each reason	Ok
		c. reasons for not completing assigned treatment	Reasons for non-completion	Add
		d. compliance in each treatment group and each subgroup	Compliance of patients, and target achievement of INR	Add
	5. Was treatment blinded?	**a.**method and efficacy of blinding patients to treatment	method of blinding of patients	Ok
			actual blinding efficacy	Add
		b. method and efficacy of blinding provider(s) to treatment	method of blinding of providers	Ok
			actual blinding efficacy	By Arg
		c. method and efficacy of blinding study nurse(s) to treatment	method of blinding of study nurses	Ok
			actual blinding efficacy	By Arg

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
		d. method and efficacy of blinding investigator(s) to treatment	method of blinding of investigators	Ok
			actual blinding efficacy	By Arg
	6. Were trial participants blinded to interim trial results?	**a.**method and efficacy of blinding patients to result	method of blinding of patients	Deferred
			actual blinding efficacy	Deferred
		b. method and efficacy of blinding provider(s) to result	method of blinding of providers	Deferred
			actual blinding efficacy	Deferred
		c. method and efficacy of blinding study nurse(s) to result	method of blinding of study nurses	Deferred
			actual blinding efficacy	Deferred
		d. method and efficacy of blinding investigator(s) to result	method of blinding of investigators	Deferred
			actual blinding efficacy	Deferred
C. Were there any confounding cointerventions?	**1. What were the cointerventions?**	**a.** description of cointerventions (type, schedule, method, duration)	Allowed: Hydralazine and Isordil, Captopril, Enalapril	Ok
		b. proportion of each treatment group taking each cointervention	Proportion of subjects on digoxin and beta-blocker in each treatment group	Ok
		c. frequency and nature of follow-up clinic visits	Table of follow-up activities	Add
	2. Was there a wash-out period?	**a.** wait between enrollment and randomization and/or treatment	Same as run-in period.	N/A
D. Were the outcome definitions valid?	**1. Were the outcome definitions clear?**	**a.** outcome definitions (when assessed, by whom, on which subjects)	Death due to any cause, on all subjects, method not reported	Ok
			Cerebral infarction on all subjects, by CT scan	Ok
			QOL in those with ICD	Deferred
			Cost	Deferred
			Functional Status	Deferred

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
		b. designation of primary and secondary outcomes	Cerebral infarction primary, death secondary	Ok
	2. Are the outcomes intermediate or final?	**a.** outcome definitions (as in I.D.1.a)	See I.Criterion D.1.a. above	N/A
		b. primary and secondary hypotheses	1 primary and 3 secondary hypotheses	Ok
		c. objective of the study	"To test the hypothesis that amiodarone can prolong survival among patients with CHF and asymptomatic but frequent and complex ventricular arrhythmia."	Ok
	3. Were the side effect definitions clear	**a.** definitions of side effects	Definition of hepatitis	Ok
	4. Did any outcome definitions change between design and execution?	**a.** any changes in outcome definition	Definition of major hemorrhage changed	Ok
E. Were outcomes assessed in a valid manner?	**1. Was the assessment method described clearly?**	**a.** description of assessment method	Definition of what qualifies as an endpoint cerebral infarction	Ok
		b. training of assessor	CT scans read by neuroradiologists	Ok
	2. How accurate is the assessment method?	**a.** validity and reproducibility of assessment method	Hypothetical example	Mod
	3. Were the outcome assessors blinded?	**a.** blinding of assessor(s) to treatment received	Neuroradiologists assessing for cerebral infarction blinded to treatment status	Ok
		b. blinding of assessor(s) to interim and final results	Not clearly stated.	Add
F. Are the outcome results valid?	**1. Were the measurements complete?**	**a.**% of patients yielding usable data at each timepoint, in each treatment group, and in each subgroup	Numbers followed-up and had outcomes assessed, for all groups	Ok

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
		b. characteristics of those who did not complete treatment as assigned and why	Hypothetical example	By Arg
		c. characteristics of those lost to follow-up and why	Hypothetical example	By Arg
	2. Were the raw results described clearly?	**a.** raw results of outcomes	Total mortality at end of study in amiodarone and placebo groups	Ok
			Total mortality at end of study in ischemic and non-ischemic subgroups	Ok
			Side effect rates in amiodarone and placebo groups	Ok
			Kaplan-Meier survival data for total death	Add
			Mean ejection fraction +/- SD in treatment groups	Ok
		b. summary descriptors, with precision	Odds ratio of total mortality, with 95% c.i.	Ok
			Relative risk reduction of cerebral infarction	Ok
			Absolute risk reduction (ARR)	By Arg
			Number needed to treat (NNT)	By Arg
		c. justification for any parameterization, or transformation		Deferred
		d. total person-years of follow-up per treatment group	440 person-years in the placebo and 456 in the amiodarone group	Add
		e. follow-up time per datapoint	Mean follow-up of 1.7 years for placebo group for total mortality	Mod
	3. Were the statistical results described clearly?	**a.** name of statistical test	Kaplan-Meier for total mortality	Ok
			T-test and Cox both done on total death outcome	Add
		b. actual result of test statistic	p=0.6 for total mortality	Ok

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
	4. Were the statistical methods valid?	**a.** intention to treat and/or efficacy analysis?	Intention to treat analysis	Ok
			Efficacy analysis only	By Arg
			Both ITT and efficacy analysis	By Arg
		b. name of computer program used	Hypothetical example	Add
		c. justification for use of statistical methods	T-test for dichotomous, Kaplan Meier for survival, etc.	Ok
		d. handling of losses to follow-up	Description of approach	Add
	5. Are the results robust to alternative analyses and inferential statistics?	**a.** raw results	As in I.Criterion.F.1.a-c and F.2.a-e	N/A
G. Was the trial design and execution valid?	**1. Was the design and execution valid?**	**a.** primary and secondary hypotheses (as in I.D.2.b), and *a priori* and post-hoc hypotheses	All hypotheses	Ok
		b. power calculation and required sample size	Alpha, power, targeted effect size, baseline rate, method used	Ok
		c. specification of *a priori* and post-hoc subgroup analyses	Ischemic subgroups were defined *a priori*	Mod
		d. current stage of trial	Complete, fully reported	Ok
		e. changes between intended and executed protocols	13 protocol changes	Add
			Change from initial to final study sites	Add
		f. reasons for the protocol changes	Hypothetical example	Add
	2. How was any interim analysis conducted?	**a.** interim analysis method, schedule, by whom, adjustment for multiple looks, reporting procedure	Method of Canner used every 6 months by Data Monitoring Committee, reporting to the Executive Committee	Ok

