December 1991

Report No. STAN-CS-92-1402

Thesis

PB96-149422

A Bayesian Computer-Based Approach to the Physician's Use of the Clinical Research Literature

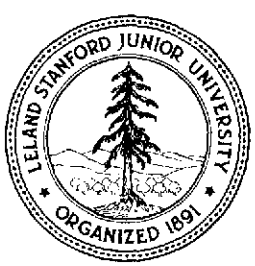
by

Harold P. Lehmann

Departments of Computer Science and Medicine

Stanford University

Stanford, California 94305



BIBLIOGRAPHIC INFORMATION

PB96-149422

Report Nos: STAN-CS-92-1402

Title: Bayesian Computer-Based Approach to the Physician's Use of the Clinical Research Literature.

Date: cDec 91

Authors: H. P. Lehmann.

Performing Organization: Stanford Univ., CA. Dept. of Computer Science.

Sponsoring Organization: *National Library of Medicine, Bethesda, MD.

Contract Nos: NLM-LM-07033, NLM-LM-05208

Type of Report and Period Covered: Doctoral thesis.

NTIS Field/Group Codes: 57E (Clinical Medicine), 44T (Data & Information Systems), 88B (Information Systems)

Price: PC A13/MF A03

Availability: Available from the National Technical Information Service, Springfield, VA. 22161

Number of Pages: 298p

Keywords: *Clinical medicine, *Information retrieval, *Bibliographic databases, Scientific literature, Physicians, *Medical research, *Bayesian decision theory, THOMAS.

Abstract: In this dissertation, the author considers the thesis that Bayesian decision theory can provide the foundation for a computer-based environment that helps physicians to use the research literature. As a basis for evaluating approaches to solving the literature problem, the author argues for the use of Bayesian statistics over classical statistics. The shift to Bayesian statistics requires a change in the paradigm within which research data are evaluated. To show that the new paradigm can be implemented in a functioning computer system, the author has developed a prototype system, called THOMAS, that gives the physician reader a number of capabilities.

A BAYESIAN COMPUTER-BASED APPROACH TO THE PHYSICIAN'S USE OF THE CLINICAL RESEARCH LITERATURE

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF MEDICAL INFORMATION SCIENCES

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

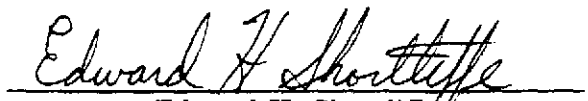
DOCTOR OF PHILOSOPHY

By

Harold P. Lehmann

December 1991

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

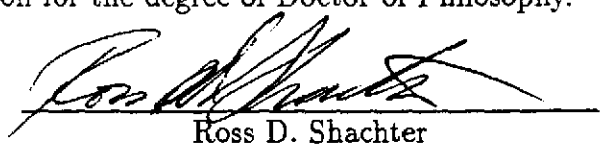


Edward H. Shortliffe
(Principal Advisor)

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

Byron W. Brown, Jr.
(Department of Health Policy and Research)

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



Ross D. Shachter
(Department of Engineering–Economic Systems)

Approved for the University Committee on Graduate Studies:

Dean of Graduate Studies

Abstract

Physicians need to understand the clinical research literature if they are to make informed clinical decisions; yet the techniques required for using the literature in this way are difficult for many clinicians to acquire and to use. I call this dilemma of needing information yet being unable to extract it the *literature problem*. To date, automated statistical methods used to solve the literature problem have been limited in the degree to which they can represent methodological and domain concepts that are crucial to the physician who must take clinical action. In this dissertation, I consider the thesis that Bayesian decision theory can provide the foundation for a computer-based environment that helps physicians to use the research literature.

As a basis for evaluating approaches to solving the literature problem, I develop a *knowledge-level analysis* of the problem. On the basis of this analysis, I argue for the use of Bayesian statistics over classical statistics. The shift to Bayesian statistics requires a change in the paradigm within which research data are evaluated.

To show that the new paradigm can be implemented in a functioning computer system, I have developed a prototype system, called THOMAS, that gives the physician reader a number of capabilities: (1) to analyze a study in a structured way, (2) to examine a study in multiple ways, (3) to incorporate domain knowledge and prior belief into an analysis, (4) to incorporate methodological knowledge into an analysis, (5) to determine the optimal therapy, (6) to examine the change in belief in any parameter of the underlying statistical model, (7) to compare the beliefs in any two

parameters, and (8) to examine the sensitivity of any posterior belief or decision to different prior beliefs. THOMAS operates in the domain of randomized clinical trials that compare the effects of different drugs on a patients' survival.

To incorporate any methodological concern, THOMAS (1) requires a statistical submodel for the concern, and (2) requires a visual metaphor though which the physician can communicate the particular concern. THOMAS contains submodels for the methodological concerns of loss to followup, withdrawal, noncompliance, crossing-over, and measurement unreliability. The system uses the visual metaphor of the *patient-flow diagram* for physician input.

In the course of each consultation, the user implicitly constructs a statistical model appropriate to the study and to the user's reading of that study. The construction process is based on representing the statistical models as *hierarchical, typed influence diagrams*, a structure that limits the interactions among parameters in a statistical model. Prespecified *construction steps* dictate how the primitive methodological submodels are pieced together. A *metadata-state diagram,* containing basic methodological knowledge assessed from a statistical expert and from the methodological literature, limits the sequence of construction steps the user is allowed.

The system has been evaluated positively by a small number of its intended users. The representational framework can be extended to deal with methodological concerns beyond THOMAS's current abilities.

This dissertation extends the Confidence Profile Method of Eddy, Hasselblad, and Shachter (1991) by automating its use. In addition, this dissertation puts on the medical-informatics agenda the question of how physicians should act on the basis of research data, and suggests novel methods for storing, using, and retrieving the contents of the biomedical research literature.

Acknowledgments

I have been fortunate to have a thesis committee in which each member has given so much direction and support. I am indebted to Ross Shachter for his acting as a true mentor in instruction, in criticism, in guidance, and in reassurance. Bill Brown has exemplified patience, intelligence, and support. His willingness to countenance contrary opinions in a positive and constructive manner will serve me as an enduring role model of the true gentleman and scientist. I thank Ted Shortliffe for his constant encouragement and for his tenacity at keeping this work on track and intelligible.

Members of the REFEREE group gave me the initial impetus to work on this problem, and offered continuing support and constructive evaluation of my work. Bruce Buchanan has always had an open mind and a bemused look to balance my enthusiasm. Diana Forsythe worked mightily to inject a human component into research that could otherwise tie itself into equational knots. Dan Feldman kept the work clinically relevant. Marty Chavez was a fellow enthusiast, and I learned much from his intense and imaginative mind.

Research always occurs in a community, and the Medical Informatics Lab (especially the "Bayesian mafia") provided the ideal community in which to work. Greg Cooper helped me to formulate this research early on, and has remained a constant source of insight. David Heckerman and Eric Horvitz have taught me much about the profound implications of taking the decision-analytic position, bolth through their conversations and their actions. Eddie Herskovits's and David Klein's abilities to see

through any nonsense and foolishness would stand me in good stead. Brad Farr and Holly Jimison gave insightful feedback during our seminars with Ross Shachter. Mark Frisse and Thierry Barsalou have been persistent cheerleaders. Mark Musen helped to organize the knowledge-acquisition material. Other members (current and past) of the lab—Ingo Beinlich, Larry Fagan, Adam Galper, Michael Kahn, Les Lenert, Richard Lin, Geoff Rutledge, Mike Shwe, and Jaap Suermondt—have been helpful in innumerable ways. I thank Blackford Middleton, John Hornberger, and Doug Owens, from "across the way," for help in placing this work into medically meaningful contexts.

I thank David Bergman and Susan Galel, along with Ted Shortliffe and Bill Brown, for consenting to be subjects in this work.

Lyn Dupré executed her typically thorough, professional, and instructional review of the dissertation. Bill Poland implemented the posterior-mode updating algorithm. I thank the National Library of Medicine, under Grant LM-07033, for its direct support of this work and, under Grant LM-05208, for its support of the SUMEX-AIM Resource. I thank Darlene Vian and Lynne Hollander for making work such a pleasure. And I thank the pediatric department at Kaiser–Permanente at Hayward for their willingness to give an considerate ear to, and a thoughtful appraisal of, ideas not generally part of the clinical environment.

Finally, this work would have been impossible without support on the home front. I thank my wife Rivka for her constant and unstinting love, support, and patience during this eventful time, and our new bundle of energy, Amalya, for giving us such joy.

Contents

List of Tables

List of Figures

Chapter 1

The Literature Problem

The clinical research literature provides important information for physicians making clinical decisions, yet clinicians generally have limited skills for appraising such studies critically. In addition, they have difficulty using statistical tools to help in analyzing such literature. In this chapter, I introduce this problem, which I call the *literature problem,* and I present my thesis, that the problem can be solved within a computer-based decision-analytic framework, specially tailored for the problem.

Section 1.1 introduces the literature problem. Section 1.2 provides a specific example, which I then use as an illustration throughout the dissertation. Section 1.3 places the literature problem in the context of medical informatics and Section 1.4 explains the need for a novel solution to the problem. My research was developed in response to this need, and in Section 1.5, I summarize my thesis, discussing its conceptual components and the representational structure of the solution I propose. This section also introduces the domain within which I evaluate the thesis, the domain of randomized clinical trials. Section 1.6 introduces the program, THOMAS, that embodies the solution, and presents a demonstration of the program in use. Finally, Section 1.7 provides a reader's guide for the remainder of the dissertation.

1.1 The Clinical Scientific Literature and Biostatistics

Physicians appeal to the clinical research literature when they want to rationalize, justify, or explain their actions. The clinical research literature is important in this regard, because research papers provide the medical community with its highest-quality information for making clinical decisions. These decisions may involve individual patients,[1] and classes of patients (Yusuf et al., 1985). The federal government (Field and Lohr, 1990) and other third-party reimbursers are increasingly demanding justifications of specific medical practices (Eddy, 1990), and they, too, look to the research literature.

The medical scientific community uses *biostatistics* as its formal framework for interpreting clinically derived, scientific information. Members of the community use statistical methods to arbitrate questions of scientific validity. The techniques involve qualitative understanding of methodology and quantitative analysis of data. Among the methods most relevant to clinicians are those used in studies that compare treatment alternatives.

Despite basic biostatistics courses in most preclinical curricula, physicians tend to lack statistical knowledge and need help in applying statistical methods. Current strategies for providing such help include seeking ways to reinforce the statistical concepts taught in medical school, offering postgraduate continuing education in statistics, publishing reviews articles, and providing a variety of methodology checklists and guidelines.

A novel strategy for providing such help is the introduction of computer-based

[1] For example, the patient-specific problems analyzed in Dr. S. Pauker's series entitled "Clinical Decision Making Rounds" in the journal *Medical Decision Making*.

expert systems. *Expert systems* constitute a class of computer program that provide users with expert-level advice in domains where such expertise tends to be ill-structured and judgmental (Hayes-Roth et al., 1983). Experience over the past 15 years shows that such programs can indeed perform at a high level of expertise (Smith et al., 1985; Heckerman et al., 1989).

In this dissertation, I shall explore the *literature problem*: How should we judge clinical action and reach patient-specific management decisions on the basis of results in the clinical research literature. My goal is to formulate a framework for helping physicians to solve the literature problem, and to describe an implementation of that formulation in a working computer program.

1.2 The Metoprolol Example

As a concrete example of the literature problem, imagine you are a physician treating a 55-year-old white man who has just had a heart attack (myocardial infarction, MI) and who has been brought into the hospital almost immediately after symptoms began. Besides needing to stabilize his acute cardiovascular status, you want to prevent worsening of his general cardiac condition. You have heard that a drug, metoprolol, which belongs to the beta-blocker class of medications, might improve his cardiac status. It has, however, serious known side effects. Should you administer the drug?

You have access to a paper by Hjalmarson and colleagues (1981) (see Figure 1.1) that reports that acute administration of metoprolol is associated with subsequent fewer deaths than placebo administration over the first three months after the acute heart attack. The observed mortality rates were 8.9 percent in the placebo group and 5.7 percent in the metoprolol group. The strength of the conclusion is suggested by the classical statistical measure, the p value, of 0.012, which is less than the

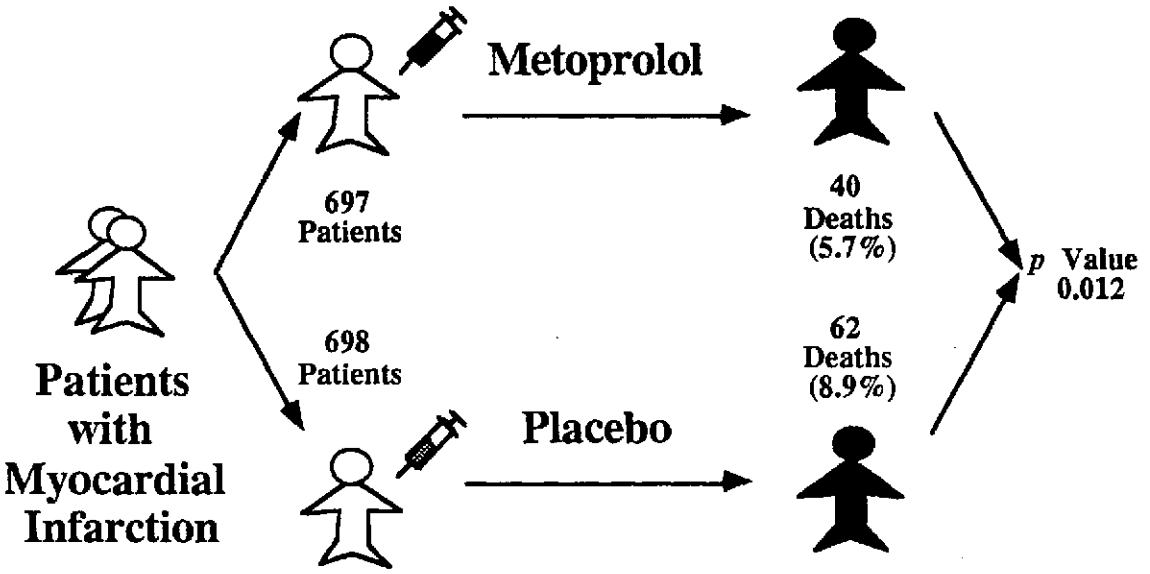


Figure 1.1: The metoprolol study. This patient-flow diagram for patients enrolled in the study shows the sequence of states study patients experienced: the initial state of suffering a myocardial infarction, the assignment state to metoprolol or placebo, and the endpoint state of surviving or not. The terminal state contains a statistical summary of the study results—the p value. (Source: adapted from Hjalmarson, Å., Herlitz, J., et al., Effect on mortality of metoprolol in acute myocardial infarction, Lancet 2(9251):823–827 (1981).)

traditional threshold of 0.05, suggesting the superiority of metoprolol. This suggestion does not answer definitively an important question: Does the observed difference in mortality rates offset the possible side effects? A further piece of information raises a methodological concern: In a close reading of the article, you find that fully 19 percent of the patients assigned to receive metoprolol in fact were not treated with the test medication. How much should this departure from protocol affect your assessment of the study's validity and your decision to give metoprolol to your own patient?

I shall use this example throughout this dissertation for clarifying the issues involved in solving the literature problem. The particular report by Hjalmarson and colleagues (Hjalmarson et al., 1981) has played an important role in the practice of

cardiology. For instance, Yusuf and colleagues (Yusuf et al., 1985) included this study in their meta-analysis of the use of beta-blockers after acute myocardial infarction. By the end of the 1980s, the use of such drugs became the standard of care (Antman and Braunwald, 1990).

1.3 Medical Informatics and the Literature Problem

The field of medical informatics has given relatively little attention to statistical issues. An informal review of the proceedings of MEDINFO and the Symposium on Medical Applications in Medical Care between the years 1984 and 1990 shows that about 50 articles out of 1200 (4 percent) could be related to issues of data analysis of scientifically collected data; even fewer refer to the problem of using the published scientific literature. Rennels (1987), whose work is based on classical statistics, comes closest to tackling the literature problem, and I shall refer to it several times in the course of this dissertation.

We can presume that the medical informatics community's general indifference to statistics derives from physicians' belief that statistics is best left to statisticians. Statisticians have, in fact, built systems to help statisticians of different levels of proficiency to perform data analyses (Gale, 1986a; Nugent, 1986; Oldford and Peters, 1988; Tierney, 1990), and even to help in the design of clinical studies (Weiner et al. 1987). I shall argue, however, that such systems are inappropriate for use by physicians in light of the demonstrated limited knowledge physicians have of sophisticated statistics. I claim that for decision making by end users (such as physicians), an additional layer of interface and semantics is needed beyond those supplied by existing programs and by classical statistics itself.

An important consequence of this indifference to statistics is that physicians

have lost control of an important source of information—clinical research data—with the result that this information—ostensibly collected to aid practitioners in their ministrations—has had less effect on daily practice than investigators have expected (Gelband, 1983). Because an important goal of medical informatics is to give physicians control over the mass of information deluging them today, this dissertation demonstrates that improving physicians' statistical reasoning and their evaluation of the clinical research literature should be on the agenda of the medical informatics community.

1.4 Problems with Classical Statistics

I propose that a novel type of computer system is needed to help solve the literature problem. Although I shall discuss the full argument for this proposal in Chapter 2, I shall discuss two important issues here: physicians have difficulty with statistics, and classical biostatistics is unable to deliver important services needed by physicians.

1.4.1 Physicians' Difficulty with Statistics

Familiarity with a minimal fund of knowledge—descriptive statistics and elementary statistical tests—would give a physician access to two-thirds of clinical research articles, according to a review by Emerson and Colditz (1983) of published reports. As shown in Table 1.1, investigators have surveyed whether physicians possess that minimal fund of knowledge. Wulff and colleagues (Wulff et al., 1987) sent a questionnaire of nine biostatistical problems to 250 subjects randomly selected from the national registry of Danish physicians. The questions covered the basic concepts to which Emerson and Colditz refer. The median score on this questionnaire was 2.4, out of a maximum of 9. Among physicians who said they did "understand all the expressions" (Wulff et al., 1987, p. 4) in the survey, the median score was 4.1. This

well-designed and executed study suggests that average clinicians are not familiar with basic statistical concepts, even if they think they are. Weiss and Samet (1980) sent a 10-item questionnaire to 141 internal-medicine house staff and attending physicians at an academic institution. This instrument garnered a mean score of 7.4, out of a maximum of 10. In a separate study, two questions were sent by Friedman and Phillips (1981) to 685 pediatric residents nationwide. Twenty percent answered the correlation question correctly; 50 percent answered the p-value question correctly.

Table 1.1: Data documenting physician statistical fund of knowledge.

Source	Subjects	Sample Size	Number of Questions	Summary Score (Method)
Wulff et al., 1987	Random Physicians	250	9	2.4 (Median)
Weiss and Samet, 1980	Academic Internists	141	10	7.4 (Mean)
Friedman and Phillips, 1981	Pediatric Residents	684	1	20 (Percent)
			1	50 (Percent)

A possible cause of the difficulty encountered when physicians use statistics is the numerical nature of the domain. The relatively better scores obtained by Weiss and Samet (1980) may result from the fact that many of the questions dealt with methodological concepts, whose qualitative nature physicians found more within their ken than the number-based items of the other investigators. Another source of difficulty, however, is the counterintuitive nature of certain constructs in classical statistics.

Pocock, Hughes, and Lee (1987) document problems investigators have with classical statistics. Their review of published reports of controlled clinical trials from major journals (see Section 1.6.2) discloses problems in a number of areas: multiple

analyses of data, misuse of p values as measures of strength of evidence, misapplication of hypothesis testing for arriving at conclusions, inappropriate analyses of subgroups of patients, and improper examination of data before the formal conclusion of the study. Note that each goal (e.g., to analyze the data in different ways) desired by the different investigators is reasonable, and that their errors lay in their misapplication of classical statistics due to their misconstrual of classical statistical notions.

The now-classical example that demonstrates the counterintuitivity of the classical statistical approach is the controversy over the University Group Diabetes Program study of oral hypoglycemic medication (UGDP, 1970). In this study, the medication, intended to help in the management of diabetes, apparently caused some patients to die: 26 of 204 patients died in the experimental group and 10 of 205 patients died in the placebo group, an apparent double death rate. The study was not designed to detect differences in mortality rates, and the trial was terminated earlier than originally intended, as a result of the examination of these data before the formal conclusion of the study. The departure from the initial protocol and the altered focus in the results both cast doubt on the validity of the statistical conclusions. Meinert and Tonascia (1986) review the history of the controversy. By its conclusion, the debate had involved several universities and national institutions. Diamond (1983) shows how the confusion and controversy resulted primarily from a basic misunderstanding of the p value. I shall explore the difficulties with the p value more extensively in Section 3.5.

1.4.2 Limitations of the Classical Paradigm

Over the past 80 years, the classical paradigm has been successful in helping physicians to distinguish useful from useless—and even harmful—therapy. The paradigm has been a linchpin in the biomedical community's drive to promote a scientific approach to medical care (Feinstein, 1985). Nevertheless, there are profound difficulties

with classical statistics. I shall discuss them extensively in Section 3.5; here, I shall preview an important problem: classical statistics does not offer capabilities that physicians need. Two such functions are the ability to express uncertainty and the ability to recommend a decision for a specific individual.

Readers have uncertainty about a domain both before and after they have read an article. The degree of readers' uncertainty plays heavily in their decision whether to act or to seek further information. Classical statistics' primary locution for expressing uncertainty is the *confidence interval* (see page 67). A study may be faulted simply for not reporting these intervals (Gardner and Bond, 1990) and some methodologists see the confidence interval as a solution to the problem of overreliance on the p value as a measure of the strength of evidence (Felson et al., 1990). Yet, it is commonly understood within the statistical community that consumers of statistical reports misconstrue the true semantics of the confidence interval (Rubin, 1984). As properly understood, confidence intervals communicate uncertainty in an estimate of the parameter involved (e.g., mortality rate); the true value may still be any number (Armitage, 1983, p.109). As commonly misunderstood, they express how likely it is that the true value of the parameter lies within the reported interval.

Readers also want to make decisions. The statistical subspecialty of *statistical decision theory* (Wald, 1950) has been developed over the past 40 years for this purpose. Its focus, however, is on making policy decisions that affect many people (or studies) over time, rather than on making an individually tailored choice. Modifying the derived global policies for individual cases involves nonobjective, heuristic procedures, as both statisticians (Brown, 1984) and expert-system designers (Rennels, 1987) know.

1.5 Thesis: An Alternative Paradigm

Would it be possible to modify classical statistics so that we could obtain the benefits, yet correct the problems? *Bayesian* statistics (de Finetti, 1974; DeGroot, 1970; Lindley, 1972; Savage, 1972; Box and Tiao, 1973; Berger, 1985) is an approach that is designed to do just that. Specifically, Bayesian statistics[2] allows for the expression of uncertainty, through subjective probability, and for the recommendation of individual action, through *decision analysis* (Howard and Matheson, 1981), a formal discipline concerned primarily with helping decision makers to take action in individual cases.

The thesis of this dissertation is that, *decision analysis and the Bayesian paradigm can form the basis of a computer-based environment to aid physicians making clinical decisions on the basis of scientific data from the clinical research literature.* I shall explore the components of this thesis in the following six subsections, I shall outline the representational structure needed in the seventh, and I shall summarize the evaluation of the thesis in the eighth.

1.5.1 Clinical Focus of Decision Making

There are many uses physicians make of scientific data from the research literature. One is to guide further reading, using an article to decide what is important to learn. Another is to take clinical action. Our concern with clinical decision making implies that our methods will be grounded in the *clinical significance* of any conclusion from a study, as opposed to the strictly *statistical* significance of the results.

[2]Named after the Reverend Thomas Bayes (1702–1761)

1.5.2 Published Scientific Data as Primary Source

There are many sources of information physicians can use for making decisions. A database of past observations and therapeutic actions is one; when such data are rigorously collected, they may form the basis for formal statistical analysis or for matching a current situation with similar circumstances that have occurred in the past. Nonscientifically collected, published observations, such as case histories, constitute another source, for which statistical methods are inapplicable. We shall narrow the scope of this dissertation to scientific studies as they are published. Rennels (1987) and Eddy and colleagues (1991) have done work in this narrowed scope that comes closest in spirit and in detail to the work presented in this dissertation. I shall discuss Rennels' work in Section 3.6, and Eddy's work in Section 4.7.

1.5.3 Applicability of Decision Analysis

The primary requirements of decision analysis are that any uncertainty of the decision maker is represented by probability, that any preference of the decision maker is represented by utility, and that the optimum decision for the decision maker is the action that maximizes the expected utility (Howard and Matheson, 1981; Berger, 1985).

Decision analysis is appropriate in domains where (1) uncertainty is a major concern, (2) the stakes are high enough that a formal analysis is worth the effort, and (3) there is an individual decision maker. My definition of the literature problem satisfies these conditions: (1) Biostatistics is, by its nature, concerned with uncertainty. (2) The stakes involved are often life and death, as well as unpleasant outcomes and substantial monetary expense. (3) The focus is the individual clinician who must take action on behalf of a particular patient. In building artificial-intelligence (AI) systems, knowledge engineers have taken a decision-analytic approach in a variety

of domains, including diagnosis (Heckerman et al., 1990), learning (Star, 1987; Buntine, 1989), vision (Levitt, 1988), and control of inference (Breese and Fehling, 1988; Horvitz et al., 1989), but not, to date, in statistical consulting.

Some underlying assumptions of decision analysis violate basic principles held by classical statisticians. Chapter 2 provides my arguments to justify violating these principles in my solution to the literature problem. I shall summarize the line of argument here. Recent approaches to knowledge acquisition in expert systems pay attention to the differentiation between the *goals* of interest in the domain and the *procedures* by which the goals are met. As difficult as it is to achieve this separation in many domains, it is even more difficult in statistics. In most domains, there is no articulation or theory of the procedures used by domain experts to achieve the desired goals. Statistics, however, seems to contain just such an articulation—the body of statistical methods we are enjoined to use—which blurs the separation between the goals and the implementation of the domain. The basic principles of classical statistics constitute the foundation of this body of methods. I have found that, to achieve the differentiation between goals and procedures necessary to solve the literature problem, we must tease apart those principles necessary for the solution from those which make it difficult to solve.