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
	3. Were the trial's conclusions supported by the data?	**a.** authors' conclusion of the trial	"Low-intensity anticoagulation with warfarin prevented cerebral infarction in patients with nonrheumatic atrial fibrillation without producing an excess risk of major hemorrhage."	Ok
		b. authors' statement of study limitations	Hypothetical example	Add
		c. authors' statement of clinical application	That patients with intermittent or chronic AF should be on warfarin; pts. with lone AF or contraindication to warfarin should be on aspirin	Ok
		d. actual sample size	N=525	Ok
		e. all the other data requirements for I.A-H.	Complete trial bank entry	N/A
	4. Why was the trial stopped?	**a.** stopping rule	Stopping rule, method of Canner	Ok
		b. interim analysis, as in I G.2.d	Early termination, following methods described under I.Criterion G.2.a	N/A
H. Was there an outside source of bias?	**1. Did the funders of the trial influence the results?**	**a.** funding source (who, what type)	Funded by both government and private industry	Ok
		b. funder's right to review or approve the manuscript	VA HSR&D reviewed manuscripts	Ok
	2. Were the investigators reputable?	**a.** investigators	Names and affiliations of all investigators	Ok
			CHF-STAT Investigators	Add
		b. area of specialization of each investigator	Specialization of investigators	Add
		c. name and contact information for contact person	Name and address of contact person	Ok
	3. Was the trial monitoring appropriate?	**a.** name and composition of data monitoring committees	Names and membership of all data committees	Ok

Table D.2 Instance tests for competency I: Judgment of internal validity (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
		b. whether any committee members were also authors	Some of the data committee members were authors also	Add
		c. background and training of committee members	Some biostatisticians, some nurses, etc. on SPINAF committees	Add
		d. area of specialization of committee members	Specialization of central committee members	By Arg
	4. Was the publication of the study influenced by outside events?	**a.** why study published when it was	Not stated	Ok
		b. details of trial publications	References to Circulation and NEJM papers, and conference abstracts	Mod
		c. submission and publication date	Dates of main result paper, and publication dates of all other papers	Ok

Table D.3 Instance tests for competency II: Judgment of generalizability

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
A. Were the patients similar to the target population?	**1. How highly selected were the patients?**	**a.** method of sampling for potential subjects	Convenience sampling	Add
			Random, consecutive sampling	By Arg
		b. recruitment flowchart	Recruitment, screening, and eligibility flowchart	Add
	2. What were the patients' clinical characteristics?	**a.** inclusion criteria	Compound rules	Ok
		b. exclusion criteria	Compound rules	Ok
		c. number who got excluded for each exclusion rule	Numbers excluded for each rule	Mod
		d. baseline characteristics, of included and excluded groups	Characteristics of included subjects	By Arg
		e. run-in or washout period?		N/A
	2. What was the baseline rate?	**a.** outcome result in control group	Baseline death rate in placebo group	Ok
B. Is the setting comparable?	**1. Where was the trial conducted?**	**a.** final study sites and enrollment	16 final VA sites	Mod
	2. What was the referral level of the study sites?	**a.** referral level of each study site	Cardiology practice, NOS	Mod
			Community practice, specialty clinic, etc.	By Arg
	3. What was the health care setting?	**a.** inpatient or outpatient treatment	Patients treated as outpatients, MDs were outpatient doctors	Add
		b. payment method of each site	VA payment in all SPINAF sites	Ok
	4. When was the study conducted?	**a.** start and end enrollment dates	Enrollment from 6/1/87 to 5/30/90, with early termination	Ok
C. Is the intervention reproducible locally?	**1. What was the objective of the intervention?**	**a.** treatment objective	Two trials in one, one primary and one secondary prevention	Add
			Acute and chronic therapy, diagnosis	By Arg

Table D.3 Instance tests for competency II: Judgment of generalizability (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
	2. Is the intervention described clearly enough for local duplication?	**a.** intervention description, as in I.B.1-2.	see I.Criterion B.1-2.	N/A
	3. How often was treatment taken as assigned?	**a.** completion of assigned treatment, as in I.B.4.	see I.Criterion B.4.	N/A
		b. method of increasing compliance	Hypothetical example	Add
		c. actual compliance as in I.B.4.e	as in I.Criterion G.2.a.	N/A
	4. What were the associated cointerventions?	**a.** cointerventions, as in I.C.1-2.	as in I.Criterion C.1-2.	N/A
		b. frequency and nature of follow-up clinic visits	Table of follow-up activities	Add
D. Are the study outomes of local interest?	**1. What was the outcome?**	**a.** outcome definitions, as in I.D.1-3.	see I.Criterion D.1-3.	N/A

Table D.4 Instance tests for contextual interpretation of a trial

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
A. Interpret the trial in its scientific context	**1. What is the biological and clinical background to the trial?**	**a.** basic science background	Hypothetical example	Add
		b. clinical background	Background for SPINAF study, previous trials discussed	Add
	2. What commentary is there on this trial?	**a.** editorials	Singer on SPINAF	Ok
		b. letters to the editor		Ok
		c. nonsystematic reviews that cite this trial		Deferred
		d. bulletin boards, discussion groups that cite this trial		Deferred
	3. What related work exists?	**a.** other primary research in the literature		Deferred
		b. ongoing related trials		Deferred
		c. planned trials		Deferred
	4. What work has formally placed this trial's results in the context of others?	**a.** systematic reviews that include this trial		Deferred
		b. decision/cost-effectiveness analyses that include this trial		Deferred
B. Interpret the trial in its ethical context	**1. If applicable, was human subjects committee clearnace obtained?**	**a.** whether Human Subjects Committee approval sought and granted	SPINAF approval sought and obtained	Ok
	2. If applicable, was informed consent obtained?	**a.** whether informed consent obtained from all subjects	Consent obtained from all CHF-STAT subjects	Ok
		b. method for obtaining informed consent	Hypothetical example	Add

Table D.4 Instance tests for contextual interpretation of a trial (Continued)

Competency Decomposition		Data Requirements	Instance Testing	
Competency	**Subcompetency**		**Criterion**	**Result**
C. Interpret the trial in its socioeconomic context		**a.** decision/cost-effectiveness analyses that include this trial		
		b. guidelines that cite this trial		

Appendix E

Glossary

This appendix provides a glossary of terms used in this dissertation. Several of these terms have different meanings within specialized communities; for the sake of simplicity, I have omitted some of these distinctions.