1.5.4 Extending the Bayesian Paradigm

The kernel of the Bayesian solution to the literature problem is as follows. The investigators summarize their results in a form called the *likelihood function* (see page 53). The reader then combines her[3] prior *knowledge* of the domain with the likelihood function to arrive at her *posterior belief*. This belief can then be used by

[3]As a convention throughout this dissertation, the physician is female; the patient, the statistician, the investigator, and the system builder are male; and the machine is neuter. The term *analyst* refers either to the statistician or to the reader, depending on context. These conventions allow the reader to follow the discussion more clearly.

the decision-analytic engine to arrive at the optimal recommendation for action.

Various approaches to *Bayesian reporting* (Hildreth, 1963; Dickey, 1973; Berger, 1985; Hilden, 1987) combine the likelihood functions conveyed by investigators and the prior beliefs expressed by readers to arrive at depictions of the corresponding posterior beliefs. These approaches assume a single likelihood function for each study, which corresponds to a single way of analyzing the data or of examining the study. This assumption makes sense if we characterize the relationship between the investigator and the reader as a separation of labor: The investigator reports what happened, the reader updates her beliefs. This premise does not, however, empower the reader to apply her knowledge of methodology (what can go wrong in a study) or of pathophysiology (how the specific medical context affects the study) in arriving at a conclusion. Therefore, the concept of representing prior knowledge must be broadened. This extension is a contribution of this dissertation. I justify my heuristic approach in Section 4.4 and describe it fully in Sections 6.3.2 and 6.3.3.

1.5.5 Dynamic Computer-Based Environment

Computers have generally been necessary for any practical application of Bayesian statistics, because solutions need numerical integration and other computation-intensive procedures (Goel, 1988). Knowledge-based methods would appear to offer a solution to the literature problem because there are multiple sources of knowledge (statistical, methodological, domain, and clinical) needed to solve the problem, and because they can provide structure to the precarious act of building a solution (Efron, 1986).

To clarify the process of structuring a Bayesian statistical analysis, I shall contrast it with the approach of classical statistics (see Section 3.5 for a more complete examination). The classical procedure calls for the statistician first to choose a statistical model appropriate for the problem, then to choose the best test suited to that model and to the data available, and then to execute the test. The final inference

is a result of the test chosen and the result of the test. Expert systems founded on this approach are *diagnosing systems,* because their primary task is to select the appropriate model and its corresponding test. *Selection* is the central task, because creating a new test for any given statistical model is too difficult an activity (worthy of a doctoral dissertation in its own right).

The Bayesian approach, however, allows the analyst to *construct* an arbitrary model that he feels is appropriate. Regardless of the model constructed, the approach calls for a single inference procedure applied to all statistical models—*probabilistic updating* (see Section 4.3.1). The final inference uses the result of that calculation in a well defined (and uniform) way to arrive at a recommendation. Thus, the central task of a Bayesian system is the construction of the appropriate model, and the system must create a model anew for each problem, doing so on demand. A Bayesian system is, therefore, primarily a *planning,* or even a *knowledge-acquiring,* system. The contrast between the two approaches is discussed more fully in Sections 3.5 and 4.6.

My solution to the problem of dynamic Bayesian statistical-model construction depends on two knowledge representations:

- *Influence diagrams* are data structures that have been used increasingly over the past 10 years for representing uncertainty in probability-based expert systems. The use I shall make of influence diagrams for creating statistical models is a novel application of this representation (see Section 7.3).

- *Metadata-state diagrams* are state-transition networks I have created specifically for this dissertation (see Section 7.2). They comprise two sets of knowledge: *what* can happen to patients at different stages of a study, and *how* those circumstances affect the growing statistical model.

1.5.6 Physicians Readers as the Target Users

We wish to empower physician readers to apply their knowledge of methodology and of pathophysiology to solving the literature problem within a computer-based environment. The *interface* between the physician and the decision-analytic approach therefore must be a major concern in this dissertation. The interface must be based on *semantics* familiar to physicians; the challenge is to find interactive metaphors that such users find intuitive and that have operations that can be translated into procedures consistent with the decision-analytic approach. I shall use two such metaphors: checklists and patient-flow diagrams.

Checklists are used by many methodologists (Warren, 1981; Gehlbach, 1982; Feinstein, 1985; Haynes et al., 1986; Sacks et al., 1987; L'Abbé et al., 1987; Reisch et al., 1989) for organizing information in studies. Physicians find intuitive the action of checking which problems need attention or of choosing among possible choices. In a computer-based environment, a checklist can be made dynamic in that different questions come into view depending on the choice made by the user. Such an environment has the advantage that the sequence of actions can be guided by the machine. My use of checklists is discussed in Section 6.2.

Patient-flow diagrams are used by many journal-article authors to communicate to readers what happened to patients over the course of a study. These diagrams group together patients who are similar in some way; I call these groups *cohorts*. Figure 1.1 showed an example of such a diagram for the metoprolol study. In a computer-based environment, these diagrams can made dynamic, allowing the reader to communicate to the machine attributes of each patient group, such as the total number of patients in the group and the methodological problems experienced by those patients. I describe these diagrams fully in Section 6.3.2.

1.5.7 Representational Structure

With this background of the components of my proposed solution to the literature problem, we shall construct the representation structure needed to solve the problem by examining the information needed for a decision-analytic solution. I shall build up the resulting framework from its components. Technical terms will be defined in Chapter 4.

We begin with the *decision analysis* (Figure 1.2). This process uses a *decision model* (not shown) which consists of the outcomes of interest, their utilities (reflecting mortality–morbidity tradeoffs),[4] and the parameters that determine their likelihoods. The analysis takes as one if its inputs probability distributions for the beliefs in the values of those parameters. Since the distributions are based on observed data, they are distributions *posterior* to the reading of the study. The analysis also takes as its input *preferences* of the patient that reflect his mortality–morbidity tradeoffs. The decision analysis produces, as its output, the *optimal decision.* In the metoprolol study, the outcome of interest is the death of a patient, whose likelihood is parameterized by a single number, the mortality rate, and the decision is whether to treat with metoprolol.

There are two ways to produce the posterior probabilities, once the reader has read the paper. She could assess her posterior beliefs directly. However, this method ignores the limited statistical sophistication of the reader, leaving implicit all the methodological considerations we want to make explicit, and ignores the known probabilistic processes that generated the data. The second way is to help the reader with this complex task by performing an analysis that takes the probabilistic processes and methodological considerations explicitly into account: a *Bayesian statistical analysis* (Figure 1.3). This analysis takes as its input a statistical model that includes the

[4]For purposes of this dissertation, financial costs of treatments or outcomes will not be considered in the utility models. The approach used could be extended, however, to handle multiattribute models.

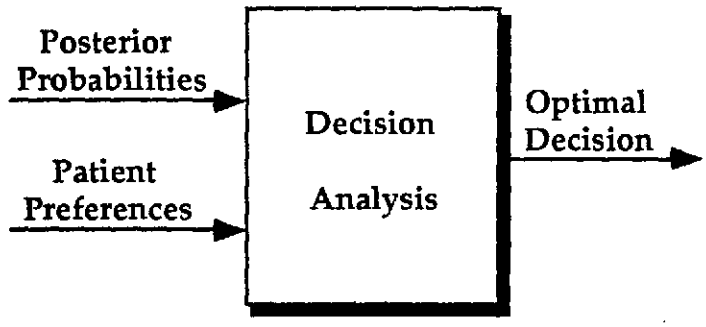


Figure 1.2: Decision analysis. This information-flow diagram depicts the data needed by the process to produce the optimal decision. The structure of the decision model (not shown) is fixed by the knowledge engineer.

parameters of interest and parameters relating to biases and errors perceived by the reader as relevant to the understanding of the report. The analysis also needs as its input the prior beliefs about every parameter. Bayesian statisticians have expended much effort in developing ways to compute the posterior distribution from a given model (Eddy, 1989).

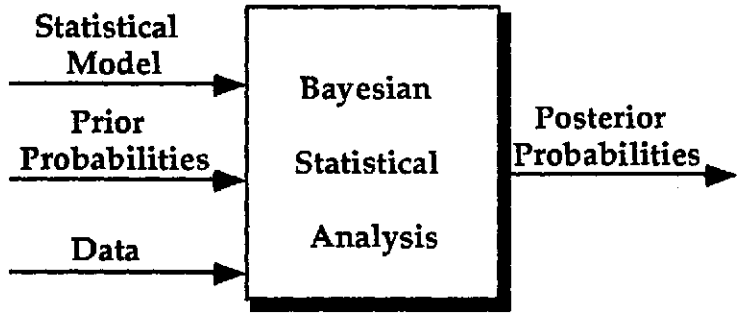


Figure 1.3: Bayesian statistical analysis. This information-flow diagram depicts the data needed by the statistical analysis to produce the posterior probabilities required as input items by the process in Figure 1.2.

The statistical model required for the Bayesian analysis (Figure 1.3) must be constructed on the basis of the methodological concerns (Figure 1.4). The structural element responsible for this task is the *Bayesian methodological formulation.* The structure and function of this element are major contributions of this dissertation,

and are the subjects of Chapter 7.

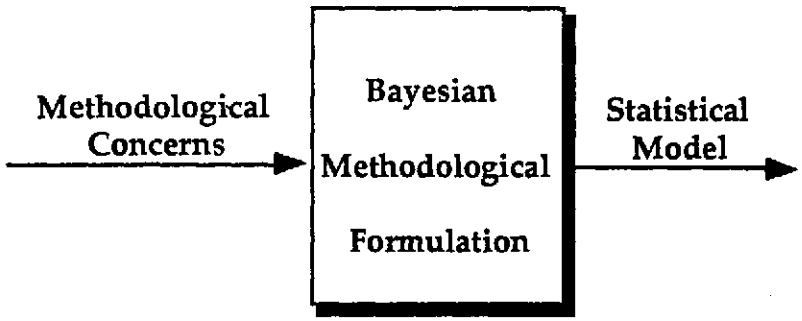


Figure 1.4: Bayesian methodological formulation.

The fundamental input into the system is information from the research report itself (see Figure 1.5). The contents of the clinical research report include numerical and text data. Because computer-based processing of the text in an article would be an unrealistic demand on current natural-language processing abilities, my approach expects the reader to interpret the contents of the paper for the machine, arriving at the appropriate numerical input and the appropriate methodological concerns, such as the identity of the central quantitative elements of the study and of threats to internal validity.

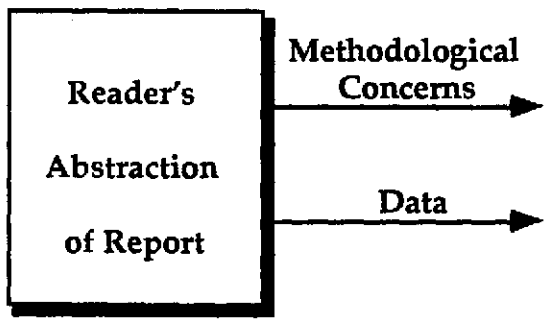


Figure 1.5: Reader's abstraction of report. The numerical and text information from the written report of the study must be transformed by the reader into numerical data and methodological concerns. The data are required by the Bayesian statistical analysis (Figure 1.3), whereas the methodologic concerns are required for the computer's formulation of the statistical model (Figure 1.4).

The entire model is given in Figure 1.6. In this completed model I have made explicit the reader's background knowledge as the source for the prior probabilities for parameters in the statistical model, and her posterior knowledge as being the net output and goal of the entire process. Her prior knowledge in fact enters everywhere in the framework: in choosing methodological concerns, in choosing parameters of interest, and in establishing the decision model.

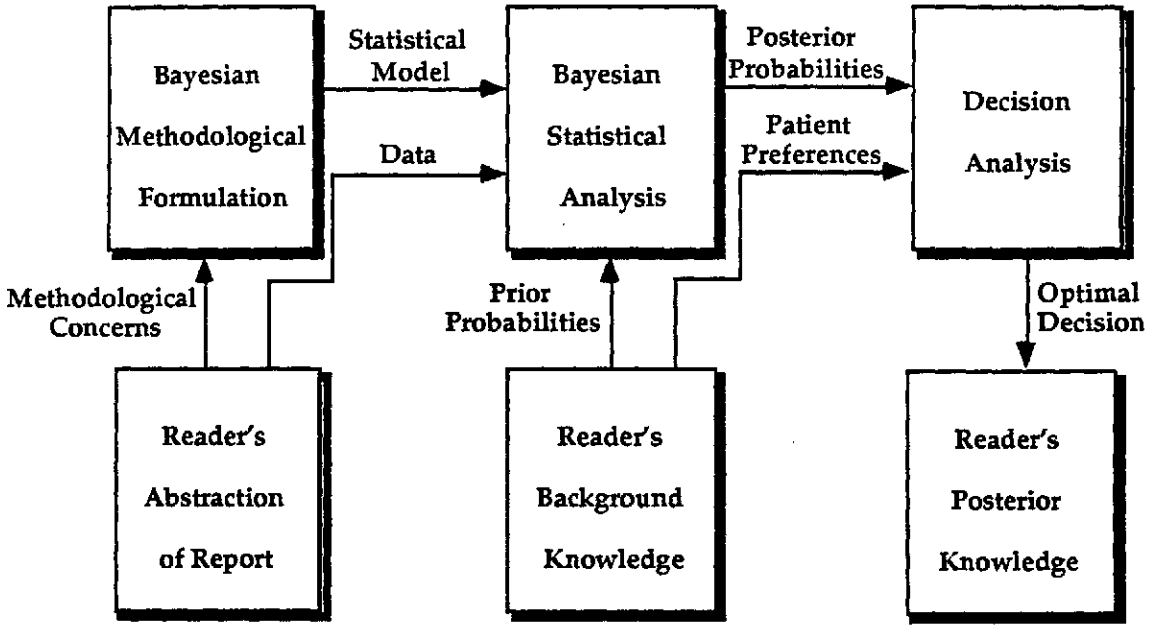


Figure 1.6: The Bayesian strategy. This diagram is a composite information-flow diagram, constructed from Figures 1.2, 1.3, 1.4 and 1.5, depicting the Bayesian model to assist with using a clinical research article for clinical decision making.

1.5.8 Evaluation

The evaluation of the thesis has two components: evaluating the representational integrity of the framework I propose, and evaluating the prototype system built to implement the framework. The representational integrity will be demonstrated in the course of this dissertation, as I show the various specifications the framework is

designed to meet. The evaluation of the prototype system has two parts: demonstrating that it meets its specifications, and that a physician can use the system for its intended purpose. I shall discuss these evaluations in Chapter 8.

The thesis, as I have presented it, encompasses large areas of medical research. In this dissertation, I shall focus on one particular domain (randomized clinical trials), and I shall demonstrate the thesis with a computer program that helps physicians to apply the results of that class of research to clinical decision making.

1.6 THOMAS

In this section, I shall describe the prototype system I have built to demonstrate the concepts of the thesis. I shall first describe the program in general terms, and then I shall present a demonstration of using the program in the context of the metoprolol example.

1.6.1 The Program

THOMAS[5] is my prototype computer system[6] that embodies the concepts in this thesis.

The system provides the physician user with the following abilities:

- To analyze a study in a structured way
- To examine a study in multiple ways
- To incorporate domain knowledge into an analysis

[5]Named in honor of Reverend Thomas Bayes. Blackford Middleton made me aware of the resonance with the concept of ***doubting Thomas***, which is appropriate in this context of uncertain information and high-stakes decisions.

[6]The system is implemented on the Macintosh computer, with Allegro Common Lisp as the language for the inference engine, and HyperTalk as the language for the user interface. The occasional crowded computer screens result from the window-size limitations of the version of HyperCard used.

- To incorporate methodological knowledge into an analysis
- To determine the optimal therapy
- To examine the change in belief in any parameter
- To compare the beliefs in any two parameters
- To examine the sensitivity of any posterior belief or decision to different prior beliefs

I shall describe, in Chapter 5, how the system delivers these services. The output comes closest in spirit to the *z-test for proportions* in classical statistics. Thus, THOMAS is a prototype for helping physicians perform Bayesian statistical analyses, obviating the need to use classical-statistical tests. The following sections give a flavor of the interaction; Chapter 6 goes into more detail.

1.6.2 Methodological Domain

Physicians can use THOMAS when reading a study report of a particular research design—the *randomized clinical trial* (RCT), a type of *controlled clinical trial* (CCT). CCTs are prospective studies in which patients are assigned to one of two or more *interventions*, such as drug and placebo, and are followed over time for the occurrence of some endpoint, such as mortality or a specific morbidity. RCTs are studies where the assignment to therapy is made randomly. The purpose of randomization is to limit possible biases in the study. RCTs are the current gold standard for clinical research (Feinstein, 1985) and, although they represent only about 1 percent of the articles published each year (Meinert et al., 1984), their influence in academic and public discourse is proportionally much greater (Gelband, 1983). RCTs are the particular domain that we shall examine. In particular, we will be concerned with RCTs that compare the effects of drug therapy on patient mortality.

1.6.3 A Sample Interactive Session

THOMAS's interaction with the user is divided into five tasks, as indicated by the master checklist in Figure 1.7: defining the problem, describing the study, examining the statistical results, examining the recommended decision, and finishing the consultation.

1.6.3.1 Definition of the Problem

THOMAS needs two pieces of information to define the problem. First is the identity of the medications involved in the study. Figure 1.8a shows the user telling THOMAS that metoprolol is the experimental drug; not shown is the user informing the machine that placebo is the control drug.

The second piece of information conveys the physician user's judgment about the mortality–morbidity tradeoff. In Figure 1.8b, the user tells the machine that, in her judgment, 6 months of increased life expectancy for the patient would be required to justify using the drug, to balance the implicit morbidities: side effects of arrhythmias, added cost, and added hassle of taking metoprolol. The input value of this *pragmatic difference* is where the physician encodes her prior knowledge about the domain as it applies to the patient at hand.[7]

1.6.3.2 Describe the Study

The process of describing the study entails selecting the study to be examined, specifying the design, communicating current knowledge about the drugs and methodology involved, and describing the study execution.

[7]Note that this evaluation is a composite indicator of utilities of the possible morbidities and other risks. Although this assessment could be approached with a formal decision analysis, for purposes of this research, I have chosen simply to request a single utility measure. This measure is discussed more fully in Section 5.3.

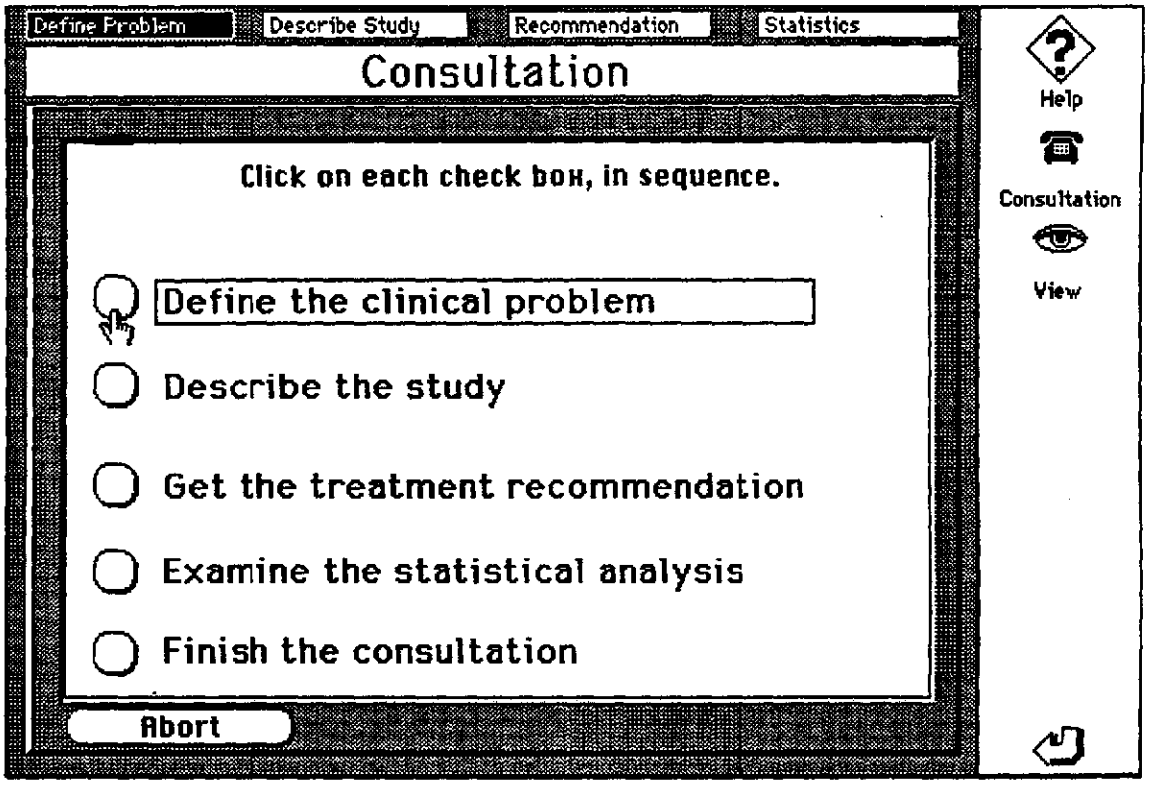
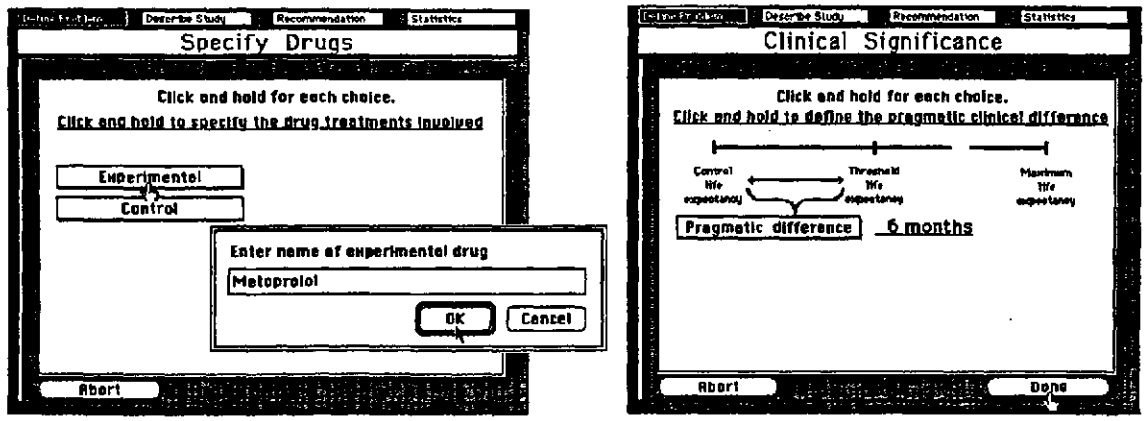


Figure 1.7: THOMAS's top level checklist. This screen image from the program shows the five tasks the user completes in performing an analysis. The first four tasks correspond to parts of the overall design of Figure 1.6: The task *Define the clinical problem* includes the task of giving the system patient preferences. The task *Describe the study* corresponds to the task in inputting methodological concerns, prior probabilities, and study data. The task *Get the treatment recommendation* comprises the system's performing the Bayesian statistical analysis and decision analysis and displaying the optimal decision. The task *Examine the statistical analysis* allows the user to examine the posterior probabilities generated by the system.

The panel of rectangles at the top of the screen helps users to keep track of their progress through the tasks in the course of completing an analysis. The icons on the right side of the screen refer to ancillary functions. Users begin their traversal through the dynamic checklist by clicking on the indicated button.



(a) (b)

Figure 1.8: Definition of the problem in THOMAS. (a) Specification of the experimental drug. This screen image is an overlay of two images. In the first, the user indicates the choice of selecting the *Experimental* drug. A pop-down menu (not shown) allows the user to select a previously known choice or to enter a new drug name. The second image shows the name the user has typed in (*Metoprolol*). (b) Specification of the pragmatic difference. The graphic on the screen suggests the meaning of this difference.

1.6.3.2.1 Selection of the Study There are two parts to the task of study selection: specification of the citation and commencement of the analysis of the study (Figure 1.9). To specify the citation, THOMAS allows the user to select from a list of citations that grows as new citations are typed in. THOMAS does not have access to the contents of the chosen article except through information entered by the user.

The machine enables the user to create a sequence of analyses. The initial analysis is called the *baseline description,* by default. The user may return several times to modify this and subsequent descriptions by dividing groups of patients initially lumped together (see Figure 1.15). The sequence of analyses creates a tree of analyses, where a descendent analysis is a modification of its ancestor. These alternative analyses enable the user to answer questions regarding the effects of different methodological concerns, either alone or in concert.

1.6.3.2.2 Specification of the Design Before getting the details about the study, THOMAS must know basic information about the general design of the study. The details of the statistical-construction algorithm depend on what design is used.

There are two components to the design: the architecture of the study (Figure 1.10a) and the outcome of the study (Figure 1.10b). THOMAS at present knows about only one design, the two-arm randomized clinical trial (depicted in Figure 1.1), and about only one outcome, mortality.

1.6.3.2.3 Communication of Current Knowledge The Bayesian paradigm demands that an agent assess her prior beliefs before viewing information that could update those beliefs. In the statistical domain, this assessment translates into evaluating beliefs about *parameters,* such as mortality rates. THOMAS requests such information before allowing the user to input data from the study.

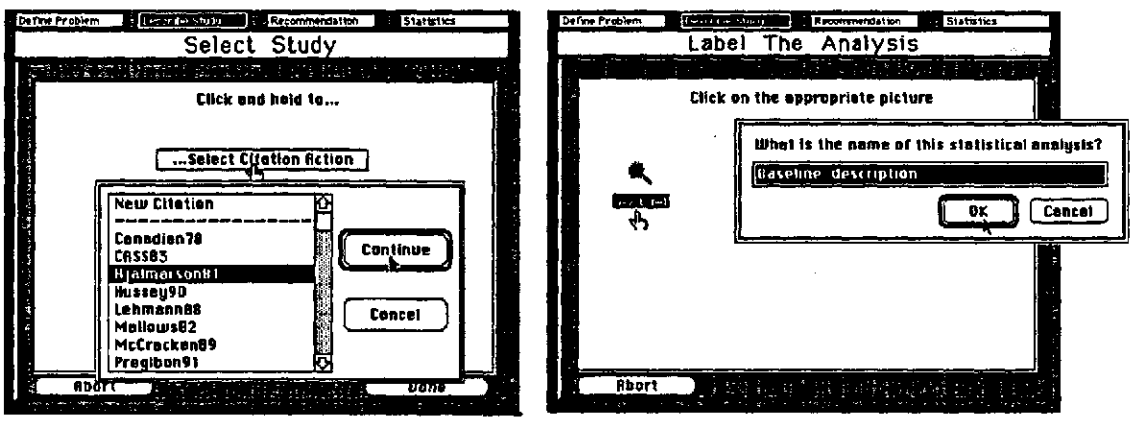


Figure 1.9: Selection of the study in THOMAS. (a) Specification of the citation. This image is an overlay similar to Figure 1.8a. (b) Specification of the name of the analysis. Another image overlay, this screen image shows the user choosing to define a new analysis with the name *Baseline description.* If she so wished, the user could return to this screen to define other analyses that modify the baseline description or each other, and thereby generate a tree of analyses.

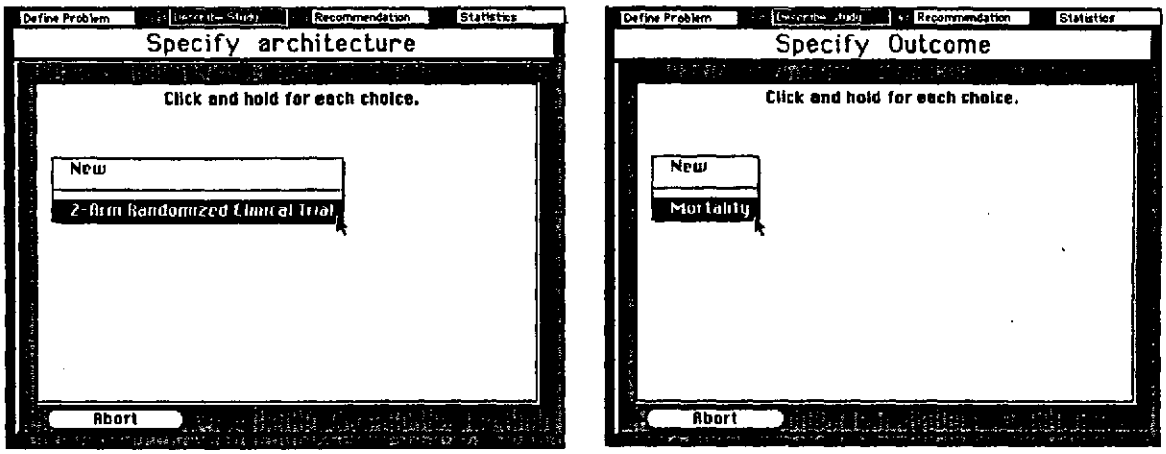


Figure 1.10: Specification of the study design in THOMAS. (a) Specification of the architecture. Using the recurrent interface of the pull-down menu, the machine ascertains the study architecture. THOMAS currently knows about only the *2-Arm Randomized Clinical Trial.* Hence, the *New* option is greyed out on the screen and disabled. (b) Specification of the study outcome. THOMAS currently allows only the outcome *Mortality.* Issues of morbidity were previously addressed in defining the required pragmatic difference (Figure 1.8b).

The user must first select the parameter she wishes to consider from a list generated by the machine. In this example, this list comprises two names (Figure 1.11a): the *population mortality rate in patients assigned to placebo* and the *population mortality rate in patients assigned to metoprolol.* THOMAS assembles the names from information already input by the user and from knowledge it has about RCTs. The names of the medications come from information input by the user (Figure 1.8a). The parameter type (*mortality rate*) derives from the name of the outcome (Figure 1.10b) and that outcome's method of assessment (count of death events).

Having chosen to consider the placebo mortality rate, the user requests help in understanding the assessment task. Figure 1.11b shows the machine's response to her request: Prior knowledge about placebo gives THOMAS domain knowledge.

In Figure 1.11c, the user has selected to claim total ignorance about mortality

rates in patients with acute myocardial infarctions treated with placebo. Note that, although this choice will result in an analysis most comparable to the classical statistical analysis (see Figure 1.11b), it is probably an inaccurate assessment of the physician user's prior knowledge. This disparity between the mathematical demand for presumed objectivity and the domain reality of intersubject variation in knowledge and disagreement is at the heart of the contrast between classical and Bayesian approaches.

Not shown in the figures is the user's similar choice to assume total ignorance of the mortality rate in patients treated with metoprolol.

1.6.3.2.4 Description of Study Execution Figure 1.12 shows the patient-flow diagram used to inform the machine about what happened to patients in the metoprolol study; this approach is unique to THOMAS, and is a major distinguishing feature with respect to programs such as that implemented for the Confidence Profile Method (Eddy, 1989) (see Section 4.7). Figure 1.12a shows the diagram at the start of the description, before the user has entered any information, and Figure 1.12b displays the diagram at the end of the description, after all the information from the study (see Figure 1.1) has been input. The diagrams are dynamic in that the name of each patient cohort depends on how the cohort was formed, and in that the user can specify a cohort's history in any order she wishes. Figure 1.12a shows the user informing THOMAS via a keypad interface about the total number of patients assigned to placebo. In Figure 1.12b the machine gives the user an opportunity to change her description before continuing on to the statistical analysis (Figure 1.15 shows such a change). When the user indicates that she is finished, the statistical model is complete; the machine automatically performs the probabilistic-updating procedure.

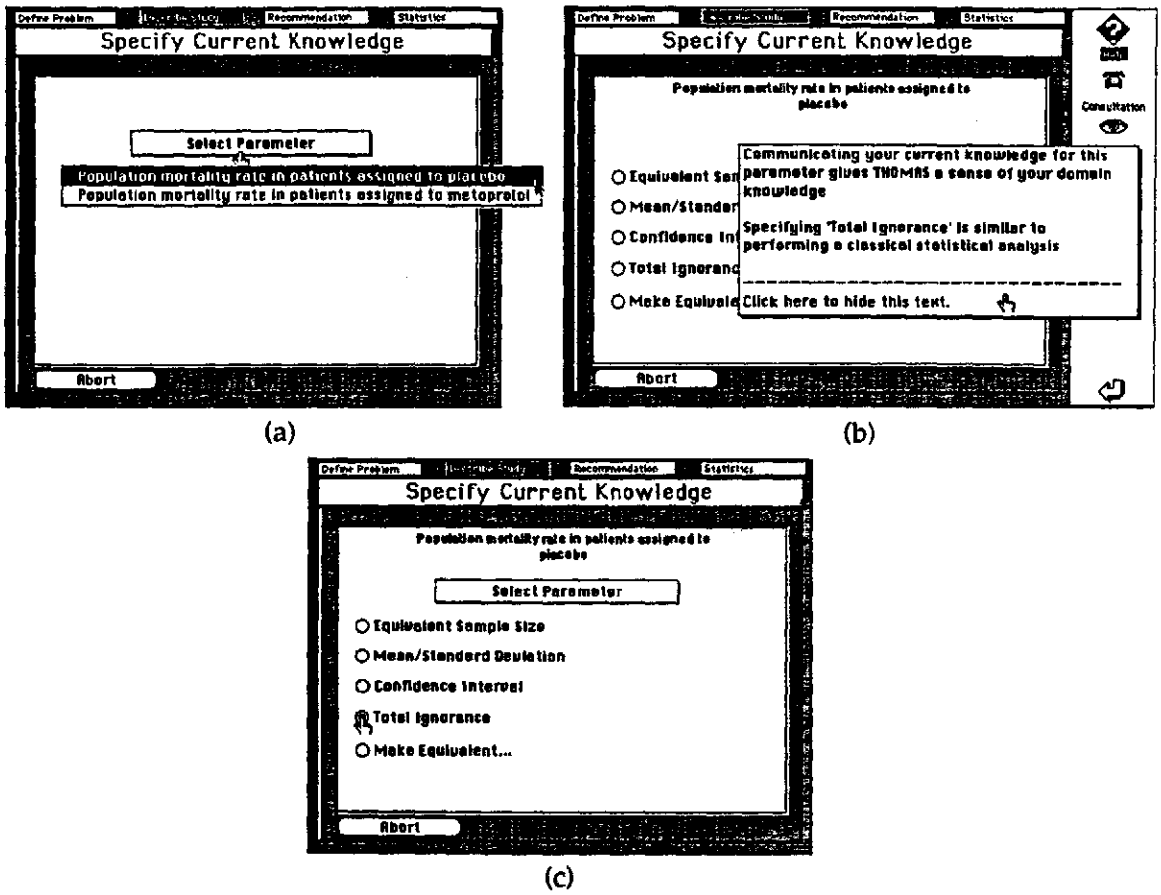
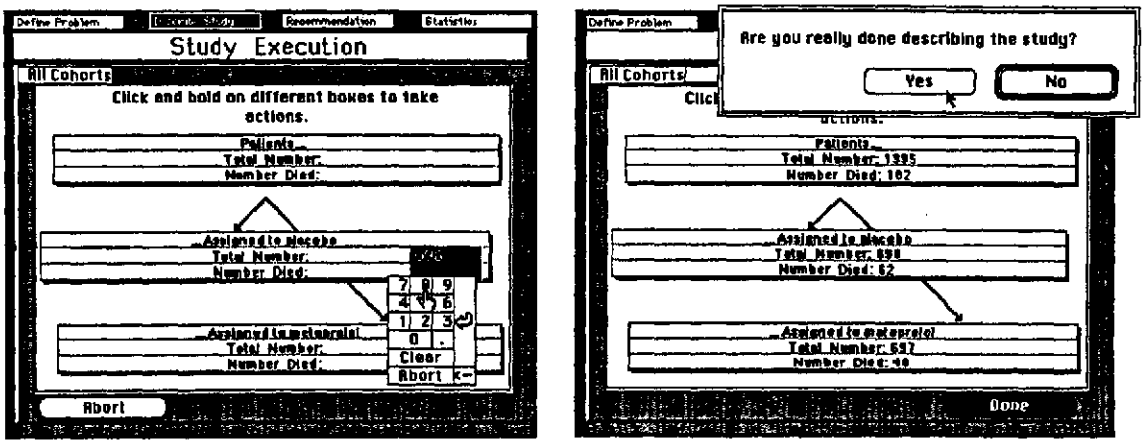


Figure 1.11: Specification of prior knowledge in THOMAS. This task is labeled *Specify Current Knowledge* to encourage the user to think about her personal experience and knowledge of the domain. (a) Choice of a parameter. An overlay of two images, this figure shows the user selecting a parameter (*Population mortality rate in patients assigned to placebo*) from a set constrained and generated by THOMAS. (b) Acquisition of help. The user has asked for an explanation of the task (note the highlighted help icon in the upper right). (c) Specification of the actual knowledge. The user has a choice of numerical and qualitative types of specification; see Section 6.3.3 for a full discussion. Once a choice is made, the machine requests information about other parameters (hence, the *Select Parameter* box), until all needed parameters are accounted for.



(a) (b)

Figure 1.12: Specification of the study execution in THOMAS. Both of these screens contain a patient-flow diagram. The boxes refer to cohorts of patients in the study. Each line of a cohort's box is active. Placing the mouse icon over the first line induces the machine to present a choice of actions for the cohort (see Figure 1.15a). Placing the icon over the second line makes the program request input about the total number of patients in the cohort. Placing the icon over the third line makes the system ask for the number of patients who experienced the endpoint of the study. (a) Specification of the number of patients in a cohort. A keypad interface pops up for mouse-based entry. (b) Completion of the description. This image shows the patient-flow diagram for the baseline description of the metoprolol study (see Figure 1.1), with all numbers entered in the appropriate lines. The machine computes the sums, and displays them in the root cohort.

1.6.3.3 Examination of the Decision

To place the results of the analysis into clinically meaningful terms, THOMAS computes the life expectancy contingent on the belief distributions calculated from the probabilistic update. Figure 1.13a depicts the graphs of these computed life expectancies. Figure 1.13b shows the results of such a computation, taking into account the threshold for clinical significance the user made when she defined the clinical problem (see Figure 1.8b). In this case, the increase in life expectancy for metoprolol over that for placebo is greater than the minimum demanded by the physician user, so the

machine recommends that she administer metoprolol.

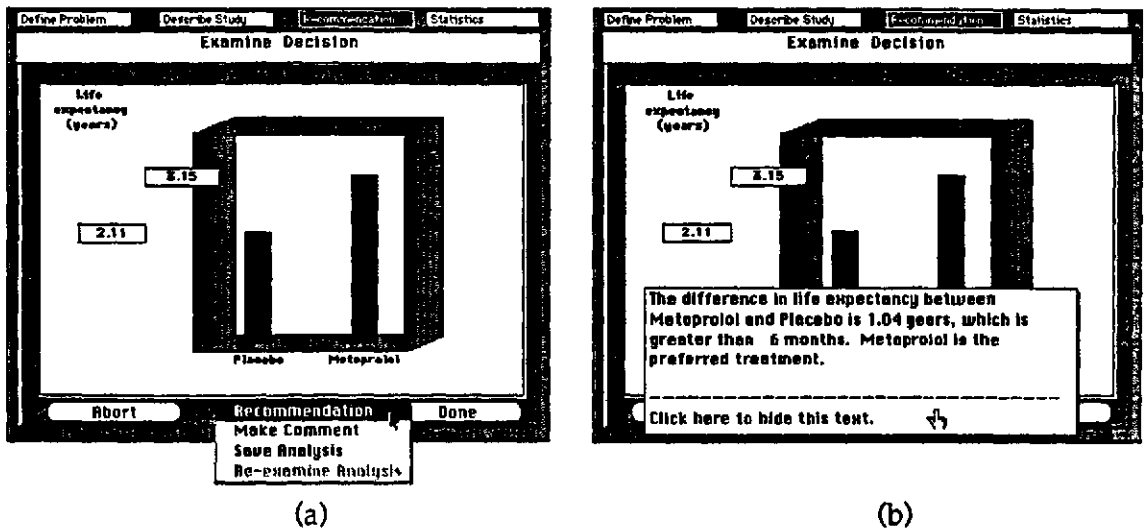


Figure 1.13: Examination of THOMAS's recommendation. (a) Life expectancy. This screen displays a bar graph of a patient's life expectancy, given each therapy, based on the belief distributions THOMAS has calculated (Figure 1.14d). (b) Recommendation. This verbal statement is based on the threshold set by the user in Figure 1.8b.

1.6.3.4 Examination of Statistical Results

The user's third major task is to review the results of the probabilistic updating, although this task is optional if she is interested in only the clinical implications. Two aspects of this examination are the review of the results and a performance of a sensitivity analysis.

1.6.3.4.1 Review of Results The user has the option of allowing the machine to guide the examination (see Figure 1.14a). During the study-description task (see Section 1.6.3.2.4), not only does the machine create the appropriate statistical model, but it applies canonical questions in which the user would probably be interested, and it associates with every question a pair of parameters to be compared. In completing

the task of examining the statistical analysis, the user can choose to pose the questions the machine has formulated for her (see Figure 1.14b). For instance, in every consultation, the user is likely to ask if the mortality rate in the treatment group is statistically significantly different from the mortality rate in the control group. THOMAS answers this question by comparing the difference in beliefs (see Figure 1.14c) in the corresponding parameters (see Figure 1.14d). This report is the Bayesian measure closest to the notion of a p value, although this measure is irrelevant in arriving at a decision.

1.6.3.4.2 Performance of Sensitivity Analysis Although it is not a formal part of the decision-analytic sequence of Figure 1.6, sensitivity analysis plays an important part in decision analysis. THOMAS allows the user to perform two types of sensitivity analysis: varying prior belief and varying the structure of the statistical model.

The user may reanalyze the study, using different prior beliefs. Such reanalysis might, for instance, show the effect on the final conclusion of the assumption of total ignorance.

The user may also reanalyze the study, taking into account different methodological problems. Figure 1.15a shows how the user would deal with the issue of the 19 percent of patients in the metoprolol study who withdrew from therapy (see page 4). The figure shows that THOMAS knows that patients assigned to the experimental treatment can undergo four types of protocol departures: They might be lost to followup, they might be withdrawn from the study, they might not comply with

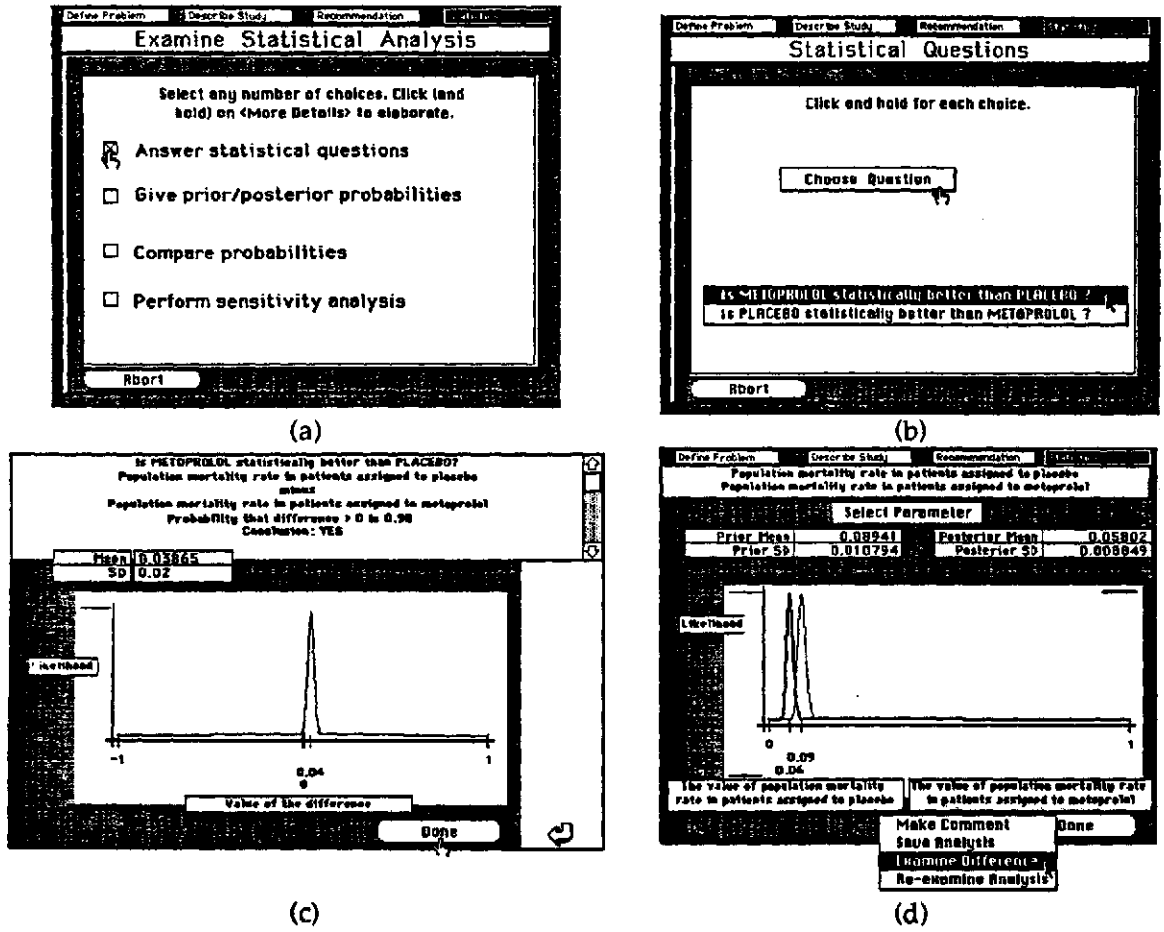


Figure 1.14: Examination of statistical results, guided by THOMAS. These screen images indicate the sequence of events. (a) When the user chooses to examine the statistical analysis, she is shown this screen, which gives her a number of options. The user has chosen to seek answers to questions posed by the machine. (b) An overlay of two images, this figure shows the user choosing the first of two questions suggested by THOMAS. (c) To answer the question, the machine presents a report at the top of the screen, containing the question selected, the parameters concerned, the probability that the difference is positive, and the conclusion (in this case, yes, the mortality rates are statistically significantly different). The examination of the difference is presented in several ways: the verbal report just described; the numerical summary, giving the mean and standard deviation of the posterior belief in the difference; and the graph, showing the distributions of the posterior belief in each parameter. (d) The user may examine the belief in each of the mortality rates individually, and may ask to review the report about the difference between them. The belief distributions shown are centered at their respective mean posterior beliefs, 0.089 (placebo) and 0.058 (metoprolol).

therapy, or they might be given the control treatment instead. The metadata-state diagram, set by the knowledge engineer but not shown, determines what protocol departures are possible for patients in different cohorts.

Figure 1.15b displays a comparison between the belief in the metoprolol mortality rate taking the withdrawals into account (the debiased *population mortality rate in patients assigned to metoprolol*) and not taking them into account (the raw *observed mortality rate in patients assigned to metoprolol*). The value of the debiased mortality rate is lower than that of the observed mortality rate, but is more uncertain. These adjustments makes sense on two accounts. First, we examine the value. The observed metoprolol mortality rate is a mixture of two debiased mortality rates: the debiased metoprolol mortality rate (patients who were assigned to metoprolol and who received it) and the debiased baseline-care mortality rate (patients who were assigned to metoprolol but who received baseline care, which is equivalent to receiving placebo). The previous analysis regarding the placebo mortality rate told us that patients not treated with metoprolol have a higher mortality rate. The debiased metoprolol mortality rate therefore must be lower than the mixture mortality rate of 5.7 percent, which indeed it is (4.5 percent). Second, we examine the uncertainty. We note that the uncertainty of the debiased mortality rate (standard deviation of 0.03) is larger than that of the observed mortality rate (standard deviation 0.009). The increase in uncertainty between the observed and debiased mortality rates makes intuitive sense, because the inference regarding the debiased mortality rate is further removed from the actual data.

Rennels' ROUNDSMAN program (Rennels, 1987) produces the same behavior, but resorts to potentially subjective heuristics to do so. THOMAS generates this behavior from a principled and formal basis.

1.7 Guide to the Reader

The remainder of the dissertation develops and fills out the concepts introduced in this chapter. The next three chapters justify the approach taken in this dissertation to solving the literature problem: In Chapter 2, I present a framework for constructing an expert system to solve the problem, based on recent approaches to knowledge acquisition. In Chapter 3, I examine the classical-statistical domain in light of that framework, demonstrating weaknesses of classical procedures for achieving desired goals. Readers acquainted with the contents of the statistical domain may wish to skip this chapter, except for those portions describing the use of influence diagrams in representing statistical models, scattered throughout the chapter, and the critique of the classical approach presented in Section 3.5. In Chapter 4, I present Bayesian concepts in more detail and demonstrate how the Bayesian paradigm is expected to solve the problems of classical statistics. Readers familiar with Bayesian notions can bypass this chapter, except, perhaps, for Section 4.6, which presents a critique of the Bayesian approach in light of the knowledge-acquisition principles developed earlier.

The subsequent three chapters present my approach to implementing the Bayesian approach: Chapter 5 delineates the design of the prototype system, including the system's intended user, restricted domain, utility model, probabilistic models, and approach to statistical parameters. Chapter 6 discusses the interface problems and solutions particular to a Bayesian statistical system intended for nonstatisticians. Chapter 7 presents the novel data structures and algorithms needed for allowing the system to perform dynamic statistical-model construction.

In Chapter 8, I discuss the evaluation of the thesis with respect to its meeting the various specifications laid out in the early chapters of the dissertation and with respect to use of THOMAS by physicians. I close the dissertation with my conclusions, in Chapter 9.

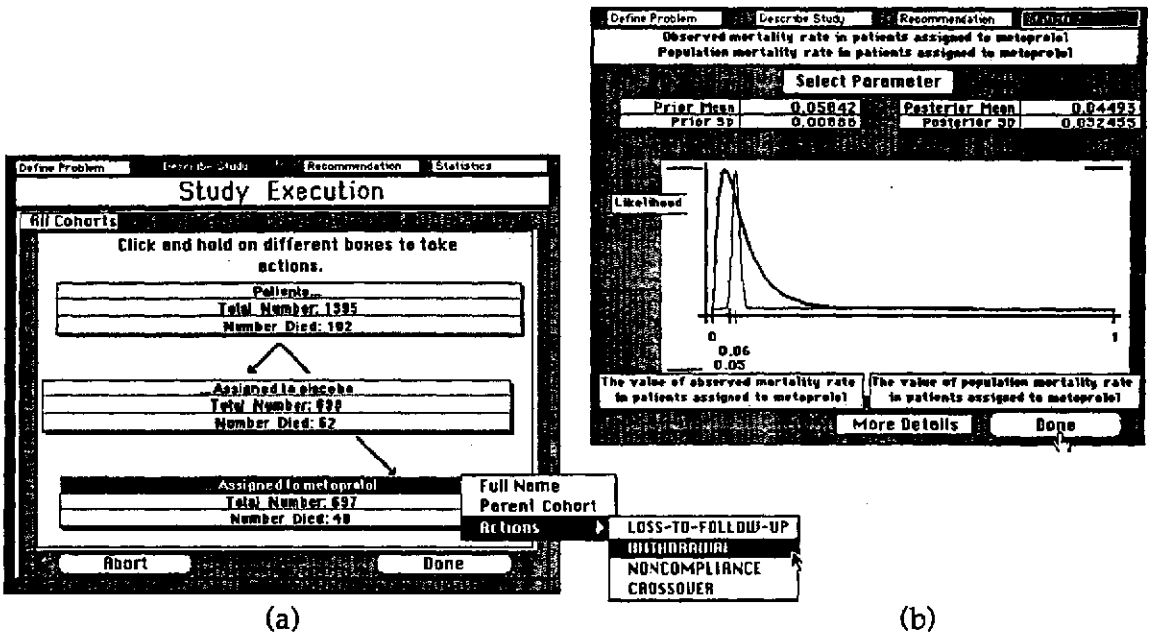


Figure 1.15: Examination of a modified model in THOMAS. In this analysis, the user is examining the effect of incorporating the methodological issue of patient withdrawals. (a) Specification of the model. The choice of modifications may be different for each cohort and is constrained by THOMAS. Not shown in this figure is the growth of the patient-flow diagram to account for patients who were assigned to metoprolol but who did not receive the drug, and for patients who were assigned to placebo but who did not continue the study. There were 131 patients in each of these withdrawal cohorts. (b) Examination of the statistical results. The two parameters of interest are the observed metoprolol mortality rate (thin line), which takes at face value the observed deaths, and the population metoprolol mortality rate (thick line), which removes bias in the observed mortality rate, taking into account the fact that patients who were included in the observed rate did not, indeed, receive metoprolol. Note that the user should believe the debiased mortality rate to be lower than the observed rate, but that the user should be more uncertain about the debiased rate than about the observed rate.