Class A generic description of a thing or concept in the object or frame **data-model**. A class can be contrasted with an **instance**, which is a particular example of a thing or concept. For example, a DRUG object that generically describes drugs as having a brand name is a class object; the AMIODARONE object whose brand-name attribute is *Cordarone* is an instance object.

Competency Decomposition An approach for specifying and evaluating conceptual models, based on decomposing a target task, or a competency, into its subtasks and methods, and specifying the domain concepts needed to accomplish those target tasks.

Conceptual Coverage The extent to which a conceptual model includes all the domain concepts needed for fulfilling its competencies as specified in its **competency decomposition.**

Conceptual Model A description of a part of the world: the concepts about that part of the world (e.g., a drug), and the relationships among those concepts (e.g., patient takes a drug). Conceptual models are expressed using a knowledge-representation language. In this dissertation, this term is synonymous with **data schema**.

Core Conceptual Model A **conceptual model** that includes the core concepts in a domain, with respect to a defined set of tasks and methods, to be accomplished by a defined user for a defined purpose.

Database An electronic collection of information that emphasizes the storing of many instances of simplified information (e.g., the prices of many cars). Databases are on a continuum with **knowledge bases**, which tend to store fewer instances of more complicated information (e.g., metabolic pathways of the *E. coli* bacterium). A database comprises a **database schema** and its **instances**.

Data-Definition Language (DDL) A computer-based language for representing knowledge in a particular **data model** format. For example, the object data-definition languages are used to define object-based data structures that represent some knowledge of the world.

Data Model The structure of the organization of data in a database. Examples of data models include the relational model and the object-based (or object-oriented) model.

Database Schema A specification of how data is organized in a database. It typically follows one of several data models — for example, the relational or the object data model — and is specified using a data-definition language. Also called a **conceptual model** or an **ontology** in this dissertation.

Direct Trial-Bank Authoring The process by which trial investigators themselves describe their trials directly into **trial banks**, using specially designed trial-bank authoring software.

Evidence-Based Medicine The use of conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients or for making population-level health-care policy.

Frame A data structure for representing declarative knowledge about the world. The data structures follow the frame data-model, which defines concepts or things as *frames* and their properties as *slots*. Slot properties are described in *facets*. In this dissertation, **Frame** is synonymous with **Object**, and *slots* with *attributes*.

Generic Frame Protocol (GFP) An emerging standard syntax for specifying and sharing frame-based ontologies.

Instance A particular example of a generic thing or concept. For example, the instance object (or instance frame) AMIODARONE is an instance of the class object (or class frame) DRUG.

Instantiate The act of replacing a variable with a constant. For example, in the SPINAF trial, the experimental intervention is instantiated by the drug amiodarone.

Interoperation The integration of a networked system of databases such that input and output into the system is uniform, and the constituent databases appear as one to the user.

Knowledge Base An electronic collection of information that emphasizes the storing of fewer instances of less simplified information, in contrast to **databases**. A knowledge base comprises an **ontology** and its **instances**.

Knowledge Representation The building of computable models of some domain for some purpose. These models are encoded in **knowledge-representation languages**.

Knowledge-Representation Language A natural or artificial language for describing conceptual models, or ontologies, and for computing with those conceptual models.

Meta-Analysis (1) A review article in which studies have been systematically identified, retrieved, and evaluated, and their quantitative results combined using meta-analytic methods. Meta-analyses are subset of **systematic reviews**. (2) A statistical method for combining the quantitative results of multiple studies.

Object A data structure for representing declarative knowledge about the world. The data structures follow the object data-model, which defines concepts or things as *objects* and their properties as *attributes*. In this dissertation, **Object** is synonymous with **Frame**.

Ocelot-CCM My implementation of a conceptual model in the Ocelot knowledge-representation system. Ocelot-CCM is also the ontology for RCT Bank.

Ontology Another term for **conceptual model** and **data schema**. An ontology can be used to define the contents of a knowledge base

Precision The fraction of retrieved documents that are relevant to the search query.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

Randomized Trial An experimental design in which subjects are randomly assigned to a treatment. The benefit of this design is that unknown confounders are randomly distributed among the treatment groups, and one can therefore make a stronger inference that

any differences in observed effects among the treatment groups are due solely to the treatment assigned.

Recall The fraction of all relevant documents in a document collection that are identified by the search query.

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents in the document collection}}$$

RCT Bank My implementation of a **trial bank**, based on the **Ocelot-CCM** conceptual model. It is the knowledge base of the **RCT Presenter** system.

RCT Presenter A system comprising a **RCT Bank** and a web server that allows a user to browse the contents of RCT Bank over the web.

Systematic Review A summary of the evidence in the literature pertaining to a particular question. A systematic review includes studies have been identified, retrieved, and evaluated using prespecified and uniform procedures. When appropriate, statistical methods such as **meta-analysis** methods are used to combine the quantitative results of the studies.

Trial Bank A structured, electronic database that includes information about a trial's design, execution, and results. It must have an explicit **data schema** that satisfies the design specification in Appendix A.

Trial-Bank Publishing The publishing of randomized-trial results concurrently as both a text article in a medical journal and as an entry in a **trial bank**.

Trial-Bank System A network of **trial banks** that **interoperate**.

Bibliography

1995 Science Citation Index Journal Citation Reports. Philadelphia: Institute for Scientific Information.

ACTIS (1996). "AIDS Clinical Trial Information Service." [Web document] 10/1/96. Available: http://www.actis.org/

Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987). "A proposal for more informative abstracts of clinical articles." *Ann Intern Med* **106**(4): 598-604.