Chapter 2

Knowledge Acquisition for the Literature Problem

In building any decision-support system, the system builder must have an understanding, or a *model,* of the domain of interest. There are at least two strategies for building such a model (Musen, 1989). Using the *extractive* strategy (Breuker et al., 1987, p. 13), the knowledge engineer assembles a model that copies as accurately as possible the concepts, methods, and strategies used by domain experts. Rennels (1987), for instance, built the ROUNDSMAN system applying this strategy.

Using the *constructive* strategy (Anjewierden, 1987), system designers divide their task into the construction of three models (see Figure 2.1). The *conceptual model* encodes the designer's comprehension of domain concepts. The *knowledge-level model* (Newell, 1981) abstracts desired domain goals. The *design model* contains specifications for a working system.

In Section 2.1, I shall summarize the sources of knowledge I have used in assembling the different models needed to develop the framework and to build THOMAS. In Section 2.2, I shall develop the knowledge-level specifications to be used for creating

the conceptual and design models for the literature problem. The use of such a cascade of models is recommended by the Knowledge-Analysis and Design-Structuring (KADS) methodology (Anjewierden, 1987).

The knowledge-level specifications, summarized in Section 2.3, are the major contribution of this chapter. The central insight is that the guidelines offered by methodologists for solving the literature problem properly belong to the knowledge-level model (see especially Section 2.2.3.2). This insight gives the knowledge engineer an extra degree of freedom in building the design model.

Although Figure 2.1 suggests that the knowledge engineer proceed along the sequence of conceptual to knowledge-level to design models, in the biostatistical domain, different approaches within the domain lead to different conceptual models for the same problem. Therefore, I shall present the knowledge level in this chapter, and the conceptual and design models together for each approach in the subsequent two chapters, where I shall describe the classical and the Bayesian models, examining them in terms of these specifications.

2.1 Sources of Domain Knowledge

I have used three sources of knowledge for this dissertation: knowledge acquisition from a domain expert, reading in the biomedical literature, and personal experience. Although this chapter and the two following it present a logical progression, achieving that linear sequence required several cycles of testing and refinement.

Much of the knowledge acquisition for this dissertation grew out of research done for the REFEREE project (Lehmann, 1988). The purpose of that project was to build an expert system that would help a reader to evaluate the credibility of a report of a randomized clinical trial. Knowledge acquisition in that project comprised observations of and interviews with a biostatistician, Byron Wm. Brown, Jr., by

Diana Forsythe (an anthropologist) and myself as he read reports of randomized clinical trials. We spent 45 hours on this process over a period of 1 year. Our analysis was concurrently reviewed by other members of the project: Bruce G. Buchanan, Dan E. Feldman, and R. Martin Chavez. The detailed results of the knowledge acquisition sessions are not used explicitly in THOMAS, but the interviews helped me to develop the specifications for the program.

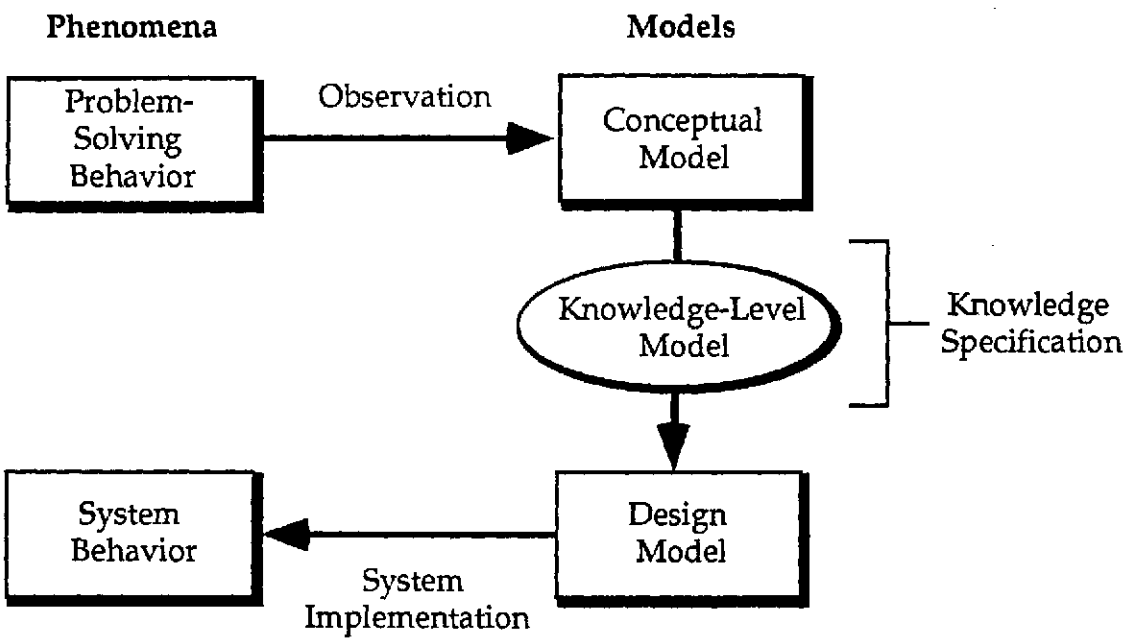


Figure 2.1: Knowledge-analysis and design-structuring (KADS) methodology. Domain problem-solving behavior is transformed into system behavior via three tasks (observation, knowledge specification, and system implementation). Knowledge engineering requires the construction of three models: the conceptual model, the knowledge-level model, and the design model. These models are explained in the text. (*Source:* Adapted from Akkermanns, G.S.H. and Wielinga, B., On problems with the knowledge-level perspective, in *Proceedings of the fifth knowledge acquisition for knowledge-based systems workshop,* Gaines, B.R and Boose, J.H. eds., University of Calgary, 1990, pp. 30/1–30/20.)

The subdomain of *statistical inference* (see Section 3.3.2) comprises the tools biostatisticians employ in offering suggestions for clinical decision making based on research data, and supplies the instruments academic physicians use in constructing decision criteria. I accordingly drew on a large body of literature in this domain, including contributions by statisticians (Peto et al., 1976; Brown and Hollander, 1977; Armitage, 1983; Meinert and Tonascia, 1986), epidemiologists (Feinstein, 1985), clinicians (Sackett, 1979; Sackett et al., 1991; Haynes et al., 1986), and meta-analysts (L'Abbé et al., 1987; Sacks et al., 1987; Eddy et al., 1990). These sources present quantitative and qualitative methods for interpreting the scientific biomedical literature. The variety of approaches available provides the knowledge engineer with the challenge of integrating them.

Finally, my background in epidemiology, statistics, and clinical medicine, including discussions with colleagues, has allowed me to serve as my own "expert," especially for the purposes of considering what notions physicians find difficult to comprehend and what forms of information physicians find manageable.

2.2 Knowledge-Level Analysis

Clancey (1989) presents a manifesto for knowledge acquisition for "second generation expert systems" (p. 285). His research programme[1] includes the following notions regarding knowledge acquisition (adapted from pages 288–289 Clancey, 1989)):

1. Expert systems are situated systems.

2. Data gathering for problem solving represents a social interaction rather than a problem solver's internal process.

[1]Following philosophers of science (e.g., Kuhn (1962) and Radnitzky (1973)), I shall denote a long-term research agenda by the British spelling. This orthographic convention is especially needed in a dissertation where computer software is a major topic of discussion.

3. Knowledge-level descriptions abstract sequences of (expert) behaviors.

4. Domain models are not the expert's "mental model."

I shall explain and use these dicta as an outline for a knowledge-level description of the domain of the literature problem. Our consideration of each point will lead to desiderata for a potential design model.

2.2.1 Situated System

The act of solving the literature problem occurs in a context comprising basic scientists, clinical researchers, statisticians, funding agencies, editors, librarians, lawyers, judges, juries, and clinicians. There are three components to this context: (1) research, (2) publication, and (3) interpretation.

The *research component* depends on the biomedical scientific community maintaining a research programme for scientific research, the agenda (Radnitzky, 1973) for which is made explicit by funding agencies and is kept implicit in the theories, methods, and aims of the scientific community (Laudan, 1984). *Theories* are the concepts (such as the relative merits of metoprolol and placebo in treating patients who have had heart attacks) that scientific studies are attempting to establish. *Methods* are the agreed-on techniques (such as randomized clinical trials) scientists employ in arriving at conclusions about theories. *Aims* are the criteria upon which different methods are judged.

Objectivity is the primary aim of the biomedical scientific community: Individual studies and whole methodologies are judged on this basis. A defining aspect of objectivity is that *disagreement between two scientists over the implications of study results for a particular theory must be accessible to external review and, therefore, must be expressed in explicit terms.* Methods that employ numerical reasoning appear objective, as do methods that apply coherent and consistent procedures for evaluating

the resulting numbers. Thus, biostatistics provides an attractive framework for, and toolbox of, methods in biomedicine.

Reproducibility is an important component of objectivity; if a study's results cannot be reproduced on the basis of information provided by its investigators, it has low credibility (Lehmann, 1988). The information provided lays out the study's audit trail. Thus, the *auditability* of a study's design and execution plays an important part in assessment of a study's reproducibility. The use of *formal* models is one way of ensuring auditability.

The *report component* of the literature problem's context comprises the publication of results in the scientific literature, which depends on peer reviewers' judgment of acceptability, and on editors' assessment of newsworthiness (Goffman, 1981). Opinion leaders in the clinical community publish secondary articles, reviewing the primary research (Williamson et al., 1989). These evaluations often contain the commentators' opinions regarding the applicability of the researchers' conclusions to clinical practice.

The *interpretation component* involves the individual clinician's reading of the article. Her decision to change her actions on the basis of the article depends on her personal reading of the article, on her regard for the opinion leaders, on the opinions of her day-to-day colleagues, and on her assessment of her legal risk in taking the action suggested by the article (Williamson et al., 1989).

An expert system that aids the clinician reading a research paper sheds light on only one aspect of this multifaceted context. While the construction of a system capable of hosting all the agents described above remains a futuristic fantasy, we need now a design model that supports the interactions among those agents. We need systems that allow for the intersubjective differences among readers such that the sources of, and reasons for, the variation are apparent.

Thus, we need a system that is grounded on the community-shared aim of objectivity and auditability, but that allows for *intersubjective differences* among readers.

2.2.2 Social Aspects of Data Gathering

There are two social features we observed in the course of our knowledge acquisition for the REFEREE project. They are the consequence of the expert's own position in the scientific community, and the importance of the social context of the problem at hand.

A statistician functions in a community of statisticians, and is, of necessity, aware of who is trustworthy and who is not. Thus, one of the first queries our expert made in reviewing every paper concerned the identity of the study's investigators and statisticians, or the level of statistical expertise available to the study investigators. Rather than being the product of simple parochial interest, this concern provided the basis for our expert's evaluation of information not available in the written report. If he considered the study statistician to be trustworthy, he would give the study the "benefit of the doubt"; if not, he would assume that any missing information meant that the corresponding methodological concern was implemented by the study investigators in the least credible way. Such knowledge of the community is difficult to include in an expert system; it is the most private, idiosyncratic, and mutable of information. I decided not to attempt to represent it explicitly in THOMAS.

The context of the problem at hand is a social issue in that different questions are raised by different user communities. Clinicians will be most concerned with questions of clinical, practical effectiveness, whereas clinical researchers will be most concerned with biological, ideal efficacy. These different concerns lead to different strategies of analysis. The ability to deal with both strategies—*effectiveness* and *efficacy* (Feinstein, 1985)—shall be part of the specification for the design model.

2.2.3 Sequence of Behaviors

Methodologists (Sackett, 1981; Feinstein, 1985; Meinert and Tonascia, 1986) provide various prescriptions for clinicians' reading of the scientific literature. I shall show, in this section, how these prescriptions serve as ready-made knowledge-level analyses, but do not supply the details necessary to implement those analyses.

These authorities view the clinician as solving the literature problem in the following steps. (1) The physician articulates a particular clinical problem related to a single patient or to a class of patients. (2) She then seeks and finds an article relevant to this problem (Scura and Davidoff, 1981). (3) In reading the paper, the physician keeps in mind concerns about the credibility of the report, the validity of the study, and the applicability of the authors' conclusions to the problem at hand (Sackett, 1981; Mosteller, 1981). (4) The physician offers the therapy suggested by the conclusions of the study, if the conclusions support that action, and if the conclusions meet a variety of criteria (Rennels, 1987). I shall first discuss the limitations of this idealization, and then shall concentrate on the currently available prescriptions for step 3.

2.2.3.1 Limitations of the Idealization

The ideal reader presumed by the methodological literature differs from a real person in important ways. (1) Physicians often do not articulate their clinical problems in as structured a way as is required by formal systems, such as online literature-retrieval programs (Walker et al., 1989). (2) Finding a relevant article is made easier by literature-retrieval programs, but they are not 100 percent sensitive in finding *all* relevant articles. (3) Clinicians often use methodological and numerical techniques incorrectly (see Section 1.4.1). (4) An individual's conclusions, after she has read a paper, are not the sole determinants of her subsequent behavior (Williamson et al., 1989).

In this dissertation, I shall make the following assumptions. (1) Because problem formulation is currently too difficult to automate, I shall presume that the reader can articulate the clinical problem formally. The machine will help the user to refine the clinical question, if necessary. (2) Because the retrieval and understanding of texts recovered by online literature-retrieval programs are difficult problems in their own right, I shall assume that the physician-user of the system selects the article on her own and informs the machine of its contents. (3) Because statistical difficulty is the major reason why naïve users employ statistical software, I shall expect the system to ease the clinician's difficulty with statistics and methodology. (4) Because of the narrowed focus we chose in Section 2.2.1, I shall assume that the physician makes her decision in isolation from the biomedical community. I shall, however, expect her to take the opinions of statisticians, investigators, and other physicians into account; for instance, such opinions will influence her answer to the systems's questions regarding domain and methodological knowledge.

2.2.3.2 Prescriptions as Knowledge-Level Descriptions

Knowledge-level descriptions focus on desired domain goals and avoid specifying methods for achieving those goals (Newell, 1981). As an example of methodological prescription for reading a clinical research article, Chalmers and colleagues (1981) provide the following, among several pages of similar injunctions:

> (1) TESTING PROCEDURES. The next group of determinants of a good study has to do with certain measurements that the investigators should undertake before or during the study.... (2) The importance of patient compliance in clinical trials has been emphasized repeatedly.... (3) Some objective methods of verifying that patients are conforming to the protocol must be described. For example, in a drug trial pill counts would be acceptable. Subjective assessments of compliance are often used.... (4) In some trials, the assessment of

> compliance is considered not applicable.... (Chalmers et al., 1981, pp. 36–37, sentence numbers added)

Sentence 1 establishes determination of a study's quality as the *goal* for the entire process, and defines *testing procedures* to be a major subtask to achieve this goal. The methodologists strive for normativity, since they tell us that they are offering testing procedures that investigators *should* undertake. Sentence 2 presents the determination of the level of *noncompliance* as a further subgoal. Sentence 3 identifies methods (pill counts and subjective assessments) for satisfying this subgoal. Finally, in sentence 4, the authors allude to instances where the subgoal is not relevant.

Regarding *procedures* for satisfying the goals, we note that, in sentence 3, the authors do *not* describe how investigators should actually use pill counts to measure noncompliance, they do *not* describe how to make the subjective assessments, and they do *not* give details as to how the determination of noncompliance would affect the overall determination of the "good study." These considerations are left as implementational details.

Sackett and colleagues (1981) are interested in helping practicing clinicians to negotiate the literature problem. They follow a strategy similar to that of Chalmers and colleagues (1981), describing further desiderata for determining a study's quality. They seem to provide, however, an implementable method for achieving the desiderata, instructing the reader simply to ignore any article that fails to meet all the criteria. They clarify, however, that this Procrustean algorithm should be applied when the physician is reading simply to "keep up" with the literature. However, "when reading up on a specific patient," to solve the literature problem, "before accepting the conclusions" of nonexperimental studies, for instance, they offer a different procedure, concluding that this procedure "is obviously a judgment call and should be tempered" (Sackett, 1981, p. 150). *How* to make the judgment call and *to what degree* that judgment should be tempered are again left as implementation

details.

Finally, in the REFEREE project, we found that our expert brought to our attention many of the same concepts detailed by these different authors. We also found that it was difficult for him to define those concepts crisply: We spent about 10 hours working on the definition of *credibility*, alone. These difficulties can be explained on the basis of a difference between the knowledge-level description of the solution to the literature problem, which is shared by the statistical community, and its procedural-level implementation, which is rarely specified by statisticians, and hence is subject to interpretation and controversy.

In conclusion, we should use as many of the prescriptions as possible in guiding the construction of the design model. I shall give the conceptual details underlying these prescriptions when I discuss the domain concepts involved in solving the literature problem (see Section 3.2.3).

2.2.4 No Expert Mental Model

Statistics is different from other domains for which expert systems have been built, because its practitioners construct formal models of their actions. In focusing on the instrumental use to which statistics is put, Armitage (1983, p. 1) defines statistics as a discipline concerned with the treatment of numbers obtained from the study of groups. Snedecor and Cochran (1980, p 1) offer the broader description that "statistics deals with techniques for collecting, analyzing, and drawing conclusions from data." Efron (1986) presents a typology of those techniques: data enumeration, data summary, data comparison, statistical inference. Fisher (1959), Neyman and Pearson (1930), and Lehmann (1986) have formalized one of these types, statistical inference, in terms of what actions statisticians should take in response to different computed results.

Despite the understanding statisticians have of their techniques, Hand (1986), in

developing an expert system to help novice statisticians perform multivariate statistical analyses, was surprised to find that the way he went about performing such analyses was different from the way he thought he did them, and also was different from the way he was taught to do them. Clayden,[2] in building a statistical consultative system at the University of Leeds, was surprised to find that statistical consultants perform the function of educators almost as often as they act as technical advisers. Other developers of statistical expert systems (Oldford and Peters, 1986; Pregibon, 1986) are equally concerned with the actual actions taken by practicing statisticians. Still other statisticians (Cox, 1977; Mallows and Walley, 1980) seek formal theories to describe the complex tasks of data (or *datum* (Good, 1980)) analysis.

Using the research literature to solve clinical problems is an activity even less rigorously defined than is data analysis, as we saw in Section 2.2.3.2 in examining methodologists' prescriptions. The various desiderata of Chalmers and colleagues (Chalmers et al., 1981), the flowcharts of Sackett and colleagues (1981, 1991), and the prescriptions of Feinstein (1985) come closest to defining that activity, but they are informal and clearly are abstracted from statistical activity; no one would claim that they describe accurately, and in detail, what happens in the mind of any single reader of the scientific literature.

2.3 Knowledge-Level Summary

From the discussion in Section 2.2, we can assemble, as specifications, desired properties of a system designed to help physicians solve the literature problem.

1. *Objectivity:* The system should depend on objective, reproducible, and auditable methods (Section 2.2.1).

[2]Personal communication.

2. *Intersubjectivity:* The system should allow for differences of opinion among readers (Section 2.2.1).

3. *Normativity:* The system should implement methodologists' knowledge-level prescriptions (Section 2.2.3.2).

4. *Flexibility:* The system should be able to evaluate both the pragmatic effectiveness and the ideal efficacy of tested therapy (Section 2.2.2).

5. *Adaptability:* The system should enable the clinician to express methodological concerns without using statistical language (Section 2.2.3).

6. *Simplicity:* A simplified system should help the physician to interpret a single article that she has selected and read; should exclude explicit knowledge about particular statisticians and investigators (Section 2.2.3.1); should exclude explicit knowledge about particular statisticians and investigators (Section 2.2.2); should assume that the physician user has the ability to express clearly the particular problem at hand (Section 2.2.3.1); and should support the decision making of a single physician, rather than that of an entire community (Section 2.2.3.1).

The first and second specifications seem at odds with each other: One calls for objectivity, whereas the other calls for subjectivity. In the next chapter, we shall examine one proposed resolution of this tension in the traditional implementation of the knowledge-level model—the classical statistical design model.

Chapter 3

Classical Design Model

The implicit solution to the literature problem currently carried out by the biomedical community is founded on classical statistics. In this chapter, I shall examine the classical solution as a design model[1] for solving the literature problem, and I shall examine the adequacy of the solution in terms of the knowledge-level specifications from the previous chapter (see Section 2.3). I shall also examine, in Section 3.6, some previous computer-based solutions to the literature problem.

In Sections 3.1 through 3.4, I shall present the basic concepts—the vocabulary—of the domain. I shall use the metoprolol example to clarify the concepts. Continuing to use the Knowledge-analysis and design-structuring (KADS) (Anjewierden, 1987) approach to knowledge acquisition, I shall organize the presentation by their fourfold division of domain knowledge into (1) task concepts, (2) domain concepts, (3) inference concepts, and (4) strategy concepts. In explicating these concepts, we will also develop the design model, from the bottom up.

Section 3.5 provides a knowledge-level critique of the classical-statistical approach, using the knowledge-level conditions of Section 2.3 as test criteria. This critique

[1]See the introduction to Chapter 2 for the definition of this and other knowledge-engineering terms.

motivates the exposition in Chapter 4—the Bayesian design model.

Readers already familiar with classical statistics may wish to read only the last three sections. However, the first sections present statistical concepts in a novel way, using influence diagrams[2] to describe the statistical concepts; the reader may wish to peruse the figures in those sections. Section 3.6 summarizes previous computer-based work.

3.1 Task Concepts

Therapy selection, which we shall define as a choice between two drugs, is one of the physician's primary clinical tasks. *Statistical inference* (see Section 3.3.2) is the major subtask that statisticians recommend for satisfying the primary task. Statistical inference is, to use an information-flow metaphor, the primary conduit between the reporting statistician and the deciding physician. The question to be answered by the design model is, What information should flow through that channel? A summary of the classical answer is given on page 71.

3.2 Domain Concepts

I shall discuss four classes of domain concepts in statistics. *Probabilistic concepts* concern the representation and management of uncertainty, *statistical concepts* concern the relationships among observations, *methodological concepts* concern the structure of the observation process, and *inference concepts* the way investigators should learn from observations. The concepts constitute the primary vocabulary needed to understand any biostatistical system.

[2]See Section B in the Appendix for a short tutorial on this knowledge representation.

3.2.1 Probabilistic Concepts

Probability embodies notions of observations, of random variables (which abstract observations), of likelihoods (which communicate the chances of observations occurring), and of parameters (which summarize likelihoods).

The concept of an **observation** is central to any model of statistics. Statisticians make two basic, implicit assumptions: observations of the world are possible, and observations of past events are related to future observations. Classical statisticians take as their fundamental axiom that probabilities are ratios of observations observed to observations that could be observed; probabilities are frequencies. This is the *frequentist axiom* (see von Mises and Geiringer (1964) and Kolmogorov (1965)). In this dissertation, observations, such as the death of a patient from a myocardial infarction (MI), shall be denoted by lowercase italic letters (e.g., x).

A **random variable** can take on a value that refers to any one of a number of outcomes; an observation is an outcome that has occurred. The *random* aspect of such a variable is that different outcomes have different chances of occurring. The *variable* aspect is that different numbers are assigned to different outcomes (or to sets of outcomes). Random variables shall be denoted by uppercase italic letters (e.g., X); vectors of random variables shall be denoted by boldface uppercase Roman letters (e.g., $\mathbf{X}$). A patient's lifespan (denoted L) is an example of a random variable.

The **probability density function (pdf)** or **likelihood function** (which is directly proportional to a pdf) assembles the chances of occurrence for different outcomes of a random variable. The hallmarks of a pdf are that the sum (or integral) of all likelihoods is 1 and that no likelihood is negative. The pdfs shall be denoted as $P(X)$ (for the random variable X), and the likelihood functions shall be denoted as $\ell(X)$; $\ell(X) = k\,P(X)$, for some constant k.

Likelihood functions can take on many different shapes. There are some shapes, however, that are canonical in that, if you were to know the values of a small number of

parameters of the likelihood function, you could derive the entire likelihood function. Such likelihood functions are called *parametric models.* Distributions we shall be using are these: the Bernoulli distribution, which we shall denote $\mathcal{B}(p)$, with the single parameter the probability of success; the binomial distribution, which we shall denote $\mathcal{BI}(p,n)$, with parameters the probability of success and the sample size; the normal distribution, which we shall denote $\mathcal{N}(\mu,\sigma^2)$, with parameters mean and variance; the the beta distribution, which we shall denote $\mathcal{BE}(\alpha,\beta)$, with the parameters the effective number of successes and the effective number of failures[3]; and the exponential model, which we shall denote $\mathcal{E}(\lambda)$, with the single parameter the instantaneous failure rate.

From a computer-science point of view, parametric models reduce computational complexity. For instance, consider the random variable that indicates the lifespan of a physician's next patient with an acute MI. That variable is continuous, with an uncountably infinite number of values, and hence, an uncountably infinite number of likelihoods. If the physician assumes that the variable is exponentially distributed, then only two facts are needed to generate the uncountably infinite number of likelihoods: the identity of the distribution (exponential), the value of the instantaneous failure (mortality) rate. This huge reduction in computational complexity is a compelling reason for using parametric models in statistical work.

A **parameter,** *therefore, is an unobserved entity, that, when specified, can be used to determine the likelihood function of a random variable.* Practitioners and students of classical statistics view parameters as real, physical quantities, like the speed of light. The purpose of research, from this point of view, is to determine the value of those constant quantities. I shall denote individual parameters lowercase by Greek letters (except for the Bernoulli success probability, which shall be denoted by p). The symbol θ shall denote the general, nonspecific parameter.

[3]These labels are my own; there are no standard names for these two parameters.

In summary, parametric models represent the uncertainty—probability—investigators and readers have in the occurrence of outcomes of interest. I demonstrate the concepts presented in this section in Figure 3.1, using an influence-diagram representation. The graphical conventions for this and future diagrams are listed in the Appendix (Section A.3). In brief, uncertain quantities are represented by ovals and deterministic quantities (constant or functional) are represented by double ovals. Arrows (arcs) represent dependency: An arrow from A to B means that knowledge (belief) in B depends on knowledge (belief) in A. In the case where B is deterministic, such an arrow means that the value of B is certain, given the value of A. I shall introduce further graphical conventions as they arise.

3.2.2 Statistical Concepts

One goal of statisticians is to learn from scientific studies. In such studies, investigators are interested in the relationships among observations; statisticians express such relationships in terms of the associations of the observations to the unobservable parameters, and, therefore, work with probabilistic concepts. The heart of these relationships is the notion of a **population** of individuals with similar features, such as all patients with MIs; a new patient with an MI is thought of as a member of this ideal population. The amalgamation of likelihoods for individuals exhibiting the outcome of interest leads to a likelihood function for the population. If that function can be parameterized, its parameters are called *population parameters*. In the course of research, the experimenters collect a **sample** of individuals from the population, such as patients with an MI admitted to a particular hospital in South Sweden. The goal of statistical analysis is to relate observations made on sample individuals to population parameters. This relationship is encoded in a **statistical model** (Ellman, 1986). The statistical model is the central knowledge representation in statistics.

We can examine the metoprolol example from this perspective. Consider the population of all patients with an MI who survive the acute attack and are admitted to the hospital; the outcome of interest is survival by 3 months after admission. The random, observable variable, X, assigns 1 to the outcome death and 0 to the outcome survival. The likelihood function for X is parameterized by a single, unobserved number, the probability (p) of the patient dying, yielding a Bernoulli, parametric, probabilistic model. (The probability of survival is $1 - p$; the two probabilities constitute the entire pdf.) Consider the experiment of observing 698 such patients over 3 months. The observed patients constitute the sample. If we assume that the outcome of each patient is related to the parameter, p, the same way, and is not influenced by the outcome of any other patient, then we can use the binomial statistical model to represent the relationship between the sample outcomes and the population parameter. The binomial model relates p to the number of deaths, D, in the 698 patients: $P(D \mid 698) = {}_{698}C_D\, p^D(1-p)^{698-D}$, where ${}_{698}C_D$ is the binomial coefficient.[4] The likelihood function for the number of deaths is $\ell(D \mid 698) = p^D(1-p)^{698-D}$. Note that, in this discussion, two types of knowledge are involved: One is the statistical technical knowledge regarding classes of probabilistic models. The second is the domain knowledge that patient deaths do not affect one another.

If the observations are independent of one another and each has the same relationship to the population parameter, then the observations are said to be **independent and identically distributed** *(iid)*. I shall say that, in such a situation, the parameter *governs* the sample observations. Figure 3.2 depicts this central notion of independence, and introduces the idea of a **statistic:** a function of observed data. One powerful result of the concept of iid is that simple functions of the different outcomes can be shown to have simple relationships to the original parameter(s) of

[4] ${}_{698}C_D = \frac{698!}{D!(698-D)!}$

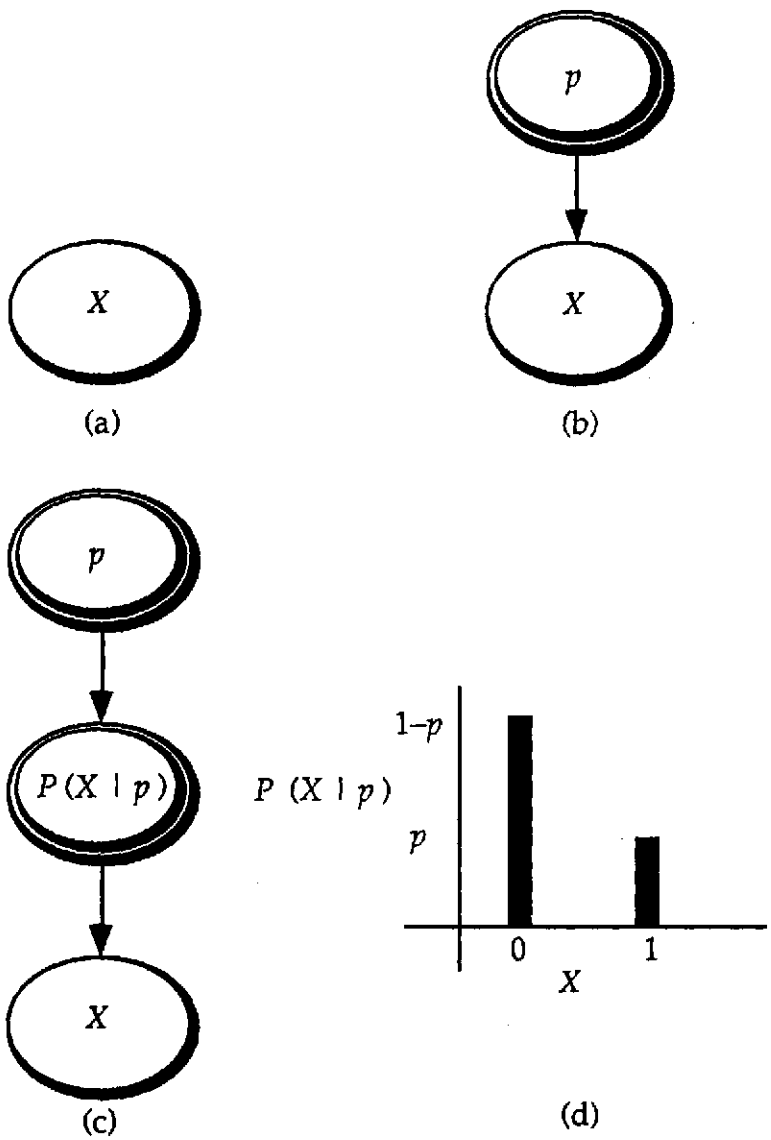


Figure 3.1: Classical random variables. (a) An oval containing a label represents the random variable, X. The variable might represent the death of a patient. (b) The uncertainty of X is dependent on one parameter, p, that is constant (and therefore represented by a double oval, indicating a deterministic node). The value of p may be unknown. The parameter p might have the semantics of mortality rate. (c) The dependence of X on its parameter is such that, if the parameter were known, then the uncertainty in X, given the values of that parameter, would be fixed; X remains uncertain. This functional dependency is indicated by the double oval. (d) This graph shows the actual values of the deterministic dependence of X's likelihood on the parameter, where $X = 1$ denotes that the proposition signified by X is true.

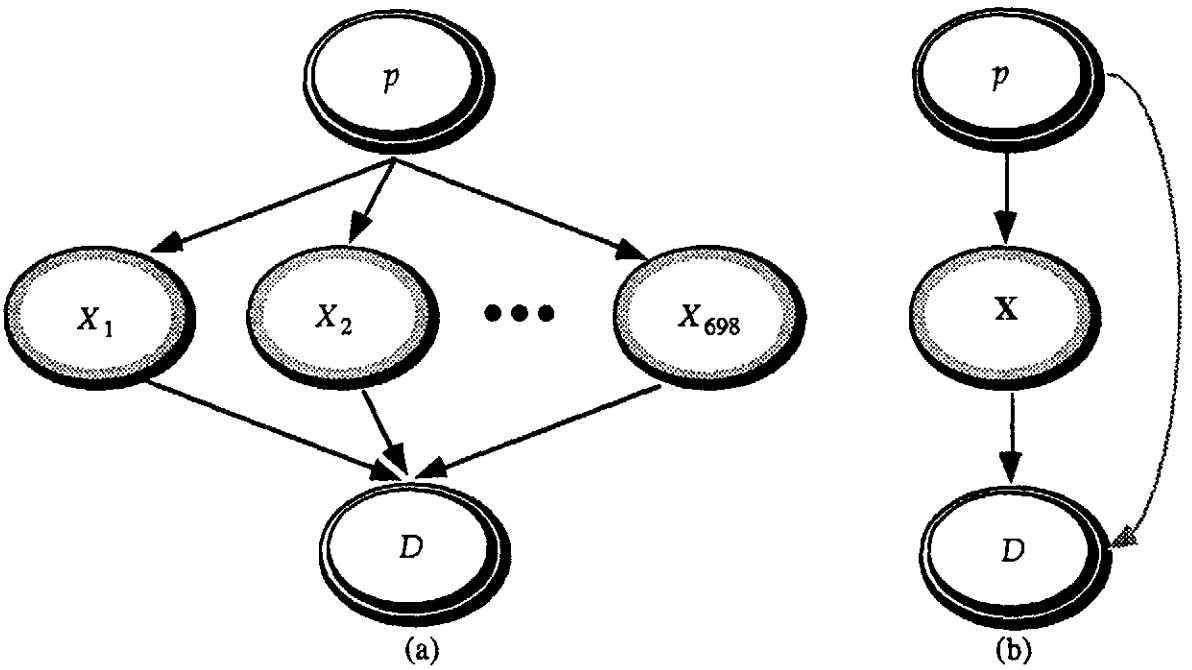


Figure 3.2: Classical model for independent and identically distributed (iid) random variables. (a) 698 patient deaths (X_i), with parameter p, are modeled as conditionally independent of one another, given the parameter, and each having the same relationship to that parameter (i.e., the Bernoulli distribution). The X_i have been observed, and are therefore represented by a shaded oval. The *binomial* distribution implied by this diagram ($P(D = \sum_{i=1}^{698} X_i \mid p)$, see the text) actually has two parameters: p, an unknown, fixed constant, and 698, the number of patients, in this case, a known, fixed constant. D is a *statistical function* (or, simply, *statistic*) of the observations. (b) An alternative graphical convention for the model in (a) is shown. Here, $\mathbf{X}$ is a vector of the values $(X_1, X_2, \ldots, X_{698})$. The speckled (derived-probability) arc indicates that the distribution of D given p and 698 is known; the identity of this distribution depends on the solid-arc relationships in the diagram.

interest. Thus, $D = \sum_{i=1}^{698} X_i$ is the total number of deaths in the sample and is a statistic (function) of the observed data. A well-known result of probability leads to the conclusion that $D \sim \mathcal{BI}(698, p)$.

3.2.3 Methodological Concepts

Baseline statistical models, such as those described in the previous section, embody the assumption that observations made on sample individuals give information directly for the population parameter. Investigators analyzing—and readers examining—the results of a study need this premise to erect a bridge between prior observations and future decisions. Methodological concerns modify this assumption. There are two classes of concerns that we shall consider (Cook and Campbell, 1979). The first comprises those concerns that consider the possibility that the parameters governing each individual may differ between subjects; these are concerns regarding *internal validity*. The second comprises those concerns that take into account the possibility that the parameters governing the sample observations are not the same as those governing the population; these are concerns of *external validity*.

Clearly, no two patients are ever identical. *Internal validity* concerns the degree to which the assumption that they are is *violated* in the study at hand, and the degree to which the violation affects the ultimate conclusion. Important questions include: Are the observations indeed independent of one another? Do the observations represent accurately the true values of the subjects' outcomes (Miettinen and Cook, 1981)? In RCTs, investigators answer these questions by examining the treatment and endpoint-assessment phases of a study (see Figure 3.3).

Many methodologists, including our expert in the REFEREE project, regard the *blindfolding* of investigators[5] as most important (Meinert and Tonascia, 1986; Sacks

[5]Blindfolding—also called *masking*—is the practice whereby patients or care providers or investigators remain ignorant of the identity of the treatment received by a subject in the course of a

et al., 1987), yet there are no mathematical models (to my knowledge) that instruct the reader how to debias the conclusions, given apparent blindfolding violations in a given study. The amount of blindfolding possible depends on the nature of the treatments under investigation: Medical treatment can often be masked, whereas surgical procedures are difficult to disguise.

For readers to take into account biases occurring during the treatment and endpoint phases of a study, authors have described the qualitative problems that might affect any inference (Sackett, 1979; Feinstein, 1985), and have specified quantitative models for debiasing the conclusions (Greenland, 1984; Greenland and Robins, 1986; Greenland, 1987). An important set of potential biases are those due to **protocol departures,** in which patients do not follow the directives of the study (e.g., because they move out of the area and lose contact with the study investigators). In the case of protocol departures, the measurements made do not necessarily represent the measurements intended: If patients move away, then the investigators lose the contributions of these patients' outcomes to the conclusions about the study parameters. Eddy and colleagues (1991) provide the relatively straightforward models that I use in THOMAS for protocol departures (see Section 5.6); Lakatos (1986) presents a variation on those models.

Even if patients within a study are identical for the purposes of the conclusions, the entire sample of patients may differ from the population of patients of interest. *External validity* concerns the generalizability of a study: Are subjects in the study the same as subjects governed by the population parameter (Antman et al., 1985)? Does the aggregate of sample subjects represent accurately subjects in the population (Chalmers et al., 1983)? In RCTs, these questions translate into the following queries: Were subjects recruited from the population of relevant patients in a representative manner? Were they assigned evenly to each therapy? (see Figure 3.3).

---

study.

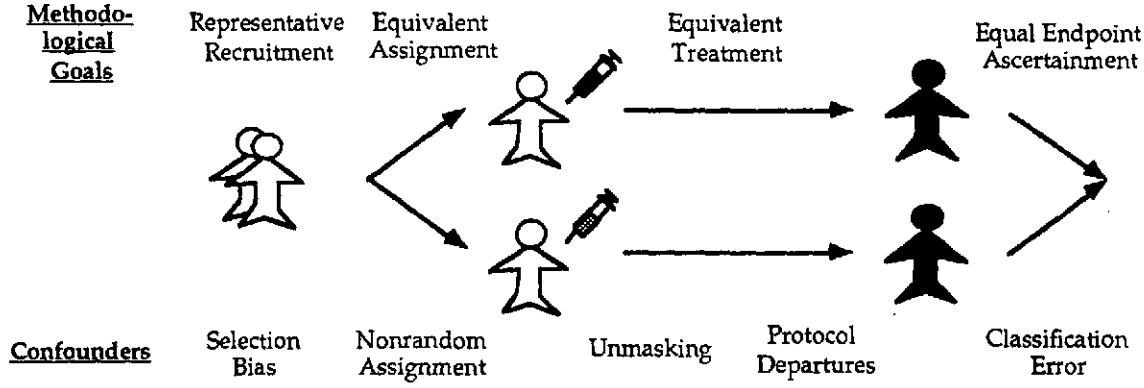


Figure 3.3: Methodological concerns in randomized clinical trials (RCTs). The central graphic shows the canonical time course of an RCT (compare to Figure 1.1). The upper line shows the methodological goals of each phase in the study. The lower line shows problems that can confound each goal. External-validity goals involve the recruitment and assignment phases of the study; internal-validity objectives involve patient-treatment and endpoint-assessment phases.

Representative patient recruitment is crucial for generalizing the results of a study (Antman et al., 1985); *selection bias* confounds this process. The general screen for detecting this bias is to compare, on the basis of important baseline characteristics, patients consenting entry into a trial with patients refusing entry. Although there are mathematical, implementable classical-statistical techniques for dealing with this confounder (Kleinbaum et al., 1981), it is difficult for investigators to assemble the data necessary to apply the methods, because it entails getting information on patients not in the study.

The primary statistical strategy for achieving equivalent assignment is *random sampling*. Random sampling and its descendent, the randomized clinical trial, have become the gold standards for judging experimental designs (Gelband, 1983; Ellenberg, 1984), although the suitability of this strategy for biomedical research has been debated from the days of Student (1937) and Fisher, early in this century, through the present (Howson and Urbach, 1989). The major attraction of randomization is

that, if a randomized sample is obtained, then, for increasingly large sample sizes, the values of simple functions of the sample observations can be shown to approach the values of the population parameters (see Figure 3.4); the functions used are then said to be statistically *unbiased.* For instance, the value of the observed mortality rate in the group of patients treated with metoprolol can be shown, under mild conditions, to approach the population mortality rate for patients treated with metoprolol, in general. Furthermore, this approach minimizes the confounding that results from uneven distribution of influential baseline characteristics (Efron, 1971).

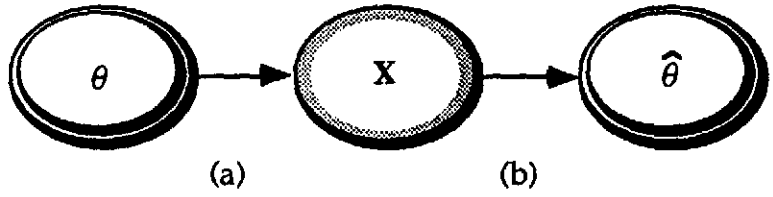


Figure 3.4: Classical-statistical parameter estimation. (a) This arc denotes the iid assumption, allowed by the process of random sampling. (b) This arc denotes the function that determines the estimate, $\hat{\theta}$, of the unobserved parameter, θ. As the sample size (i.e., as the number of components in the vector $\mathbf{X}$) approaches infinity, the value of $\hat{\theta}$ approaches that of θ.

In summary, statistical models are build out of a web of concepts: observations, probabilities, parameters, independence, identity of distributions, populations, samples, and biases. The interrelations among these concepts determine how we should use the results of a study to take action in future circumstances.

3.3 Inference Concepts

Inference concepts describe how the basic domain concepts should be used to achieve the goal—making drug choices, in our case. I shall discuss four areas of inference concepts in statistical analyses: metadata, classical hypothesis testing, statistical significance, and adjustments.

3.3.1 Metadata

The term **metadata** *refers to data about data* (Chytil, 1986). There is much information an analyst can glean about a study by just knowing the *type* of data involved. A simple example is that, for instance, only a continuous quantity may be exponentially distributed. *Metadata* is an inference concept, because it relates the type of the data the analyst is analyzing to the types of actions he may take with those data. The inference is made on the basis of the metadata *before* the data themselves are examined. A number of expert systems for statistics embody such rules (Chytil, 1986; Jasinski, 1986). *Domain knowledge* is important as well, suggesting, for instance, the numerical values that a datum can take. Thus, metadata are useful for making basic inferences about the data that are about to examined.

Expert-system researchers are paying attention to the concept of metadata, because it is the metadata that strongly affect the choice of statistical analysis. This is especially true in classical statistics, where the analyst should not examine the observed data in choosing the appropriate statistical analysis for the study (Wittkowski, 1986). Metadata are important in making these higher-level decisions. For instance, an analyst might know the type of distribution appropriate for cancer deaths or for cardiac deaths, without examining the actuarial table of deaths observed. Furthermore, metadata can suggest what adjustment models might be appropriate. For instance, the assessment of death rarely produces a false-positive report; measurement reliability is probably less important in studies measuring this outcome than in studies assessing morbidity, where study results are less certain.

3.3.2 Hypothesis testing

Statistical inference provides the rules for making statistical decisions on the basis of study results, relating observations to unobservables. *Classical hypothesis testing*

is the now-traditional method of statistical inference. A complete description and analysis of this approach is beyond the scope of this dissertation; the reader is referred to classic texts, such as that of Lehmann (1986).[6] A broad outline of the technical details is necessary, however, if we are to appreciate the differences from the approach I shall introduce in the next chapter, where we will show how the Bayesian approach replaces the concepts of this section. Novel in this presentation of hypothesis testing is the use of influence diagrams to clarify what information is used by this approach and where it is applied.

The goal for the analyst taking the hypothesis-testing approach is to rule in or out statements about real-world entities. This goal is implemented in terms of ruling in or out a preferred value of a fixed, unknown parameter, such as the difference in mortality rates of patients treated with metoprolol and those treated with placebo. Figure 3.5 shows how a general hypothesis test is represented in the influence-diagram representation. The state of the world in which the parameter has the preferred value is called the *null hypothesis*; one null hypothesis is that the difference in mortality rates is zero—the preferred value. The decision regarding the truth of the null hypothesis is based on the calculation of a *test statistic,* such as the *z-score* or the *t-test* statistic, whose value is a function of the observed data, and of a particular type of probability. That probability is called the p **value,** *the probability that the experimenters would have observed data resulting in that value of the test statistic, or data more extreme, were the null hypothesis true.* If the p value is under a certain threshold, then the experimenters were *un*likely to have observed the data that led to the calculated value of the test statistic. If the data (or the corresponding test statistic) appear unlikely, we should reject the null hypothesis in favor of the alternative hypothesis. The rule may be written as follows, for the random variable X and observed data, x:

$$\text{if } P\left(t(X) > t(x_0) \mid H_0\right) \leq \alpha, \text{ then reject} H_0$$

[6] No relation to me.

The threshold, α, determines the *type I error*—the threshold for rejecting the null hypothesis when, in fact, it is true. A second threshold, β, determines the *type II error*—the threshold for accepting the null hypothesis when, in fact, an alternative is true.

As an example of this process, let us expand on the metoprolol example, where the *z-test for proportions* (Snedecor and Cochran, 1980), [7] provides the appropriate measure. Figure 3.6 shows this test in influence-diagram representation; it is a specialization of Figures 3.4 and 3.5. The question here is whether the the mortality rate for patients treated with metoprolol and that for patients treated with placebo are the same. The parameter tested is a third parameter—the difference between the two mortality rates, which are proportions. The preferred value for this parameter of interest is zero, because the semantics of a zero difference are that the two mortality rates are the same. If, as a result of the hypothesis test, we reject the null hypothesis, then we are rejecting the notion that the mortality rates are the same.

The details of the z-test are given in the caption to Figure 3.6. This hypothesis test has four components: (1) the choice of the parametric model (in this case, the Bernoulli distribution) for the iid observations in the samples; (2) the choice of an appropriate statistic (the z-score), or function, of the observed data; (3) the knowledge regarding the distribution (the normal distribution) of that statistic given some value of the fixed, unknown parameters; and (4) the inference rule (rejecting the null hypothesis that the means are equal if the observed probability is less than some threshold), given the results of the observed probability. The choice of the parametric model (component 1) is often made on the basis of metadata and of prior domain knowledge. The knowledge regarding the distribution (component 3) depends crucially on the assumption that the investigators obtained observations by randomly sampling from a population governed by the population parameter. The sensitivity

[7] This test is more properly called the *likelihood ratio test for proportions,* where asymptotic properties of the process and nonextreme values of the true proportions allow for the assumption of a normal distribution.

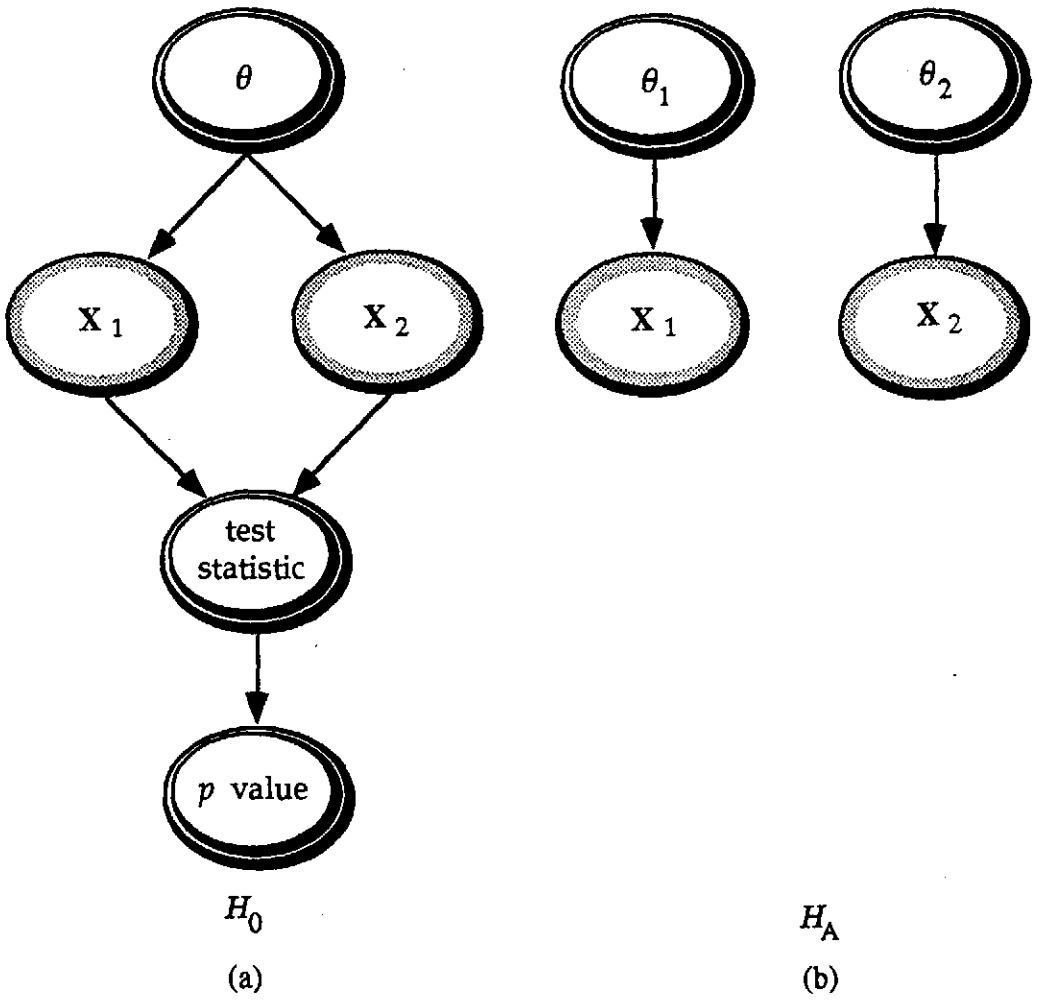


Figure 3.5: Hypothesis testing. Two samples ($\mathbf{X}_1$ and $\mathbf{X}_2$) are observed, each presumed to have probability-distribution functions governed by a general parameter, θ. (a) The null hypothesis, H_0 states that the two population parameters are identical ($H_0 : \theta_1 = \theta_2 \equiv \theta$). The test statistic is computed from the observed data; the p value is computed assuming the null hypothesis to be true. (b) The alternative hypothesis, H_A states that the two population parameters are different ($H_A : \theta_1 \neq \theta_2$).

of this process to this sampling assumption implies that, if it were known that the sampling occurred in some other way, an entirely different choice of sample statistic would have to be made. For this reason, much classical statistical research entails creating the appropriate test (component 2) for different sampling situations. Finally, the thresholds (component 4) and the test must be chosen *before the beginning of the study,* and the data must be examined only once (see Section 3.3.3). These requirements also make the process sensitive to departures from design.