Altman DG (1994). "The scandal of poor medical research." *BMJ* **308**(6924): 283-4.

Anderson J, Aydin C, Jay S, Eds. (1994). *Evaluating Health Care Information Systems.* Thousand Oaks, CA: Sage Publications.

Andrew E (1984). "Method for assessment of the reporting standard of clinical trials with roentgen contrast media." *Acta Radiol [Diagn]* **25**(1): 55-8.

Angell M (1997). "The ethics of clinical research in the Third World [editorial]." *N Engl J Med* **337**(12): 847-9.

Antman E (1992). "A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction." *JAMA* **268**(2): 240-8.

Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature (1996). “Checklist of information for inclusion in reports of clinical trials.” *Ann Intern Med* **124**(8): 741-3.

The AVID Investigators (1995). “Antiarrhythmics Versus Implantable Defibrillators (AVID)--rationale, design, and methods.” *Am J Cardiol* **75**(7): 470-5.

Bailar JC, Mosteller F (1988). “Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations.” *Ann Intern Med* **108**(2): 266-73.

Bailar JC, Mosteller F (1992). *Medical Uses of Statistics.* 2nd edition. Boston: NEJM Books.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. (1996). “Improving the quality of reporting of randomized controlled trials. The CONSORT statement.” *JAMA* **276**(8): 637-9.

Bennett KJ (1987). “A controlled trial of teaching critical appraisal of the clinical literature to medical students.” *JAMA* **257**(18): 2451-4.

Biswas A (1996). “Low-dose amiodarone in severe chronic heart failure.” *Indian Heart J* **48**(4): 361-4.

Black N (1996). “Why we need observational studies to evaluate the effectiveness of health care.” *BMJ* **312**:1215-8.

Blattner FR (1997). “The complete genome sequence of Escherichia coli K-12.” *Science* **277**(5331): 1453-74.

Boguski MS (1995). “Hunting for genes in computer data bases.” *N Engl J Med* **333**(10): 645-7.

Brown SA (1991). “Measurement of quality of primary studies for meta-analysis.” *Nurs Res* **40**(6): 352-5.

Burks C (1992). “GenBank.” *Nucleic Acids Res* **11**(20): 2065-9.

Burks C, Tomlinson LJ (1989). “Submission of data to GenBank.” *Proc Natl Acad Sci U S A* **86**(2): 408.

Burnham JC (1990). “The evolution of editorial peer review.” *JAMA* **263**(10): 1323-9.

Cochrane Collaboration (1997). “Baltimore Cochrane Center -- Medline Retagging.” [Web document] October 3, 1997. Available: http://www.cochrane.org/retag.html

Chalmers I (1990). “Underreporting research is scientific misconduct.” *JAMA* **263**(10): 1405-8.

Chalmers I (1990). “A cohort study of summary reports of controlled trials.” *JAMA* **263**(10): 1401-5.

Chalmers I, Altman D, Ed. (1995). *Systematic Reviews.* BMJ Publishing Group.

Chalmers I, Haynes B (1994). “Reporting, updating, and correcting systematic reviews of the effects of health care.” *BMJ* **309**(6958): 862-5.

Chalmers TC, Lau J (1993). “Meta-analytic stimulus for changes in clinical trials.” *Stat Methods Med Res* **2**(2): 161-72.

Chalmers TC (1981). “A method for assessing the quality of a randomized control trial.” *Controlled Clin Trials* **2**(1): 31-49.

Chandrasekaran B, Johnson T (1993). *Generic tasks and task structures: History, critique, and new directions.* Second-Generation Expert Systems. Springer-Verlag. 232-272.

Cho MK, Bero LA (1994). “Instruments for assessing the quality of drug studies published in the medical literature.” *JAMA* **272**(2): 101-4.

Cimino JJ (1996). “Review paper: coding systems in health care.” *Methods Inf Med* **35**(4-5): 273-84.

Cinkosky MJ, Fickett JW, Gilna P, Burks C (1991). “Electronic data publishing and GenBank.” *Science* **252**(5010): 1273-7.

Cohen A (1997). “Peer Review in Theoretical Physics: An Electronic Study.” In D. Rennie (Ed.), *3rd International Congress on Biomedical Peer Review and Global Communications*, page 11. Prague, Czech Republic: JAMA, The BMJ, and Project Hope.

Colditz GA, Miller JN, Mosteller F (1989). “How study design affects outcomes in comparisons of therapy. I: Medical.” *Stat Med* **8**(4): 441-54.

Cole WG (1995). “Semantic visualization of oncology knowledge sources.” *Proc Annu Symp Comput Appl Med Care* 67-71.

Connect (1996). “Cyberspace Law for Non-Lawyers.” [Web document] 10/1/96. Available: http://www.counsel.com/cyberspace/

Cook DJ, Mulrow CD, Haynes RB (1997). “Systematic reviews: synthesis of best evidence for clinical decisions.” *Ann Intern Med* **126**(5): 376-80.

Cook DJ, Sackett DL, Spitzer WO (1995). “Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis.” *J Clin Epidemiol* **48**(1): 167-71.

Covell DG, Uman GC, Manning PR (1985). “Information needs in office practice: are they being met?” *Ann Intern Med* **103**(4): 596-9.

Delamorthe T (1996). “Posting Pre-prints on the Internet.” [Web document] 9/30/96. Available: http://www.tecc.co.uk/bmj/preprint.htm#1

DerSimonian R, Charette LJ, McPeek B, Mosteller F (1982). “Reporting on methods in clinical trials.” *N Engl J Med* **306**(22): 1332-7.

Detsky AS (1992). “Incorporating variations in the quality of individual randomized trials into meta-analysis.” *J Clin Epidemiol* **45**(3): 255-65.

Dickersin K (1990). “The existence of publication bias and risk factors for its occurrence.” *JAMA* **263**(10): 1385-9.

Dickersin K (1992). “Why register clinical trials?--Revisited.” *Control Clin Trials* **13**(2): 170-7.

Dickersin K, Min YI, Meinert CL (1992). “Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards.” *JAMA* **267**(3): 374-8.

Doll WJ, Torkzadeh G (1988). “The measurement of end-user computing satisfaction.” *MIS Quarterly* **12**:259-274.

DuMouchel W, Harris J (1983). “Bayes method for combining the results of cancer studies in humans and other species.” *J Am Stat Assoc* **78**:293-308.

Easterbrook PJ (1992). “Directory of registries of clinical trials.” *Stat Med* **11**(3): 345-59.

EB Medicine Working Group (1992). “Evidence-based medicine. A new approach to teaching the practice of medicine. Evidence-Based Medicine Working Group.” *JAMA* **268**(17): 2420-5.

Eddy DM (1990). “The challenge.” *JAMA* **263**(2): 287-90.

Eddy DM, Hasselbald V, Schachter R (1992). *Meta-Analysis by the Confidence Profile Method: The statistical synthesis of evidence.* San Diego, CA: Harcourt Brace Jovanovich.

Emerson JD (1990). “An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials.” *Control Clin Trials* **11**(5): 339-52.

Evans M, Pollock AV (1985). “A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection.” *Br J Surg* **72**(4): 256-60.

Ezekowitz MD (1992). “Warfarin in the prevention of stroke associated with nonrheumatic atrial fibrillation. Veterans Affairs Stroke Prevention in Nonrheumatic Atrial Fibrillation Investigators.” *N Engl J Med* **327**(20): 1406-12.

Felciano R, Altman R (1996). “RNA Secondary Structure as a Reusable Interface to Biological Information Resources” Section on Medical Informatics, Stanford University, SMI-96-0641.

Field MJ, Lohr KN, Ed. (1992). *Guidelines for Clinical Practice: From Development to Use*. Washington, DC: Institute of Medicine.

Fikes R, Cutkosky M, Gruber T, van Baalen J (1991). “Knowledge sharing technology project overview.” Knowledge Systems Lab, Stanford University, KSL-91-71.

Fletcher RH, Fletcher SW (1979). “Clinical research in general medical journals: a 30-year perspective.” *N Engl J Med* **301**(4): 180-3.

Friedman L, Furberg C, DeMets D (1985). *Fundamentals of clinical trials.* Littleton, MA: PSG Publishing Company.

Friedman PJ (1990). “Correcting the literature following fraudulent publication.” *JAMA* **263**(10): 1416-9.

Froom P, Froom J (1993). “Deficiencies in structured medical abstracts.” *J Clin Epidemiol* **46**(7): 591-4.

Funk ME, Reid CA, McGoogan LS (1983). “Indexing consistency in MEDLINE.” *Bull Med Libr Assoc* **71**(2): 176-83.

Genesereth M (1990). “Knowledge Interface Format.” Department of Computer Science, Stanford University, 1990. Logic-90-04.

Gennari J (1993). “A brief guide to MAITRE and MODEL: An ontology editor and a frame-based knowledge representation language.” Knowledge Systems Lab, June 1993. KSL-93-45.

Glantz SA (1980). “Biostatistics: how to detect, correct and prevent errors in the medical literature.” *Circulation* **61**(1): 1-7.

Gomez-Perez A, Juristo N, Pazos J (1995). *Evaluation and Assessment of the Knowledge Sharing Technology.* Towards Very Large Knowledge Bases. IOS Press. 289-296.

Goodman SN, Berlin J, Fletcher SW, Fletcher RH (1994). “Manuscript quality before and after peer review and editing at Annals of Internal Medicine.” *Ann Intern Med* **121**(1): 11-21.

Goodman SN, Berlin JA (1994). “The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results [published erratum appears in Ann Intern Med 1995 Mar 15;122(6):478].” *Ann Intern Med* **121**(3): 200-6.

Gore SM, Jones G, Thompson SG (1992). “The Lancet's statistical review process: areas for improvement by authors.” *Lancet* **340**(8811): 100-2.

Gotzsche PC (1989). “Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis.” *Controlled Clin Trials* **10**(1): 31-56.

Greene HL (1994). “Antiarrhythmic drugs versus implantable defibrillators: the need for a randomized controlled study.” *Am Heart J* **127**(4 Pt 2): 1171-8.

Greenland S (1994). “Invited commentary: a critical look at some popular meta-analytic methods.” *Am J Epidemiol* **140**(3): 290-6.

Gregoire G, Derderian F, Le Lorier J (1995). “Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias?” *J Clin Epidemiol* **48**(1): 159-63.

Gruber T (1993). “Toward principles for the design of ontologies used for knowledge sharing.” Stanford Knowledge Systems Lab, August 23, 1993. KSL 93-04.

Gruninger M, Fox M (1995). “Methodology for the Design and Evaluation of Ontologies.” University of Toronto, April 13. Dept. of Industrial Engineering Technical Report.

Guyatt GH, Rennie D (1993). “Users' guides to the medical literature.” *JAMA* **270**(17): 2096-7.

Harbourt AM, Knecht LS, Humphreys BL (1995). “Structured abstracts in MEDLINE, 1989-1991.” *Bull Med Libr Assoc* **83**(2): 190-5.

Harlan WR (1994). “From the National Institutes of Health.” *JAMA* **271**(22): 1729.

Haynes R (1994). “Developing optimal search strategies for detecting clinically sound studies in Medline.” *Journal of the American Medical Informatics Association* **1**(6): 447-458.

Haynes RB (1990). “Loose connections between peer-reviewed clinical journals and clinical practice.” *Ann Intern Med* **113**(9): 724-8.

Haynes RB (1986). “How to keep up with the medical literature: I. Why try to keep up and how to get started.” *Ann Intern Med* **105**(1): 149-53.

Haynes RB (1990). “Online access to MEDLINE in clinical settings. A study of use and usefulness.” *Ann Intern Med* **112**(1): 78-84.

Haynes RB (1990). “More informative abstracts revisited.” *Ann Intern Med* **113**(1): 69-76.

Hayward RS (1993). “More informative abstracts of articles describing clinical practice guidelines.” *Ann Intern Med* **118**(9): 731-7.