A variation on the hypothesis-testing approach is the *confidence-interval technique,* which biostatisticians and medical investigators use to communicate the degree of uncertainty in a parameter estimate (see Figure 3.4). In calculating confidence intervals, the analyst determines the values of the parameter estimate that would lead to rejection of the null hypothesis at the specified α threshold. The range of values derived is called a *confidence interval.* One way of expressing a given confidence interval's semantics are these: If the experiment were repeated n times after the current experiment, then, even if the true parameter lay in the given interval, 100α percent of those times the experimenter would observe data that would lead to a calculated confidence interval that would *not* overlap the given confidence interval; the experimenter would mistakenly believe that the parameter's value lay elsewhere. The narrower the given confidence interval, the more confident are we about the validity of the parameter estimate. It is false to say that the experimenter believes, with $100(1-\alpha)$ percent confidence, that the value of the true parameter lies within the given confidence interval, because, by the frequentist axiom (see page 53), all probabilities are frequencies, and not measures of belief. In this case, the frequency being counted is the proportion of overlaps of calculated intervals. We shall see, in the Section 4.3.5.2, that a Bayesian version is more straightforward.

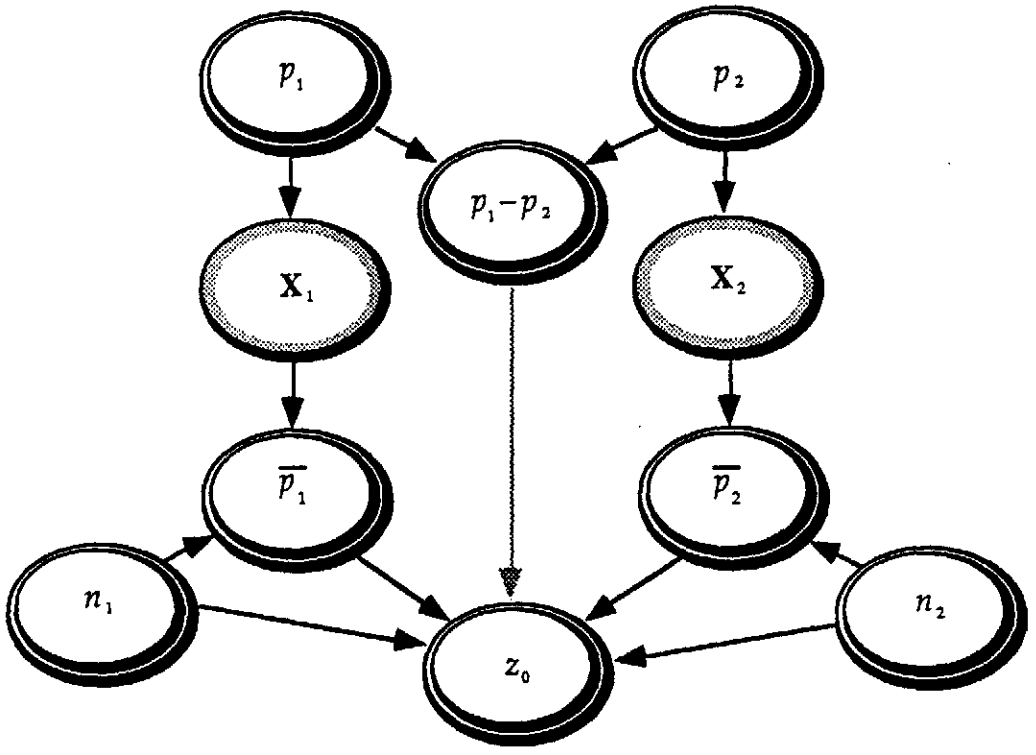


Figure 3.6: The z-test for proportions. The constant population parameters for two Bernoulli-distributed populations are p_1 and p_2. The two samples are $\mathbf{X}_1$, with n_1 iid observations, and $\mathbf{X}_2$, with n_2 iid observations. The z-score statistic is a function of the observed proportions of success, $\overline{p_1}$ and $\overline{p_2}$, (which are parameter estimates (see Figure 3.4) for p_1 and p_2 respectively), and of the (constant) sample sizes, n_1 and n_2. ($z_0 = \frac{\overline{p_1} - \overline{p_2}}{s}$, where $\mathrm{s}^2 = \hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, $\hat{\mathrm{p}} = \frac{n_1\overline{p_1} + n_2\overline{p_2}}{n_1 + n_2}$, and $\hat{q} = 1 - \hat{p}$.) The pdf for the z-score statistic, given that the difference, $p_1 - p_2$, is truly zero, is, asymptotically, the normal distribution. Note that the identity of this distribution depends crucially on all assumptions about the relationships between different entities in the remainder of the diagram, including the values of p_1 and p_2. If $P(Z \geq z_0 \mid p_1 - p_2 = 0)$ is less than some threshold probability, then the statistician accepts the hypothesis that, in fact, the two population parameters are not equal to each other.

3.3.3 Statistical Significance

Inferences require criteria for decision making. The inference concept of *statistical significance* supplies the criterion in classical statistics. When the hypothesis-testing approach is used for answering questions such as, Is the mortality rate due to metoprolol different from the mortality rate placebo?, the statistical implementation of this question is, Are the two mortality rates *statistically significantly* different from each other?

The purpose of the criterion is to set *bounds* on errors over the course of several tests (either between studies or within a single study): If a clinician acts on the basis of the type I- and II-error thresholds, then, on 100α percent of studies, she will make a clinical error if she acts on the basis of rejecting the null hypothesis, and she will err 100β percent of the time if she acts on the basis of accepting the null hypothesis. The test must be selected on the basis of the study design, which is implied by information available at the outset of the study, before any data are known; the bounds protect the reader regardless of what data are observed. Because the errors are expressed as a percent of a total number of actions, the true bounds depend on how often the tests are made. Therefore, the analyst should examine the data (and calculate a test statistic) only once in a study; if not, the thresholds must be modified.

Many authors distinguish statistical significance from *clinical* significance. Statisticians leave to the domain experts the heuristic judgments regarding the clinical meaning of a particular statistically significant difference. For instance, if the confidence interval for the mortality rate due to metoprolol lay between 0.015 and 0.020, and that for the mortality rate due to placebo lay between 0.020 and 0.021——many clinicians would state that, although the difference is statistically significant, because the two intervals do not overlap, it is not clinically meaningful, because the difference in midpoints is only 0.003. This notion of clinical significance is the final limb of the classical solution to the literature problem.

As an example of an heuristic judgment translating statistical into clinical significance, some statisticians advocate explicit adjustment of the p value to take into account prior opinion. Consider the following prescription from an often-referenced article by leading British biostatisticians (Peto et al., 1976), regarding a report's claim of significance made on the basis of p values by investigators of a clinical trial:

> Suppose...that before you saw these results you had no opinion, but on reflection the claim seems reasonable.... Now, a $P = 0.05$ result would not in itself be convincing, although it would make you more receptive to future such claims; a $P = 0.01$ result would be difficult to dismiss; while a $P < 0.001$ result would be extremely convincing.
>
> Suppose, finally, that, had your opinion been sought before reading the published report, you would have thought that there was little prospect of such a treatment being of value. Now, a $P = 0.05$ result would leave you almost as sceptical as before; and although a $P < 0.001$ result would change your mind, you would still retain a secret little doubt. (Peto et al., 1976, p. 595)

We note that these methodologists leave important notions, such as *more receptive* and *convincing,* undefined; they are left to the reader's heuristic judgment.

3.3.4 Adjustments

The fourth class of inference concepts concern the heuristic judgments just mentioned. Expert-system researchers have endeavored to define those heuristics that enable a reader to apply the conclusions of a study to specific contexts. The expert-system that embodies the most comprehensive set of heuristics is Rennels' ROUNDSMAN program (Rennels, 1987) (see Section 3.6).

The discipline of meta-analysis is apposite in this context. The discipline was created for the purpose of extending the classical statistical approach to enable an

analyst to integrate the conclusions of multiple studies. Although we are interested in single studies in this dissertation, meta-analysis is relevant because it, too, deals with issues such as problems with methodology and difficulty in determining which actions to take on the basis of research studies. There are three ways that meta-analysts address these difficulties; all involve weighting the statistical measures from each study, and computing an overall statistical measure from the base measures and the corresponding weights. In one approach (called the *Peto method* by Berlin and colleagues (1989)), the weights are a measure of the uncertainty in the statistic due to within-study sampling. These weights are simple functions of the sample sizes and the results of the component studies. In the second approach, due to DerSimonian and Laird (1986), the weights take into account between-study variation, as well. In the third approach, taken by many meta-analysts (Sacks et al., 1987; L'Abbé et al., 1987), a score is computed based on a scale reflecting the study's methodological integrity. These scores are heuristic, because their values have no normative basis, except that, on the basis of items on the scale, a higher score implies that a study is of higher quality than one with a lower score. Different meta-analysts may use different scales.

In summary, the classical answer to the question raised in Section 3.1—What information should be conveyed by the investigators to the reader?—is that the information should include the identities of the hypothesis tests used, the thresholds assumed, the results of the tests, as well as the qualitative assessment of factors that may have confounded those results and that thus may indicate a need for adjusting the conclusion.

3.4 Strategy Concepts

The fourth class of concepts needed for the design model for solving the literature problem prescribes how the analyst should use the information provided by the investigators, which is based on the task, domain, and inference concepts. The classical statistical strategy is as follows:

1. Select the appropriate statistical model and test for the study, based on the study's original design.
2. Calculate the test statistic and the p value for the study.
3. Adjust the threshold for rejecting the null hypothesis on the basis of heuristics that represent prior methodological and clinical knowledge.
4. Take the action dictated by the adjusted threshold.

The sequence of these steps is depicted in Figure 3.7, in a manner analogous to the Bayesian solution of this dissertation (see Figure 1.6).

3.5 Critique of the Classical Approach

Having just completed an overview of classical statistics, we need to evaluate the suitability of the classical approach as a design model for solving the literature problem. I shall organize this critique in terms of the desiderata for systems to help physicians with the literature problem, as outlined in Section 2.3.

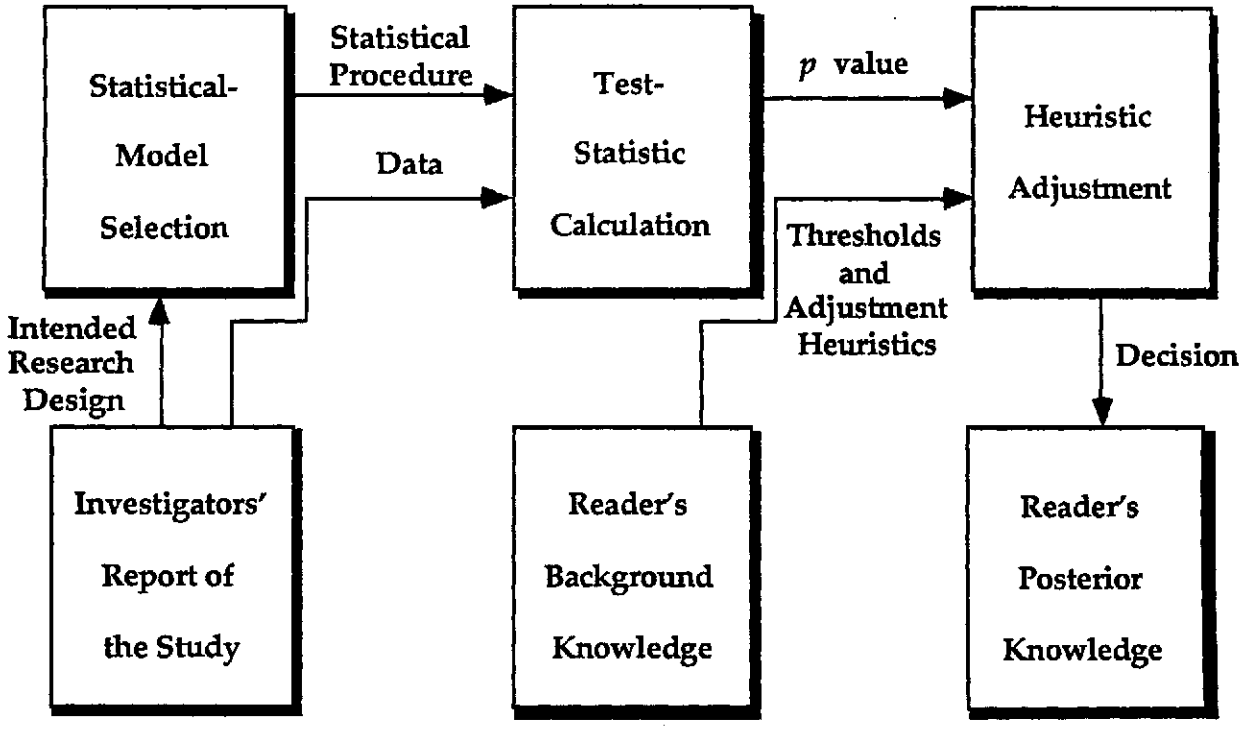


Figure 3.7: Information flow-diagram depicting the classical-statistical strategy for solving the literature problem. Although the sources of inputs are similar to the Bayesian flow diagram in Figure 1.6, the information needed, the steps taken, and the calculations made in the classical approach differ from those in the Bayesian framework.

3.5.1 Objectivity

The system should depend on objective, reproducible, and auditable methods.

The apparent objectivity offered by the classical-statistical approach is its most attractive feature. The procedure of hypothesis testing, for instance, is automatic, which ensures that two users will arrive at the same statistical conclusion; a user is, therefore, less likely to make a mistake using this approach than using an approach in which the user is expected to make a series of statistical decisions (Efron, 1986). I shall show that this automaticity is not enough, however, to keep the method objective, because of the inherent need for nonobjective heuristics (see Figure 3.7) to solve the literature problem.

The problem with the needed heuristics is that they cannot be audited. Because they have no formal basis, there is no way of recording the reasoning that goes into a heuristic such that a second analyst would arrive at the same numerical estimate for the necessary adjustment. I have pointed out in a number of places (see pages 45, 69, and 70) that prescriptions for different adjustments are too vague to be implemented, which leaves much room for variation among analysts, and, so doing, contradicts the aim of objectivity.

Thus, although the statistical procedure might be objective, the overall process of solving the literature problem is not. In this light, the p value itself, the product of the statistical procedure, is seen to be just another heuristic for decision making, rather than the determining criterion classical statisticians designed it to be.

Yet, even the objectivity of the statistical component may be called into question. Although hypothesis testing might not be objective as a way of providing the final conclusion, we might think it is objective in providing a partial answer. An alternative role for p values, for instance, is as a measure of the strength of evidence that observed data have for or against the null hypothesis. This role was, for instance, implied in the quotation (see page 70) regarding the heuristic for dealing with prior knowledge.

Most statisticians agree that the strength of evidence provided by a study should be a function of the study alone, and, hence, would be an objective measure of that study (Edwards, 1972). Bayesian statisticians (Cornfield, 1966a; Cornfield, 1966b) have analyzed this role of the p value. Berger (1985) has shown that this alternative interpretation is flawed in a number of ways. First, it is inconsistent with simple axioms of evidential support. Specifically, the p value takes into account data that were not observed,[8] rather than examining the evidential power of the observed data alone (Berger and Berry, 1988). Second, the traditional categories of strength—α of 0.05, 0.01, and 0.001, apparent support, against the null hypothesis, of 1:20, 1:100, and 1:1000—can be shown to represent evidential support, against the null hypothesis, of roughly 1:4, 1:8, and 1:244 (Berger, 1985, p. 152), values much weaker than the apparent support.[9] This disparity between the apparent strength and true strength of the p values invalidates the measure's use as an objective report of evidential strength.

3.5.2 Intersubjectivity

The system should allow for differences of opinion among readers.

If the evidential support of a study should be independent of a reader, then two readers may rationally disagree in their conclusions only if they disagree about the propriety of the statistical model, if they differ in their beliefs about the domain prior to having read the study, or if they differ in their interpretation of the study results.

If a reader thought that the reported statistical model were wrong, then she would have to choose her own model. From the classical point of view, however, there are

[8]Recall that the computation of the p value involves considering how likely were both the data actually observed *and data more extreme* to have been observed under the null hypothesis.

[9]The assumptions in this example are that the tested parameter is the mean of a normal distribution, where the variances under the null and alternative hypotheses are the same, where the prior belief is evenly divided between the null and alternative hypotheses, and where the equivalence between p values and posterior probabilities is based on a two-tailed test. The results are most sensitive to the assumption that the variances are the same in the two hypotheses.

difficulties associated with reanalyzing data. First, the correct test may be difficult to discern or to derive (component 2 in Section 3.3.2). Second, retesting data potentially violates the single-review assumption of the hypothesis-testing framework (see Section 3.3.3). Wittkowski (1982) shows how this potential violation also invalidates the construction of backward-chaining expert systems for classical statistics, which review study data multiple times before deciding on the appropriate test.

Disagreements about prior beliefs are even more difficult to resolve in the classical-statistical framework. Classical statistics does not allow for the representation of personal, prior belief, except in the form of heuristics such as those mentioned on page 70. This strategy is untenable in general, however, because it necessitates the construction of a new heuristic for each user reading each research report. As much as the assessment of prior probabilities for Bayesian systems (see Section 6.3.3) is criticized by classical statisticians for its nonobjectivity, the acquisition of such heuristics in classical systems is open to even more reproach on this same score and is less likely to be implemented successfully in an computer program that would be expected to operate automatically.

Reinterpreting the hypothesis test leads to a more specific criticism regarding prior belief. A number of Bayesians (Cornfield, 1966a; Berger, 1985) have shown that a hypothesis test can be recast in terms of the belief a reader has in the null hypothesis before and after the test. In this reformulation, the test must assume that the reader has a particular type of prior belief if it is to remain probabilistically coherent. That belief must be that the data from the study constitute the only information the reader has ever had about the parameters of interest. This assumption is equivalent to saying, in regards to the metoprolol study, that the reader, in advance of reading the study, has neither knowledge nor experience with patients who have had heart attacks, and, therefore, believes that the likelihood that the mortality rate due to metoprolol is the same as that due to placebo is equal to the likelihood that the two

3.5.3 Normativity

The system should implement methodologists' knowledge-level prescriptions.

A classical-statistical system could represent methodologists' prescriptions. Section 3.4, and Figure 3.7, in particular, suggested the information flow in such a system. The larger question of normativity is, What action *should* the physician take on the basis of the results of the study? The difficulty of the classical-statistical approach to answer this question was discussed in Section 3.5.1.

3.5.4 Flexibility

The system should be able to evaluate both the pragmatic effectiveness and the ideal efficacy of tested therapy.

The fact that the design of a study determines the appropriate analysis for the experiment leaves the reader little choice of techniques for examining the results of the study. Yet, different readers have different expectations regarding what the data of a study will tell them. Some readers will expect to assess the effectiveness of the proposed treatment, while others will expect to determine the biological efficacy. Classical methodologists argue (see Sackett and Gent (1979)) over how to extract such different functions in a formalism that does not ostensibly allow for simultaneous diversity.

3.5.5 Adaptability

The system should enable the clinician to express methodological concerns without using statistical language.

The studies cited in Table 1.1 provide testimony to the difficulties physicians have with classical statistical measures; physicians need help in understanding, at least, the basic concepts of p value and confidence intervals. A decision-support system for

rates are different. Many clinicians find this assumption clearly incorrect, because they are able to say exactly which statement they believe to be true, and how firm their belief is. Therefore, the hypothesis-testing approach makes assumptions about prior beliefs out of touch with the knowledge of the physician readers.

Finally, we have seen (Section 3.3.3) that classical statistics relegates the interpretation of clinical import to heuristic adjustment. There are problems in creating a *formal* framework for this interpretation. First, the appropriate alternative action, in the wake of rejecting the null hypothesis, will not always be apparent. For instance, rejecting the hypothesis that two mortality rates are equal implies that the two rates are not the same, but it does *not* specify which rate is better. Second, mistakes in different clinical contexts have different costs, which cannot always be folded into a single threshold. The formal solution to this last problem is to choose the levels of the thresholds depending on the type of problem. A reader might choose a stricter threshold for higher-stakes decisions, for instance; the values of α and β chosen are supposed to take into account the cost of actions based on false-positive and of false-negative conclusions. Unfortunately, investigators and reviewers generally do not make cost-dependent choices of statistical thresholds. Rather, they stick to a traditional set of criteria (e.g., an α of 0.05 for any primary clinical effect); they would otherwise be open to the charge of choosing an α to make the data "significant." Presumably, this adherence to a fixed set of thresholds is due to the difficulty in choosing different thresholds and the difficulty in auditing and reproducing (and justifying) the process of choosing them. Thus, even the accepted formal solution is heuristic and cannot be audited.

In conclusion, although there are feasible techniques to allow for intersubjectivity within the classical-statistical framework, they perforce must be nonobjective and cannot be audited, negating the very aims the framework was set up to protect.

classical statistics would need a semantic layer to provide this help. Unfortunately, *what* those semantics should be is not clear, because they are neither decision criteria nor evidential strengths, as we saw in Section 3.5.1.

3.5.6 Simplicity

A simplified system should help the physician to interpret a single article that she has selected and read; should exclude explicit knowledge about particular statisticians and investigators; should assume that the physician user has the ability to express clearly the particular problem at hand; and should support the decision making of a single physician, rather than that of an entire community.

A system designer could build a classical-statistical system to solve the literature problem in the specified, simplified context. The classical-statistical design model is tailored to the problem of working with the individual study and is easily capable of excluding idiosyncratic knowledge. The assumption that the user is able to formulate the problem at hand is made implicitly in every statistical expert system. The notion that the reader is making the decision on her own is the central pillar of hypothesis testing. Thus, these simplifications do not bias our conclusion against the classical-statistical approach.

However, if we re-examine the classical-statistical strategy as embodied in Figure 3.7, we will realize that a computer-based implementation of the strategy would stretch the limits of available technology. As suggested in Section 1.5.5, the classical approach requires a *selection*-based strategy, which could be implemented as a diagnostic expert system. The output of this step is a statistical procedure, which may be any one of a number of complicated algorithms. The test-statistic calculation is straightforward, given a program that implements the statistical procedure. The final step of heuristic adjustment is not formal and would require another type of expert system. The assessment of thresholds and adjustment heuristics—all needed by this

system—is even less formal, and is probably no easier than the problem of automating knowledge acquisition for expert systems in general. Finally, the decision produced by the adjustment step bears no certainty of being optimal. Systems geared towards statisticians have been built that perform different aspects of these tasks. However, the variety of subsystems, interfaces, and assessments needed explains the difficulty system designers have had in building general-purpose, classical-statistical expert systems for providing statistical support to domain clients.

3.6 Previous Systems

Most computer-based statistical systems have been aimed at statisticians. A few systems have been built for helping investigators to design a study. Weiner (1987) describes a research programme to develop a suite of programs for aiding in the design of clinical trials. Wyatt et al. (1991) have developed a rule-based expert system that critiques prospective designs. Their system is noteworthy in its targeting physician users.

The vast majority of systems, however, are targeted to novice statisticians who are confronted with the task of analyzing data after the completion of a study. The statistical packages, such as SPSS (SPSS, 1983) and BMDP (Dixon, 1985) are well known. A number of systems are so flexible that they are programming languages in their own right, such as the C-based S (Becker et al., 1988) and the LISP-based Lisp-Stat (Tierney, 1990). It is precisely the power of these programs that has led researchers to devise "intelligent" front ends to these systems that would constrain novice users from abusing that power. Gale's (1986) REX system was built as a front end to S, using a rule-based approach to statistical strategy (Pregibon, 1986). Nelde and Wolstenhome (1986) built a rule-based front end for the general linear modeling program, GLIM. As a final entry in this brief survey, Oldford and Peters (1988) built

an object-oriented system, DINDE, that allows the construction and visualization of multiple analyses of the same set of data.

The less constrained these systems are, the less bound they are to the classical-statistical paradigm. Wittkowski (1982) has made the point, as I have already mentioned, that systems that permit multiple analyses of the same set of data violate basic assumptions of the classical approach. Furthermore, the more general a statistical system is, the less realistic it is to expect clinicians to use them. On the other hand, there are systems that advise user's as to the propriety of different classical-statistical tests. Chavez (unpublished) devised such a rule-based expert system as part of the REFEREE research.