Hearst M (1995). “TileBars: Visualization of term distribution information in full text information access.” *CHI*, Denver, CO:

Hearst M, Pedersen J (1996). “Reexamining the cluster hypothesis: Scatter/Gather on retrieval results.” *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 76-84. Zurich, Switzerland: SIGIR Forum.

Hemminki E (1982). “Quality of clinical trials--a concern of three decades.” *Methods Inf Med* **21**(2): 81-5.

Hersh WR, Hickam DH (1993). “A comparison of two methods for indexing and retrieval from a full-text medical database.” *Med Decis Making* **13**(3): 220-6.

Hetherington J, Dickersin K, Chalmers I, Meinert CL (1989). “Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians.” *Pediatrics* **84**(2): 374-80.

Hripcsak G (1995). “Unlocking clinical data from narrative reports: a study of natural language processing.” *Ann Intern Med* **122**(9): 681-8.

Hubbard SM, Martin NB, Thurn AL (1995). “NCI's cancer information systems--bringing medical knowledge to clinicians.” *Oncology* **9**(4): 302-6, 309 discussion 309, 313-4.

Humphrey SM (1992). “Indexing biomedical documents: from thesaural to knowledge-based retrieval systems.” *Artificial Intelligence in Medicine* **vol.4, no.5**:p. 343-7.

Huth EJ (1985). “Needed: an economics approach to systems for medical information [editorial].” *Ann Intern Med* **103**(4): 617-9.

Huth EJ (1989). “The underused medical literature [editorial].” *Ann Intern Med* **110**(2): 99-100.

ICH (1995). “Structure and content of clinical study reports.” International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, November 28. ICH - Efficacy Topic 3.

Imperiale TF, McCullough AJ (1990). “Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials.” *Ann Intern Med* **113**(4): 299-307.

Jeng GT, Scott JR, Burmeister LF (1995). “A comparison of meta-analytic results using literature vs. individual patient data. Paternal cell immunization for recurrent miscarriage.” *JAMA* **274**(10): 830-6.

Johnson ED, Sievert MC, McKinin EJ (1995). “Retrieving research studies: a comparison of bibliographic and full-text versions of the New England Journal of Medicine.” *Proc Annu Symp Comput Appl Med Care* 846-50.

Justice AC (1994). “Do readers and peer reviewers agree on manuscript quality?” *JAMA* **272**(2): 117-9.

Justice AC (1997). “Masking Author Identity in Peer Review: Does It Improve Peer-Review Quality.” In D. Rennie (Ed.), *3rd International Congress on Biomedical Peer Review and Global Communications*, page 18. Prague, Czech Republic: JAMA, The BMJ, and Project Hope.

Karp P, Riley M, Paley S, Pellegrini-Toole A (1996). “EcoCyc: Electronic encyclopedia of E.coli genes and metabolism.” *Nucleic Acids Research* **24**(1): 32-40.

Karp PD, Lowrance JD, Paley SM (1995). “Generic Knowledge-Base Editor.” [Web document] 9/27/97. Available: http://www.ai.sri.com/~gkb/welcome.shtml

Karp PD, Myers K, Gruber T (1995). “The Generic Frame Protocol.” In Mullish CS (Ed.), *International Joint Conference on Artificial Intelligence*, 768-774. Montreal, Quebec: Morgan Kaufmann.

Kassirer JP, Angell M (1995). “The Internet and the Journal [editorial].” *N Engl J Med* **332**(25): 1709-10.

Kassirer JP, Angell M (1995). “Redundant publication: a reminder.” *N Engl J Med* **333**(7): 449-50.

Kleijnen J, Knipschild P, ter Riet G (1991). “Clinical trials of homeopathy.” *BMJ* **302**(6772): 316-23.

Knoll E (1990). “The communities of scientists and journal peer review.” *JAMA* **263**(10): 1330-2.

Koes BW (1991). “Spinal manipulation and mobilisation for back and neck pain: a blinded review.” *BMJ* **303**(6813): 1298-303.

Lang T, Secic M (1997). *How to Report Statistics in Medicine*. Philadelphia, Pennsylvania: American College of Physicians.

Lau J (1992). “Cumulative meta-analysis of therapeutic trials for myocardial infarction.” *N Engl J Med* **327**(4): 248-54.

Lehmann H (1991). A bayesian computer-based approach to the physician's use of the clinical research literature. Doctoral dissertation in Medical Information Sciences, Stanford University.

LeLorier J (1997). “Discrepancies between meta-analyses and subsequent large randomized, controlled trials.” *N Engl J Med* **337**(8): 536-42.

Lenat D, Prakash M, Sheperd M (1986). “CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks.” *AI Magazine* **6**:65-85.

Levy A, Rousset M (1996). “Verification of knowledge bases based on containment checking.” *American Association for Artificial Intelligence*, Portland, Oregon:

Lindberg DA, Humphreys BL, McCray AT (1993). “The Unified Medical Language System.” *Methods Inf Med* **32**(4): 281-91.

Lomas J (1991). “Opinion leaders vs. audit and feedback to implement practice guidelines. Delivery after previous cesarean section.” *JAMA* **265**(17): 2202-7.

Mallery JC (1997). “Common Lisp Hypermedia Server (CL-HTTP).” [Web document] October 22, 1997. Available: http://www.ai.mit.edu/projects/iiip/doc/cl-http/homepage.html

Marshall E (1996). “Genome researchers take the pledge.” *Science* **272**(5261): 477-8.

McConnell J, Horton R (1996). “The message, the medium, and The Lancet.” *Lancet* **348**(9020): 74.

MedConnect (1996). “MedConnect.” [Web document] 10/1/96. Available: http://www.medconnect.com

Meinert CL, Tonascia S, Higgins K (1984). “Content of reports on clinical trials: a critical review.” *Controlled Clin Trials* **5**(4): 328-47.

Miller ET, Wieckert KE, Fagan LM, Musen MA (1995). “The development of a controlled medical terminology: identification, collaboration, and customization.” *Medinfo* **1**:148-52.

Moher D, Dulberg CS, Wells GA (1994). “Statistical power, sample size, and their reporting in randomized controlled trials.” *JAMA* **272**(2): 122-4.

Moher D (1996). “Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews.” *Lancet* **347**(8998): 363-6.

Moher D (1995). “Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists.” *Control Clin Trials* **16**(1): 62-73.

Moher D, Olkin I (1995). “Meta-analysis of randomized controlled trials. A concern for standards.” *JAMA* **274**(24): 1962-4.

Morris RD (1992). “A clinical trials database as a research tool in health care.” *Online J Curr Clin Trials* (14): [3293 words; 48 paragraphs].

Mulrow CD, Thacker SB, Pugh JA (1988). “A proposal for more informative abstracts of review articles.” *Ann Intern Med* **108**(4): 613-5.

Musen MA (1992). “Dimensions of knowledge sharing and reuse.” *Comput Biomed Res* **25**(5): 435-67.

Musen MA (1993). “PROTEGE-II: An Environment for Reusable Problem-Solving Methods and Domain Ontologies.” In Bajcsy R (Ed.), *IJCAI-93*, Chambery, Savoie, France:

Narine L, Yee DS, Einarson TR, Ilersich AL (1991). “Quality of abstracts of original research articles in CMAJ in 1989.” *Can Med Assoc J* **144**(4): 449-53.

National Breast Cancer Coalition (1997). “Clinical Trials Project.” [Web document] May, 1997. Available: http://www.natlbcc.org/clinic_trials.htm

National Library of Medicine NIH (1993). “Fact sheet, technical services division.” National Library of Medicine, March.

National Library of Medicine NIoH (1996). “NLM Online Databases and Databanks.” National Library of Medicine,

Neches R (1991). “Enabling technology for knowledge sharing.” *AI Magazine* **Fall 1991**:36-56.

NEJM (1996). “Information for authors.” *N Engl J Med* **334**(22): 1484.

Noy NF, Hafner CD (1997). “The State of the Art in Ontology Design. A Survey and Comparative Review.” *AI Magazine* **18**(3): 53-74.

Nurmohamed MT (1992). “Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis.” *Lancet* **340**(8812): 152-6.

Oliver DE, Shortliffe EH, Collaboratory I (1996). “Collaborative Development of the InterMed Vocabulary Model.” Section on Medical Informatics, Stanford University, SMI-96-0652.

Olkin I (1995). “Meta-analysis: reconciling the results of independent studies.” *Stat Med* **14**(5-7): 457-72.

Onghena P, Van Houdenhove B (1992). “Antidepressant-induced analgesia in chronic non-malignant pain: a meta-analysis of 39 placebo-controlled studies.” *Pain* **49**(2): 205-19.

Ontolingua (1996). “Library of Ontologies Table of Contents.” [Web document] 9/12/96. Available:http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/index.html

Ontolingua (1996). “Ontolingua Overview.” [Web document] 9/12/96. Available: http://ksl.stanford.edu/knowledge-sharing/ontolingua/ontolingua.html#documentation

Orfali R, Harkey D, Edwards J (1996). *The Essential Distributed Objects Survival Guide.* New York: John Wiley & Sons, Inc.

Peto R (1977). “Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples.” *Br J Cancer* **35**(1): 1-39.

Pfeifer MP, Snodgrass GL (1990). “The continued use of retracted, invalid scientific literature.” *JAMA* **263**(10): 1420-3.

Pitkin RM, Branagan MA (1997). “Can the Accuracy of Abstracts be Improved by Providing Specific Instructions?” In D. Rennie (Ed.), *3rd International Conference on Bio-*

medical Peer Review and Global Communications, page 26. Prague, Czech Republic: JAMA, The BMJ, and Project Hope.

Pocock SJ, Hughes MD, Lee RJ (1987). "Statistical problems in the reporting of clinical trials. A survey of three medical journals." *N Engl J Med* **317**(7): 426-32.

Powsner SM, Miller PL (1992). "Automated online transition from the medical record to the psychiatric literature." *Methods Inf Med* **31**(3): 169-74.

Powsner SM (1989). "Automated bibliographic retrieval based on current topics in hepatology: hepatopix." *Comput Biomed Res* **22**(6): 552-64.

Pratt W (1997). "Dynamic Organization of Search Results Using the UMLS." In D. R. Masys (Ed.), *American Medical Informatics Fall Symposium*, 480-484. Nashville, TN: Hanley & Belfus.

Public Law (1993). "Government Performance and Results Act of 1993." United States Government, Public Law 103-62.

Purcell GP (1996). Contextual document models for searching the clinical literature. Doctoral dissertation in Medical Information Science, Stanford University.

Raiffa H (1968). *Decision Analysis: Introductory Lectures on Choices Under Uncertainty.* New York: Random House.

Reisch JS, Tyson JE, Mize SG (1989). "Aid to the evaluation of therapeutic studies." *Pediatrics* **84**(5): 815-27.

Rennels G (1987). *A computational model of reasoning from the clinical literature*. Berlin: Springer-Verlag.

Rennie D (1994). "Reporting randomized controlled trials. An experiment and a call for responses from readers." *JAMA* **273**(13): 1054-5.

Rennie D, Yank V, Emanuel L (1997). “When authorship fails. A proposal to make contributors accountable.” *JAMA* **278**(7): 579-85.

Roberts JC, Fletcher RH, Fletcher SW (1994). “Effects of peer review and editing on the readability of articles published in Annals of Internal Medicine.” *JAMA* **272**(2): 119-21.

Rothman K (1986). *Modern Epidemiology.* 1 1. Boston: Little, Brown.

Rothman KJ, Michels KB (1994). “The continuing unethical use of placebo controls.” *N Engl J Med* **331**(6): 394-8.

Sackett D, Oxman A (1996). *Cochrane Collaboration Handbook.* Cochrane Collaboration, and the McMaster Health Informatics Research Unit. [Web Document] 4/5/96. Available: http://hiru.mcmaster.ca/cochrane/handbook/default.htm

Salton G (1991). “Developments in automatic text retrieval.” *Science* **253**(August 30): 974-980.

Sanders GD (1996). “Presentation and explanation of medical decision models using the World Wide Web.” In J. J. Cimino (Ed.), *American Medical Informatics Fall Symposium*, page 60-4. Washington, DC: Hanley & Belfus.