Rennels' (1987) system comes closest, however, to one that helps physicians to go beyond simple hypothesis testing, and to solve the literature problem. Specifically, the ROUNDSMAN system allows a user to evaluate a patient-specific therapeutic plan in terms of research articles already saved in the program's literature database. The articles have been read previously by a domain expert and are stored in a program-specific format. The format includes heuristics that compute numbers to be used to adjust the study's conclusion as it pertains to an individual patient, for instance, through the heuristic *distance* between the patient at hand and patients examined in the study; such a heuristic are measures of external validity. Other heuristics attenuate the results of the study for such problems as protocol departures; these heuristics encode issues of internal validity. The heuristics are acquired from the domain expert by a knowledge engineer. Thus, the system requires two phases: knowledge acquisition with a domain expert reading a set of articles, followed by interaction with a clinician who has in mind a particular patient.

We shall see that THOMAS and ROUNDSMAN complement each other. Rennels has tackled a wider range of methodological problems, whereas I have focused on making the system usable by less expert readers and on ensuring that the system's

recommendations have a normative basis.

3.7 Summary

The classical design model provides the essential ingredients for implementing a decision-support system designed to help physicians solve the literature problem. Such a system must represent a number of concepts: the probabilistic concepts of observations, random variables, parameters, likelihoods; the statistical concepts of samples, populations, and statistical models; and the methodological concepts of internal validity, external validity, and bias. We have found, however, that the bedrock inference concept of classical statistics—hypothesis testing—does not provide an appropriate foundation for constructing such a system: The difficulty physicians have with the inference concepts makes the techniques inaccessible. The statistical strategy of applying possibly opaque heuristics makes the approach violate important scientific aims of objectivity and auditability. The inability of the approach to represent prior belief and the inability of readers to use the results for multiple purposes in a coherent manner both limit the utility of the approach.

In the next chapter, I shall discuss a Bayesian solution that builds on the basic probabilistic, statistical, and methodological concepts, and that answers the criticisms of the classical design model that we have detailed.

Chapter 4

Bayesian Design Model

In this chapter, I shall present the framework for my decision-theory–based solution to the literature problem. I shall do so by developing a sequence of concepts that are represented in terms of influence diagrams. Beyond serving as a mechanism for illustrating those concepts, the influence diagrams here provide the knowledge representation of a design model that acts as the conceptual framework, as the specification for a computer program, and as a structure actually to be used by that program. As before, I shall use the metoprolol example to illustrate the concepts.

The concepts are presented in the same manner as in Chapter 3 to show the comparisons with the classical approach: task (Section 4.1), domain (Section 4.2), inference (Section 4.3), and strategy (Section 4.4) concepts. In particular, the comparison shows how many classical concepts are reinterpreted in the Bayesian contexts. (1) Statistical parameters are random variables, rather than fixed, unknown constants (page 87). (2) Exchangeability replaces the concept of independent, identically distributed random variables (page 90). (3) Likelihood debiasing is an important way of producing adjustments to the conclusions, taking into account methodological concerns (page 95). (4) Calculation of posterior-probability distributions replaces test-statistic calculation (page 100). (5) Utility maximization is the implementation of the

notion of making decisions on the basis of clinical significance, rather than statistical significance (page 103).

Readers familiar with Bayesian statistics may wish to skip to Section 4.5, where the entire Bayesian design model is presented, along with references to locations earlier in the chapter where relevant concepts are discussed. The model, depicted in Figure 4.12, represents the culmination of a sequence of influence diagrams presented in the course of the chapter (Figures 4.1, 4.3, 4.4, and 4.7). All readers will find useful in this section the application of the design model to the metoprolol example.

In Section 4.6, I shall examine the Bayesian design model in terms of the knowledge-level criteria established in Section 2.3. The key discussion in this section revolves around the notion of objectivity that most critics claim to be lacking in Bayesian statistical approaches. Finally, in Section 4.7 I shall summarize work of other researchers in this field.

4.1 Task Concepts

The primary task—therapy selection—is the same in the Bayesian as in the classical context. The Bayesian framework assumes the following relationships:

- The choice of the optimal therapy (e.g., nothing versus metoprolol) depends on how that choice affects the happiness, or total utility, of the patient; maximizing patient utility is the goal of therapy selection.

- The effect the treatment choice has on patient utility depends on an important outcome, such as lifespan.

- The choice to be made depends on information regarding the outcome that the reader has available from a study report.

These relationships are depicted in the initial influence diagram of the framework, shown in Figure 4.1. The relationships among therapy choice, outcome, and patient utility are represented in a *utility model,* which is indicated implicitly by the arcs incident on the patient-utility node. Typical models balance mortality gains against morbidity losses, such as gains in lifespan versus side effects from medication. Patient-specific risk and time preferences can be included in such models as well.

Figure 4.1 shows the heart of the influence-diagram–based Bayesian design model. The diagram depicts the relationships just discussed. The diagram will be extended in the course of this chapter.

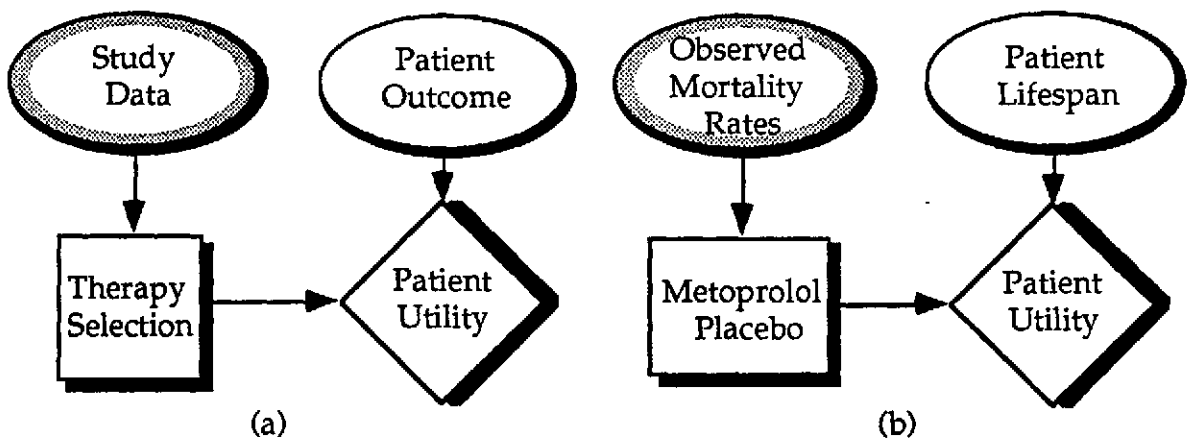


Figure 4.1: Decision component of the Bayesian design model. (a) The general model. The patient's total, expected utility (happiness) depends on the choice of therapy and the outcome that results from that choice. The utility model is implicit in the arcs incident on the patient-utility node. The fact that the clinician has the study data at hand before making her decision is represented by the arc between the study-data and therapy-selection nodes. (b) The model for the metoprolol problem. The therapy choices are metoprolol and placebo. The outcome of interest is the patient's lifespan. The study data are the observed mortality rates reported in the paper. (An alternative depiction of the relationship between therapy choice and lifespan would make the outcome node directly dependent on the decision node. However, we could not modify such a diagram as we shall need to do in subsequent figures.)

The *primary likelihood function* of the patient outcome, given the study data, is the Bayesian answer to this question: What are the contents of the channel between the reporting statistician and the reading physician? I shall explain how the likelihood is the proper channel, on page 102.

4.2 Domain Concepts

Most probabilistic, statistical, and methodological concepts—the domain concepts—are similar in the Bayesian context to those in the classical paradigm, but with a few fundamental, philosophical changes. These changes result in radically different inference and strategy concepts, as we shall see in Sections 4.3 and 4.4.

4.2.1 Probabilistic Concepts

The Bayesian approach views probability as a measure of personal uncertainty.[1] The agent's personal beliefs are the atomic probabilistic concepts in the Bayesian approach. Those beliefs, however, are still expressed in terms of probability. *Random variables,* therefore, differ in the Bayesian paradigm only in their semantics—probability representing uncertainty, not frequency—but not in their mathematical manipulation and calculation. As a result, *observations* in the world that are expressed as numerical frequencies—the atomic concepts of classical statistics—are simply one type of information that informs an agent's personal beliefs. Thus, the probability that a physician states regarding a patient's surviving an MI 3 months is a measure of a physician's knowledge and belief about the patient and his disease.[2] Study data, epidemiological results, and actuarial rates may each be used by the

[1] More specifically, we are using the *subjective* Bayesian approach. There are other Bayesian views of probability (Good, 1983).

[2] We shall use the same statistical model for the metoprolol example, given on page 55.

physician in her constructing that probability. Figure 4.2 shows the influence-diagram representation of Bayesian probabilistic concepts.

Likelihood functions and *pdfs* are the same as in classical statistics. We shall be concerned with two properties of these functions: their mean and their uncertainty. The *mean* of a pdf is the value that is needed for decision making. In a symmetric, unimodal pdf (one with only one peak), it is the value of the random variable that we *believe* is most likely; in this case, the mean is also the *location* of the pdf. The *uncertainty* of a pdf is how uncertain it is that the variable actually will take on the value at the location. This uncertainty is calculated as the *variance* or *standard deviation* (the square root of the variance) of the pdf. The *higher* the variance, the more *un*certain it is that the variable will take on the value at the location. Thus, the location of a pdf for X is $\langle X \rangle = \int x P(x)\, dx$, and the variance is $\int (x - \langle X \rangle)^2 P(x)\, dx$. A graph of the likelihood function is called a *belief curve*, to highlight its subjective use.

Parameters perform the same function in Bayesian statistics as in classical statistics: They reduce computational complexity. The semantics differ, however, because there is no need to view parameters as real, fixed, physical quantities; parameters are random variables in their own right. Because they summarize probabilities for primary variables, the uncertainty in a parameter's belief is a *second-order probability* (Kyburg, 1987; Pearl, 1987) in the primary variable. This second-order probability is useful in expert systems for controlling the system's sequence of actions (Heckerman and Jimison, 1987), and is useful in planning research and guiding further investigation. The purpose of research, from the Bayesian point of view, is to refine belief in the value of a parameter—to narrow the uncertainty of the parameter's pdf. As an example, we may be uncertain about p, the probability of a patient's dying in 3 months after an MI, and we can express our uncertainty about p in a pdf; Figure 4.2d gives an example of such a belief curve.

One implication of the random-variable nature of parameters is that the belief a person has in those parameters must be modeled and expressed *before* any research data pertaining to the parameters have been considered. Such *prior probabilities* allow an analyst to include in a Bayesian analysis just the sort of personal domain knowledge missing from classical analyses. Sets of prior probability may be assessed subjectively, or, because they are probability distributions, they may be parameterized, and only a small number of parameters need to be assessed; these new parameters are called **hyperparameters.** For instance, one common tactic for dealing with prior belief about a binomial rate, such as p, is to represent the prior belief as a beta distribution, $\mathcal{BE}(\alpha, \beta)$; α and β are the hyperparameters. The combined model of data, $p, \alpha,$ and β is called the *beta-binomial* model. The mean belief (location) and the variance, or uncertainty, are calculated from the following equations,

$$\begin{aligned} \nu &= \alpha + \beta, \\ \mu &= \tfrac{\alpha}{\nu}, \\ s^2 &= \tfrac{\mu(1-\mu)}{\nu+1}. \end{aligned} \tag{4.1}$$

These terms have the following semantics: ν is the *effective sample size,* the analyst's total prior experience of situations similar to what she is about to observe, so μ (the mean) is the proportion of "success" events that occurred in that prior experience, and $1-\mu$ is the proportion of "failure" events that occurred in that same experience. s^2 is the variance of the distribution. As an example, a prior of $\mathcal{BE}(10, 90)$ for the 3-month mortality rate in patients assigned to placebo means that the analyst's experience is that, out of the 100 patients with myocardial infarction who constitute her experience with this disease, 10 died within 3 months after the initial event. With a beta prior model, the larger the denominator, the more certain the analyst is about her prior belief; the variance of the beta distribution narrows as the denominator increases.

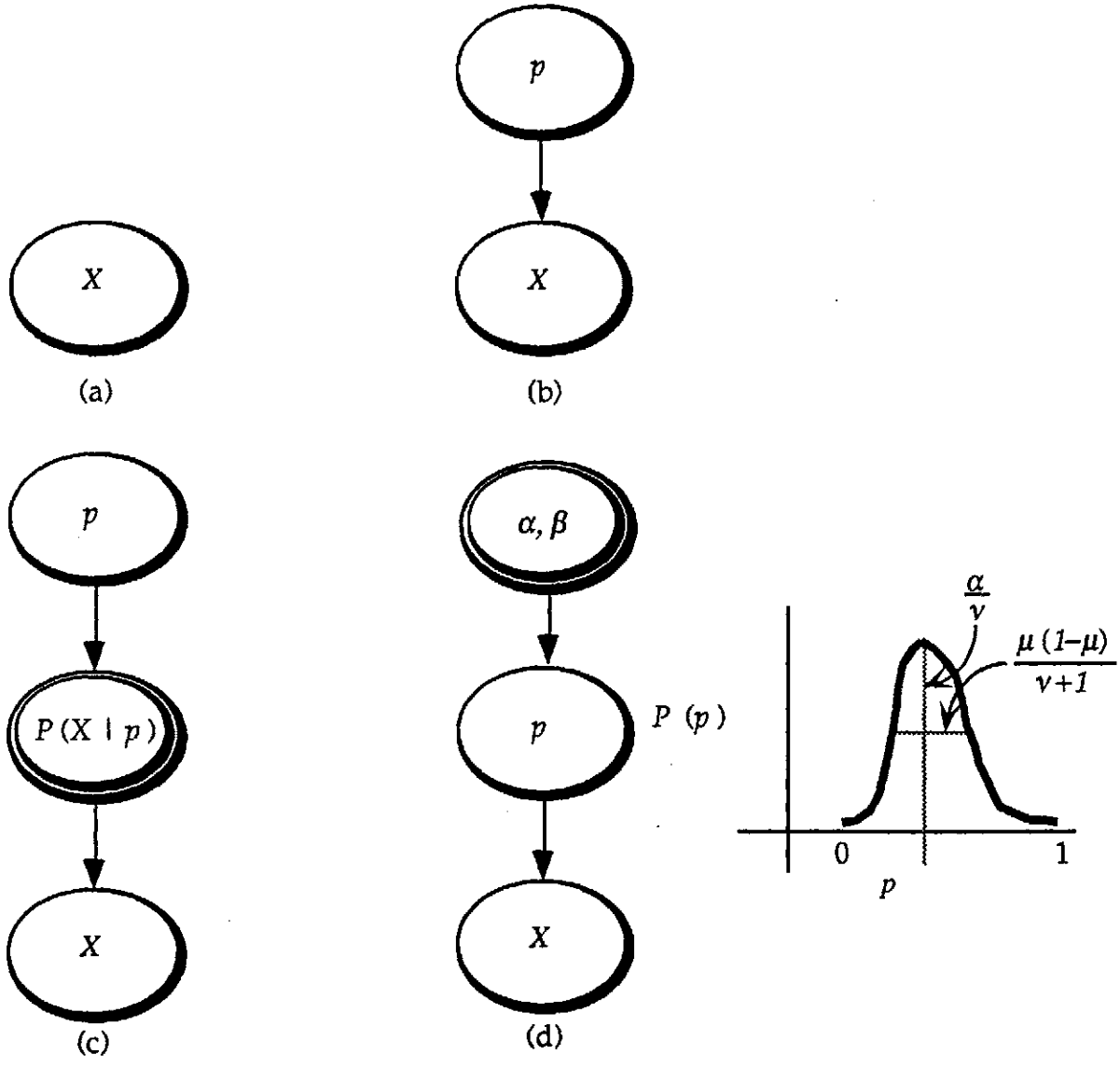


Figure 4.2: Bayesian random variables. (a) A chance node containing a label represents the random variable, X, the survival status of a patient 3 months after acute survival of an MI. (b) The uncertainty of X is dependent on a parameter, p, which is, itself, a random variable. (c) The dependence of X on p is such that, if p were known, then the uncertainty in X, given the values of that parameter, would be fixed; X remains uncertain. The content of the dependence is the same as in the classical case (Figure 3.1d). (d) The prior belief in p may be defined by the hyperparameters, α and β (placed together in a single node)—two fixed, constant parameters of a beta distribution. The graph displays the prior distribution for particular values of α and β, with shaded lines showing the location ($\mu = \frac{\alpha}{\nu}$) and uncertainty ($\frac{\mu(1-\mu)}{\nu+1}$) of the distribution, where $\nu = \alpha + \beta$ is the effective sample size.

These probabilistic concepts extend the Bayesian framework of Figure 4.1 by modeling the uncertainty of the patient's outcome (see Figure 4.3). The arc between the parameter and the outcome represents a likelihood function. The patient's outcome is a random variable that the physician reader has not observed at the time of the decision, but about which she has subjective belief, whereas the study data are random variables whose values she has observed, through reading the report. Finally, a chance node is used in the figure to indicate explicitly the random-variable nature of the parameter.

To complete the "circuit" in this figure, we need to relate the parameters to the study evidence. The combination of study data with prior belief in the parameters results in *posterior probabilities* that represent posterior beliefs for the parameters. We shall see, in Section 4.3, how to calculate these posterior probabilities. First, however, we need to understand, from the Bayesian perspective, the relationship between parameters and study data.

4.2.2 Statistical Concepts

Bayesians have the same interest in learning from study data as do classical statisticians. Bayesians, however, cannot appeal to physical reality in defining the relationship between observations and parameters; instead, they must invoke subjective reasoning to permit their use of statistical models. The key tool for translating subjective modeling into statistical modeling is the notion of *exchangeability.*

Consider the metoprolol statistical model (page 55). In that example, I made the assumptions that each patient death is individually related to the parameter, p, in the same way, and that each such outcome is not influenced by the outcome of any other patient. Classical statisticians interpret these assumptions as reflecting physical reality. However, the assumption is actually subjective knowledge, because there is no assurance that it is, in fact, true. For instance, it may be that nurses in the hospital

become more experienced over time, learning from the survivals and deaths of patients under their care; later outcomes *would* then be dependent on earlier outcomes. Or, some of the patients assigned to metoprolol may not have complied with therapy—for example, by not having taken the drug for the entire course of therapy; the outcomes of these patients are dependent on a parameter different from the parameter that governs the outcomes of those who received the assigned medication in the manner prescribed. The Bayesian way of expressing this assumption is through the concept of **exchangeability** (de Finetti, 1974): *If you believe that the conclusion about p would be the same after observing the patients* in any order, *then the patient deaths are said to be* exchangeable observations. In particular, *any future death is equivalent to any previous one.* Thus, observations in the past have implications for the future. de Finetti (de Finetti, 1974) proved that, in the case of an infinite number of observations, the assumption of exchangeability is identical to the statistical notion of independent, identically distributed random variables (iid; see page 56). This theorem has been extended to finite samples (Ericson, 1969; Diaconis and Freedman, 1980). These theorems provide for Bayesian statisticians a justification for the use of classical-statistics modeling techniques. Thus, the analyst makes, *subjectively* the assumption of exchangeability. Domain experts can understand and make this assumption, as well.

In extending the influence diagram for the literature problem, statistical models define the relationships among the parameters and the study data, and the relationships among the parameters and the patient outcomes; see Figure 4.4a. For the metoprolol problem, we assume an exponential distribution, with the parameter, λ—the instantaneous mortality rate.[3] The influence diagram expresses the notion that study data and the patient outcome are exchangeable.

[3]The relationship between p and λ is that $\lambda = \frac{1}{\Delta t} \cdot \ln \frac{1}{1-p}$, where Δt is the period of observation; see Equation 5.5.

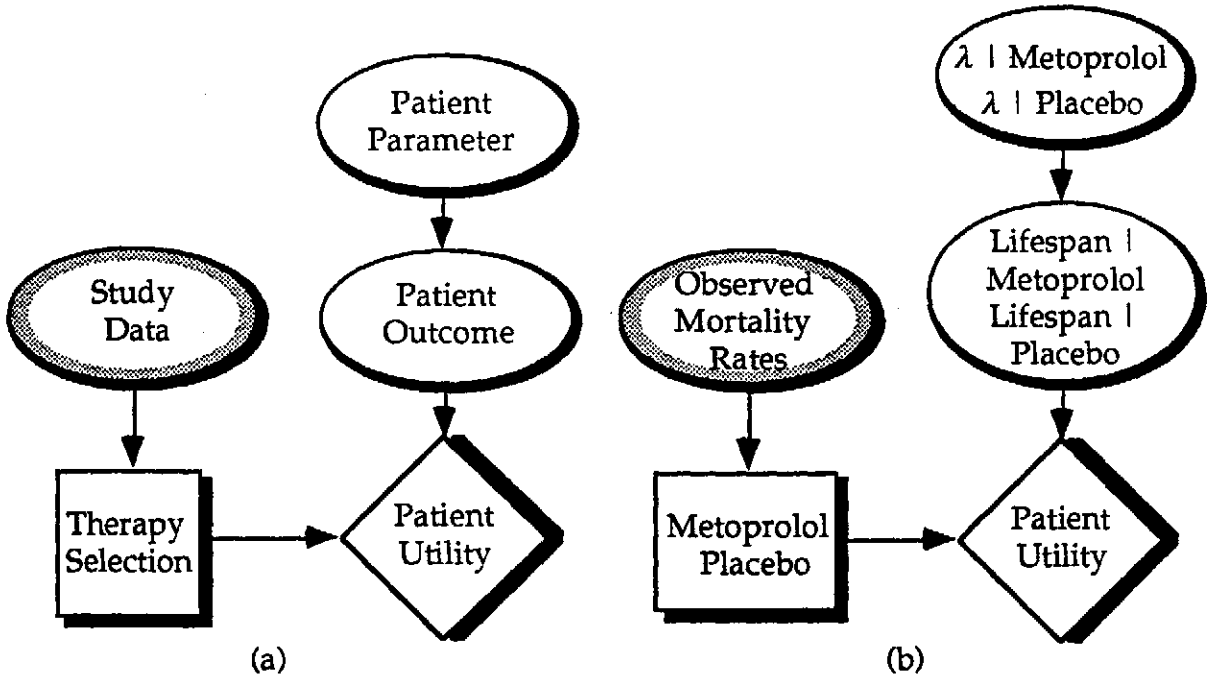


Figure 4.3: The parameter component of the Bayesian design model. (a) The general model. The uncertainty in the patient's outcome is modeled as depending on a parameter. The actual parametric model is indicated implicitly by the arc between the parameter and the outcome nodes. (b) The model for the metoprolol problem. The uncertainty in a patient's lifespan is modeled as depending on a single parameter, the instantaneous mortality rate, λ. There is one such parameter for each therapy selection.

The influence-diagram representation of exchangeability for a set of observations, a parameter, and any other variables has three conditions:

1. The observations are conditionally independent of one another, given the parameter (i.e., the same parameter governs each outcome).

2. The probabilistic dependence of each observation on the parameter is the same.

3. If any variable is dependent on one observation, then that variable must be dependent on all the observations. Its likelihood function, for the case of a dependent chance-node, or its deterministic function, for the case of a dependent deterministic-node, given those observations, must be symmetric in all the observations. So, statistical functions of observations must be symmetric in the observations.

The case where nurses' abilities to care for heart-attack patients improve over time violates condition 1, and the case of noncompliance violates condition 2. The graphical representation of exchangeability (Figure 4.5) is identical to that of iid (see Figure 3.2), except for the random-variable nature of the parameters; the hyperparameters for p are indicated in the figure. Note further the implicit equality of the likelihoods for the observations (each is $\mathcal{B}(p)$), and the symmetry of the dependent statistic $D = \sum_{i=1}^{698} X_i$, in its parents, the X_i.

The notion of exchangeability defines the *relevance* of a study for a particular decision problem, and lies at the heart of the reader's dilemma: Are patients in the study "like" the patient in the clinician's care? If the clinician thinks that the two sets of patients are *not* the same, then the assumption of exchangeability does not

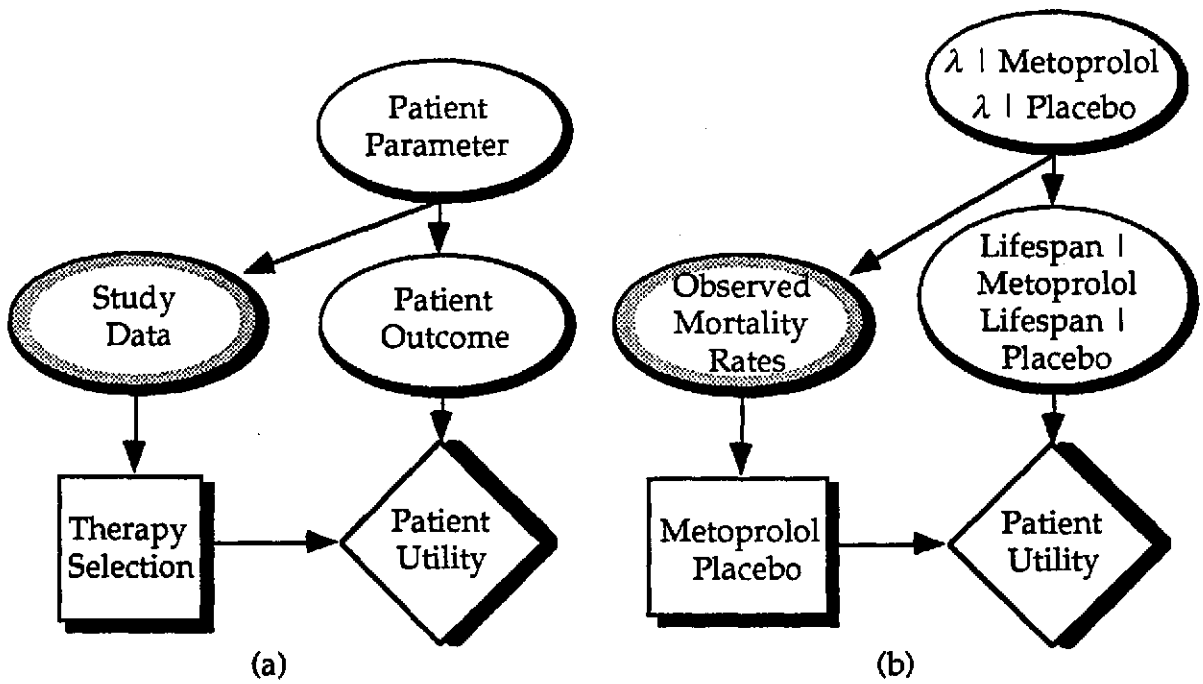


Figure 4.4: The statistical component of the Bayesian design model. (a) The general model. The statistical model for the study data is implicit in the arc between the parameter and the study-data nodes. In this diagram, the study data and patient outcome are exchangeable. (b) The model for the metoprolol problem.

hold. In our framework, the way of dealing with this assumption violation is to model the reasons for the violation. This consideration brings us to the issue of modeling methodological concerns.