Scherer RW, Dickersin K, Langenberg P (1994). “Full publication of results initially presented in abstracts. A meta-analysis.” *JAMA* **272**(2): 158-62.

Schulz KF (1995). “Subverting randomization in controlled trials.” *JAMA* **274**(18): 1456-8.

Schulz KF, Chalmers I, Grimes DA, Altman DG (1994). “Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals.” *JAMA* **272**(2): 125-8.

Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995). “Empirical Evidence of Bias. Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials.” *JAMA* **273**(5): 408-412.

Science (1996). “Science's Next Wave.” [Web document] October 20, 1996. Available: http://193.54.210.193/nextwave/nextwave.html

Sim I (1995). “A meta-analysis of randomized trials comparing coronary artery bypass grafting with percutaneous transluminal coronary angioplasty in multivessel coronary artery disease.” *Am J Cardiol* **76**(14): 1025-9.

Sim I, Hlatky M (1996). “Growing pains of meta-analysis.” *BMJ* **313**(7059): 702.

Sim I (1997). “Quantitative Overview of Randomized Trials of Amiodarone to Prevent Sudden Cardiac Death.” *Circulation* **96**:2823-2829.

Simes RJ (1986). “Publication bias: the case for an international registry of clinical trials.” *J Clin Oncol* **4**(10): 1529-41.

Simes RJ (1987). “Confronting publication bias: a cohort design for meta-analysis.” *Stat Med* **6**(1): 11-29.

Singh SN (1995). “Amiodarone in patients with congestive heart failure and asymptomatic ventricular arrhythmia. Survival Trial of Antiarrhythmic Therapy in Congestive Heart Failure.” *N Engl J Med* **333**(2): 77-82.

Smith K (1992). “Respiratory muscle training in chronic airflow limitation: a meta-analysis.” *Am Rev Respir Dis* **145**(3): 533-9.

Smith R, Roberts I (1997). “An Amnesty for Unpublished Trials.” *BMJ* **315**:622.

SORT, The Standards of Reporting Trials Group (1994). “A proposal for structured reporting of randomized controlled trials. The Standards of Reporting Trials Group.” *JAMA* **272**(24): 1926-31.

Spitzer WO (1990). “Links between passive smoking and disease: a best-evidence synthesis. A report of the Working Group on Passive Smoking.” *Clin Invest Med* **13**(1): 17-42; discussion 43-6.

Steinbrook R, Lo B (1990). “Informing physicians about promising new treatments for severe illnesses.” *JAMA* **263**(15): 2078-82.

Stewart LA, Parmar MK (1993). “Meta-analysis of the literature or of individual patient data: is there a difference?” *Lancet* **341**(8842): 418-22.

Strang N, FICHTRE (1994). “Trial Summary Report Software and Data Base: Specifications and Preliminary Results.” *Controlled Clinical Trials*, page 77S. Seattle:

Strauss A, Corbin J (1990). *Basics of qualitative research: grounded theory procedures and techniques.* Newbury Park, CA: Sage Publications.

Teo KK, Yusuf S, Furberg CD (1993). “Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. An overview of results from randomized controlled trials.” *JAMA* **270**(13): 1589-95.

ter Riet G, Kleijnen J, Knipschild P (1990). “Acupuncture and chronic pain: a criteria-based meta-analysis.” *J Clin Epidemiol* **43**(11): 1191-9.

Thompson SG (1994). “Why sources of heterogeneity in meta-analysis should be investigated.” *BMJ* **309**(6965): 1351-5.

Tramer MR, Reynolds DJM, Moore RA, McQuay HJ (1997). “Impact of Covert Duplicate Publication on Meta-Analysis: A Case Study.” *BMJ* **315**:635-40.

Uschold M, Gruninger M (1996). “Ontologies: principles, methods and applications.” *Knowledge Engineering Review* **11**(2):93-136.

van der Weyden M, Bingham C (1997). “The Medical Journal of Australia Internet Peer-Review Trial.” In D. Rennie (Ed.), *3rd International Congress on Biomedical Peer*

Review and Global Communications, page 11. Prague, Czech Republic: JAMA, The BMJ, and Project Hope.

Walsh S (1994). "Assessing the quality of reports of randomized controlled trials used in systematic overviews abstracted." *Annu Meet Int Soc Technol Assess Health Care* 068): Abstract No. 068.

Wessely S, Joyce J, Rabe-Hesketh S (1997). "Reviewing the Reviews: The Case of Chronic Fatigue Syndrome." In D. Rennie (Ed.), *3rd International Congress on Biomedical Peer Review and Global Communications*, page 21. Prague, Czech Republic: JAMA, The BMJ, and Project Hope.

Wielinga B, Schreiber A, Breuker J (1991). "KADS: A modeling approach to knowledge engineering." *Knowledge Acquisition* **4**:5-53.

Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB (1995). "Preliminary assessment of the effect of more informative (structured) abstracts on citation retrieval from MEDLINE." *Medinfo* **2**:1457-61.

Wilkes MS, Kravitz RL (1995). "Policies, practices, and attitudes of North American medical journal editors." *J Gen Intern Med* **10**(8): 443-50.

Williams J, Jr., Holleman D, Jr., Samsa GP, Simel DL (1995). "Randomized controlled trial of 3 vs. 10 days of trimethoprim/sulfamethoxazole for acute maxillary sinusitis." *JAMA* **273**(13): 1015-21.

Williams N (1995). "Europe opens institute to deal with gene data deluge [news]." *Science* **269**(5224): 630.

Williamson J, Goldschmidt P, Colton T (1986). *The quality of medical literature: An analysis of validation assessments.* Medical Uses of Statistics. Boston: Massachusetts Medical Society. 1st, ed. 370-391.

Williamson JW (1989). “Health science information management and continuing education of physicians. A survey of U.S. primary care practitioners and their opinion leaders.” *Ann Intern Med* **110**(2): 151-60.

Wyatt JC, Altman DG, Heathfield HA, Pantin CF (1994). “Development of Design-a-Trial, a knowledge-based critiquing system for authors of clinical trial protocols.” *Comput Methods Programs Biomed* **43**(3-4): 283-91.

Zelen M (1979). “A new design for randomized clinical trials.” *N Engl J Med* **300**(22): 1242-5.