4.2.3 Methodological Concepts

Statistical models akin to those of classical statistics can be used in Bayesian statistics *if the analyst assumes exchangeability.* When exchangeability can*not* be assumed, the Bayesian statistician has two options. The first is the *predictive-modeling* approach. Using this strategy, the statistician ignores parameters altogether, and models directly his belief in a future observation in terms of previous data (Hill and Weisman, 1991). Thus, the analyst might be asked to state the probability of a future death given the deaths observed in the study, without reference to population mortality rates. This strategy obviates the need for statistical models, which we have specified as representing the relationship between the outcome of the patient at hand and the data from the study; we shall not use this strategy in solving the literature problem. The second is the *model-refinement* approach, of which there are two strategies: likelihood debiasing and hierarchical modeling. Using these approaches, the statistician encodes knowledge about methodological concerns. We shall find that, when we apply this strategy, the notions of population, sample, and bias recur. Because only likelihood debiasing is used in THOMAS, I shall describe that strategy.[4]

Likelihood debiasing requires the statistician to express the original parameter in terms of other parameters. This approach calls into play new parameters. I distinguish three types of parameters: (1) the *population* parameter, which is the parameter governing the outcome for a general subject, or the parameter of interest; (2) the *study* parameter, which is the parameter governing the patients in the study,

[4]See page 226 for a discussion of hierarchical modeling.

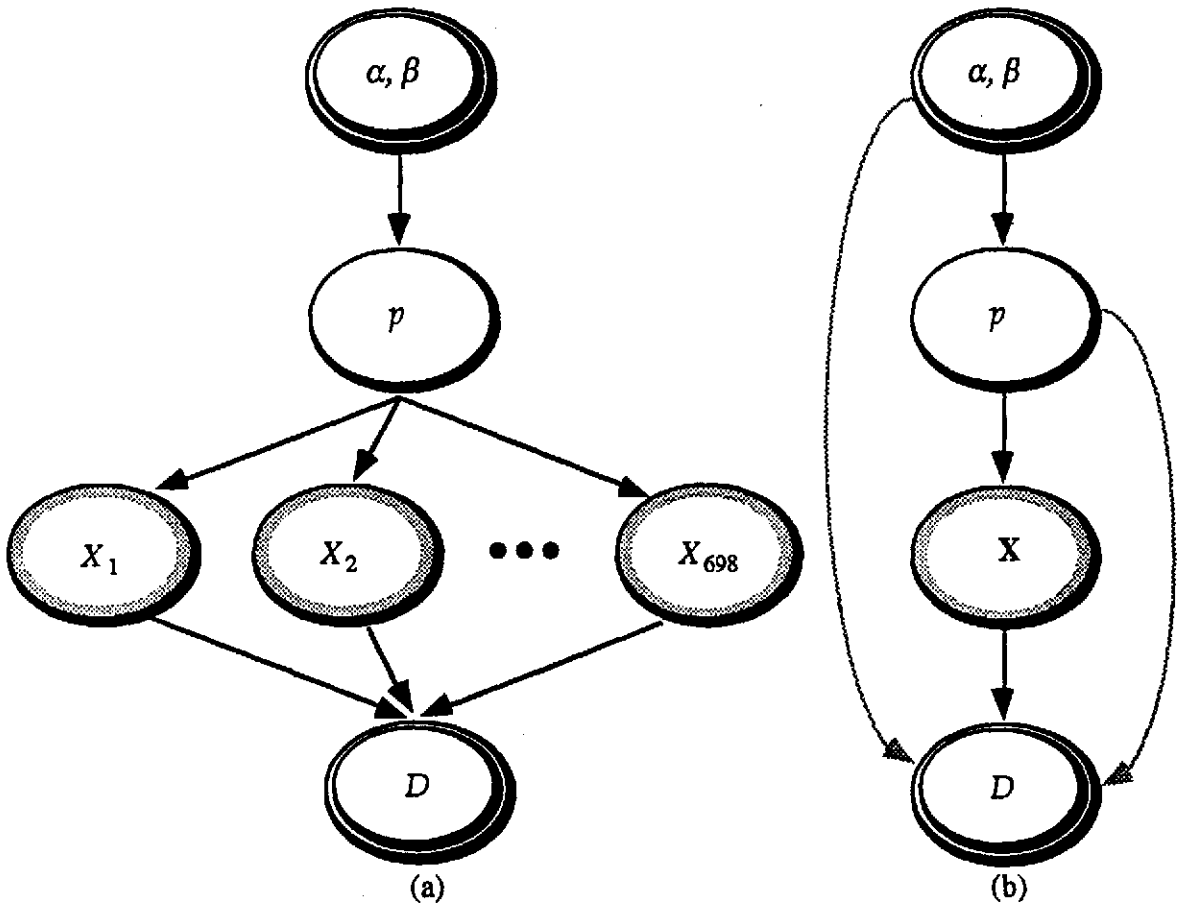


Figure 4.5: Bayesian model for exchangeable observations. (a) 698 patient outcomes (X_i) are modeled as conditionally independent of one another, given the parameter, p. The likelihood of each outcome given p is the same (i.e., the Bernoulli distribution, $\mathcal{B}(p)$). The hyperparameters, α and β, govern the prior-probability distribution for p. The X_i have been observed. D, the number of deaths, is a statistical function of the observations, $D = \sum_{i=1}^{698} X_i$. (b) An alternative graphical convention for the model in (a) is shown. Here, $\mathbf{X}$ is a vector of the values $(X_1, X_2, \ldots, X_{698})$. The derived-probability arcs denote the binomial distribution (on the right) and the beta-binomial distribution (on the left).

or a subgroup of those patients; and (3) the *effective* parameter, which is the parameter governing the observation for a specific subject of the experiment. Consider the metoprolol example, in the case of noncompliance (page 90), where condition 2 of exchangeability (page 93) is violated. If we have to dispense with exchangeability, then the Bernoulli parameter for each patient is different, and we need 698 such parameters, as in Figure 4.6a. Instead, however, we might use the model shown in Figure 4.6b, where these parameters constitute two groups: One group shares a common *study* parameter (the mortality rate of patients assigned to metoprolol who received it) and another shares a different study parameter (the mortality rate of patients assigned to metoprolol who were noncompliant for a common period of time). Then, the *effective* parameter for each patient in a group is dependent on—and identical to—the group's study parameter; the patient outcomes within each group are exchangeable. The study parameters, in turn, depend on the *population* parameters. For the metoprolol group, the dependence is again that of identity. For the noncompliance group, the dependence is that the study parameter is a functional mixture of the population parameters and an ancillary *methodological* parameter—the proportion of time the patients were compliant. The relationship between the population parameters—the parameters of interest for making the decision—and the study data is now fully specified. The pieces of the relationship are straightforward. I refer to this strategy as *likelihood debiasing*,[5] because the resulting likelihood function for an observation given the population parameters is more complex than it was before the introduction of the extra parameters, but it remains well defined.

Figure 4.7 suggests how we can extend the Bayesian design model in a number of ways, using likelihood debiasing. For instance, we can model methodological issues of internal validity. Protocol departures—ways in which the study as executed deviated from its design—divide the study patients into different groups: patients

[5] I thank David Spiegelhalter for suggesting this expression.

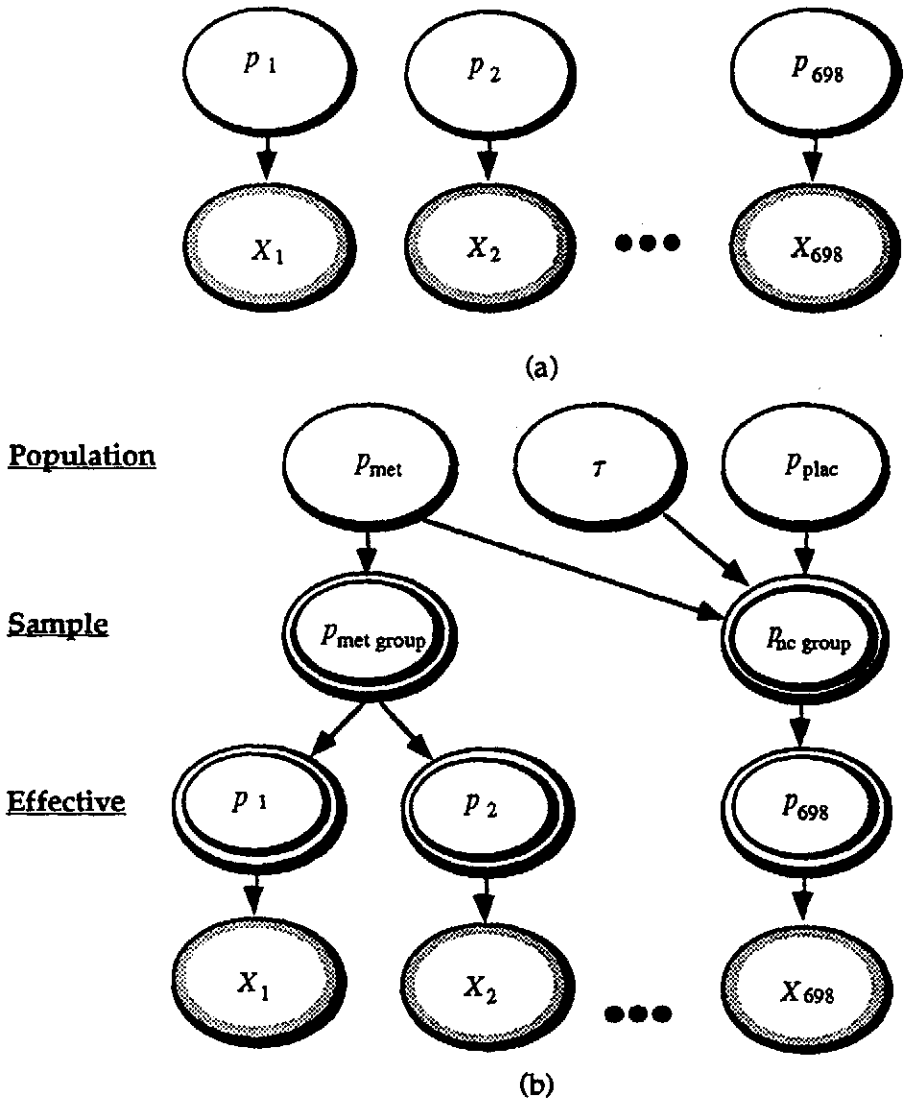


Figure 4.6: Likelihood debiasing. 698 patient outcomes, X_i, that are *not* exchangeable, because of noncompliance. (a) The likelihood for an individual patient is $X_i \sim \mathcal{BI}(p_i)$; the relationship between an individual p_i and the population parameter of interest, p (not shown) is unclear. (b) A likelihood-debiasing model for the nonexchangeability of the observations employs the notion of different *levels* of parameters: the *effective* parameters for the individual observations, the *study* parameters for groups sharing properties in common, and *population* parameters for the parameters of interest. In this example, the effective parameters are identical to the corresponding study parameters. The compliant group's study parameter is identical to the population metoprolol mortality-rate parameter. The *non*compliant group's study parameter is equal to a mixture of the metoprolol and placebo mortality-rate parameters. Thus, $X_{698} \sim \mathcal{BI}(p_{\text{nc group}})$, where $p_{\text{nc group}} = \tau p_{\text{met}} + (1 - \tau)p_{\text{plac}}$; the relationship between p_i and the population parameters is fully specified.

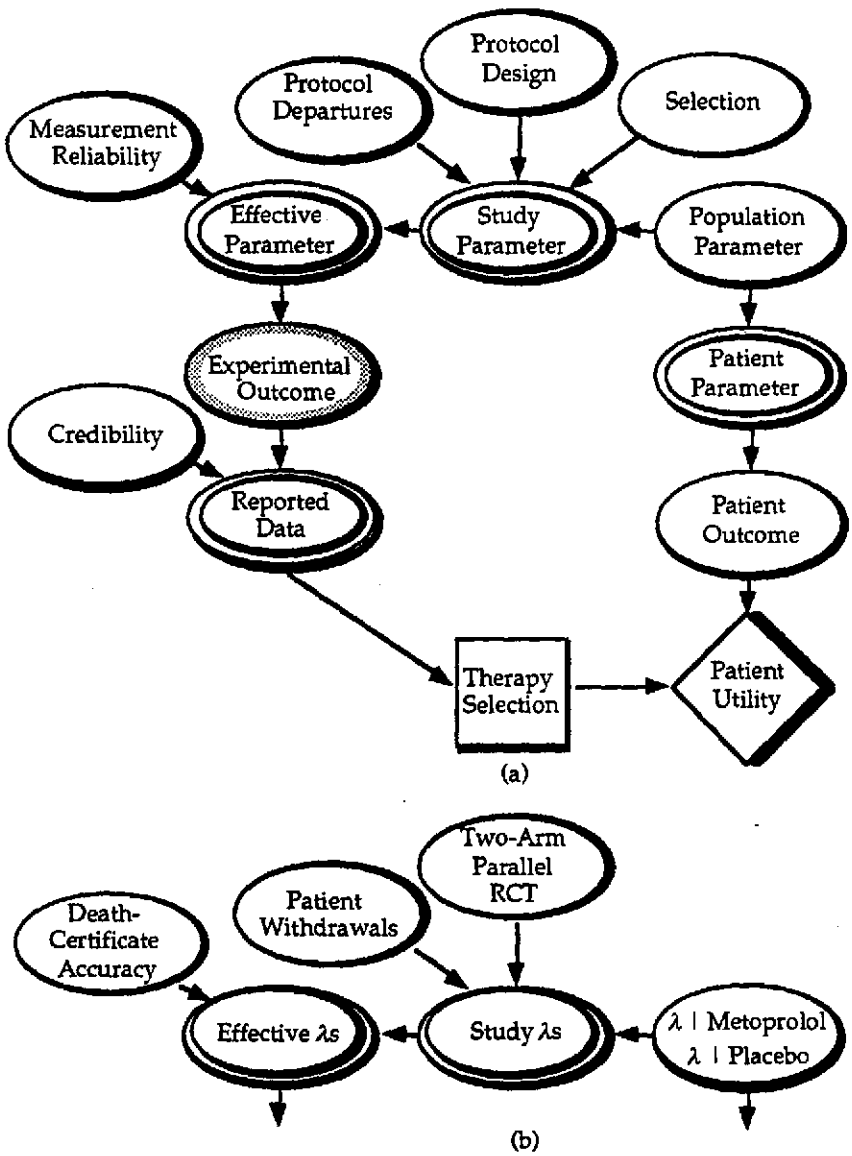


Figure 4.7: Likelihood debiasing in the Bayesian design model. (a) The general model. The relationship between the patient parameter and the study data has been expanded in terms of the population, study, and effective parameters. The methodological concerns of selection, protocol design, protocol departures, and measurement reliability are represented in the model in the indicated places. (b) The likelihood debiasing model for the metoprolol example. This model contains the following specializations of the general model: The study is a two-arm parallel randomized clinical trial, a protocol departure of concern is withdrawal of patients from their intended therapy, and a source of measurement error is the inaccuracy with which death certificates establish a person's mortality status. *RCT* denotes randomized clinical trial.

who violated protocol and patients who did not. The two groups have different study parameters, and likelihood debiasing is used to define those parameters, as in Figure 4.6b. Measurement error—ways in which is the results observed differ from the true outcomes—manifest themselves in differences between study parameters and effective parameters; likelihood debiasing is used to define those differences.

4.3 Inference Concepts

Inference, in Bayesian statistics, involves questions about the parameters and about the decision of interest. The procedure that answers the first set of questions is probabilistic updating, and the procedure that answers the second set is utility maximization. I shall discuss these procedures; then, to complete the parallel with the classical design model, and to lay the groundwork for THOMAS, I shall discuss the Bayesian perspectives on metadata and adjustments.

4.3.1 Probabilistic Updating

The goal of probabilistic updating is to find values of the parameters that are most consistent with the prior belief, with the data observed, and with the known relationships among parameters. Probabilistic updating, using influence diagrams, has shown itself useful in domains such as medical diagnosis (Heckerman et al., 1990), robot vision (Agosta, 1988), and medical technology assessment (Shachter, 1990). Figure 4.8 depicts the influence-diagram representation of probabilistic updating.

Given a statistical model, the Bayesian investigator is interested in the *posterior-probability distributions* for the parameters, given the data observed. These are "posterior" beliefs, because they describe the beliefs the investigator should have *after*

study data have been observed. Bayes' theorem gives the method of calculating posterior belief from prior belief and from the evidence:

$$P(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta)\, P(\theta)}{P(\mathbf{X})}, \tag{4.2}$$

where θ is the parameter of interest, and $\mathbf{X} = \{X_1, \ldots, X_n\}$ are the observed data. Equation 4.2 states that the posterior belief in a particular value of the parameter, given the observed data, is a function of three pieces of information: (1) the likelihood of the observed data, given the value of the parameter; (2) the prior belief in that value of the parameter; and (3) the overall probability of having observed the data. Equation 4.2 follows directly from the axioms of probability, once we allow for a parameter to be a random variable. Because the new belief is the old belief updated by the new evidence, the process of calculating the posterior belief is called *probabilistic updating.*

Note that the overall probability of the evidence is a constant *once the data are observed.* If we were to ignore the denominator in Equation 4.2, we would have, for each possible value of the parameter, a number equal to the posterior probabilities, scaled by the same number, $P(\mathbf{X})$. These new numbers are called the *posterior likelihoods* for the parameter, and, in fact, many Bayesians refer to a **likelihood function for the parameter,**

$$\ell(\theta \mid \mathbf{X}) = P(\mathbf{X} \mid \theta)\, P(\theta). \tag{4.3}$$

The Bayesian literature uses the same expression, *likelihood function,* to refer to the probabilistic dependence of the data on the parameter, and to refer to the posterior likelihood of the parameter. To avoid confusion, I shall thus use **primary likelihood function** to refer to the former ($\ell(\mathbf{X} \mid \theta)$), and *likelihood function for the parameter* to refer to the latter ($\ell(\theta \mid \mathbf{X})$).

The likelihood function for the parameter plays a significant inferential role in Bayesian statistics. Note that, once the data are observed, $\ell(\theta \mid \mathbf{X})$ is a deterministic

function of the data; this likelihood is thus a *statistic.* The fundamental principle of Bayesian statistics, called the **Likelihood Principle,** is that *the likelihood function for the parameter is the statistic necessary and sufficient to update a person's belief in the parameter of interest.* This principle may be viewed as an axiom of Bayesian statistics (Berger and Wolpert, 1984), or may be derived as a theorem (Birnbaum, 1962; Berger, 1988), based on commonly accepted axioms. The importance of the Likelihood Principle cannot be overstated, for the principle obviates the need to create any of the well-known statistical tests, such as the z-score and t-test, for evaluating the inferential implications of data. We should not be surprised to find, then, that Bayesian inference has no need for statistical tests.[6] Instead, Bayesian statisticians examine the posterior-probability distribution functions, calculated via the same procedure—probabilistic updating—regardless of the statistical model employed. The difficulties encountered in Bayesian updating are numerical, such as how to ensure the stability of an update, or how to perform the update for an arbitrary function; they are not conceptual.

Conjugate prior-probability distributions form a class of probability distributions that lack some numerical difficulties of generalized probabilistic updating. **Conjugate** *distributions are pdfs where the prior- and posterior-probability distribution have the same form, with respect to a class of primary likelihood function.* Thus, the beta distribution is the conjugate prior for binomial evidence. There are two advantages to using this class of prior pdfs. First, there are closed-form solutions to the probabilistic updating. Second, updated belief in a parameter can be expressed as altered values of the parameters modeling the belief in that parameter. Thus, if prior belief in a mortality rate were expressed as $\mathcal{BE}(\alpha, \beta)$, and if s "successes" were observed along with f "failures," then the posterior distribution would be expressed as $\mathcal{BE}(\alpha + s, \beta + f)$.

[6]This statement is not true, however, with regard to model selection; see Section 4.4.

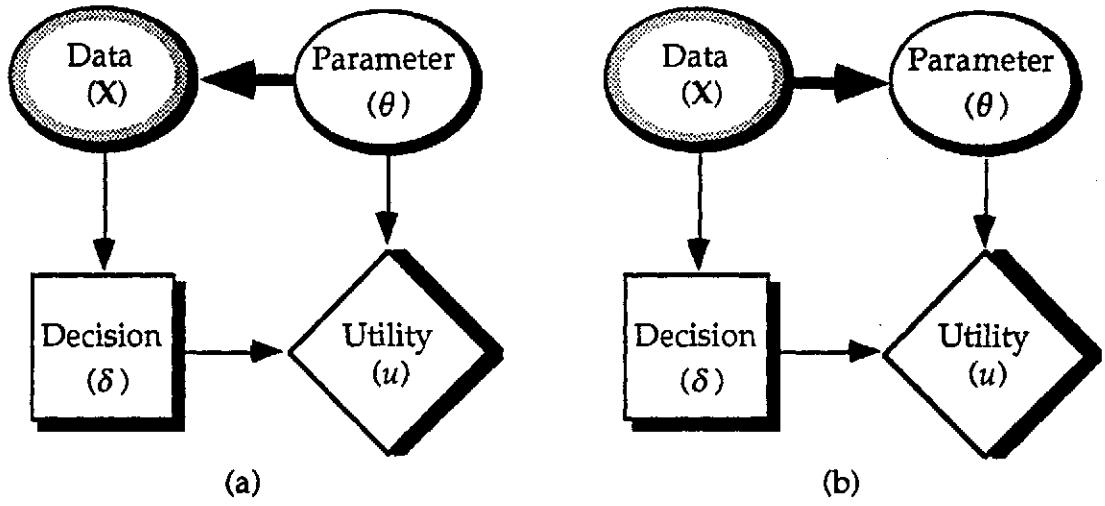


Figure 4.8: Probabilistic updating. (a) Initial structure. This structure is similar to that in Figure 4.4a. (b) After updating. The node for the parameter θ contains the posterior belief in the parameter, given the data, $\mathbf{X}$; utility maximization can now proceed. The arcs between the parameter and data nodes are emphasized to point out that probabilistic updating entails creation of the likelihood function for the parameter, $\ell(\theta \mid \mathrm{X})$.

4.3.2 Utility Maximization

Decision analysis is the field concerned with discerning the optimal, normative decision among a series of choices, generally under uncertainty. The representation of uncertainty in this field is provided by Bayesian, subjectivist probability. The representation of optimality is provided by utility theory. von Neumann and Morgenstern (1947), in their seminal book, developed an axiomatic system to encode the preferences of individuals to allow for normative decision making. Utility is a finite, subjective scale that is isomorphic to the closed interval [0, 1]. The normative, optimal decision is defined as that decision that maximizes the expected utility of the decision maker:

$$\delta^* = \max_{\delta \in \mathcal{A}}{}^{-1} \int_{\Theta} u(\delta, \theta)\, P(\theta \mid \mathbf{X})\, d\theta, \tag{4.4}$$

where δ^* is the optimal decision, $\mathcal{A}$ is the space of all possible decision alternatives, and $u(\delta, \theta)$ is the utility of taking action δ, when the parameter of interest takes on the value, θ; the other variables are defined as previously. The decision maker's attitudes to risk, time, and other tradeoffs are represented in the utility function.

4.3.3 Bayesian Metadata

In addition to the types of data involved in a particular study, Bayesian metadata include information about the types of parameters concerned in the problem, and about the availability of information about either data or parameters. Knowledge about the nature of a parameter might be used, for instance, in choosing the shape of the parameter's belief curve. Information about data availability is important, because just knowing that a particular type of datum is available is information that can update a Bayesian analyst's belief. The use of such metadata is a central theme of Chapter 7.

4.3.4 Adjustments

As discussed in Section 3.3.4, the classical design model asks us to adjust conclusions, post hoc, on the basis of information not included directly in the analysis. Such information is often the subjective type of knowledge encoded in Bayesian prior-probability distributions. Thus, the Bayesian approach does not require post hoc adjustments. Rather, "adjustments" to any estimates or conclusions result from the structure of the model employed in the analysis and from the form of the prior-probability distributions used.

4.3.5 Reformulation of Classical Measures

Many classical-statistical measures are reformulated in the Bayesian context; THOMAS uses these reformulated measures in its interaction with the user. We shall find that the posterior-probability distribution contains the information needed to construct the Bayesian equivalent of each classical statistic.

4.3.5.1 Parameter Estimation

The classical task of parameter estimation is to derive a statistical function that, on average, gives an estimate of the true parameter that is unbiased and has the least variance. Figure 3.4 depicted this process. By *unbiased,* we mean that, with increasing sample size, the estimate approaches the true value of the parameter. The Bayesian equivalent is expressed in Figure 4.9—choosing a value that maximizes the user's utility. In statistics, the utility function often used is the domain-independent least-squares function: The further the estimate is from the true value, the lower the utility. In this context, the *mean* of the posterior-probability distribution is the function that produces the desired estimate (Berger, 1985).

4.3.5.2 Confidence Intervals

Consumers of classical-statistical analyses often have a misguided understanding of how a confidence interval expresses the statistician's uncertainty in his parameter estimate. The posterior-probability distribution can offer the semantics the consumers want: The area under the pdf bounded by two values of the parameter has exactly the semantics of the degree to which the user believes the value of the parameter lies between those two values. However, the implied question, "What interval contains the desired amount of the user's belief?" (e.g., the ubiquitous 95 percent) actually has an ambiguous answer. The interval may lie, for instance, from the left-hand limiting

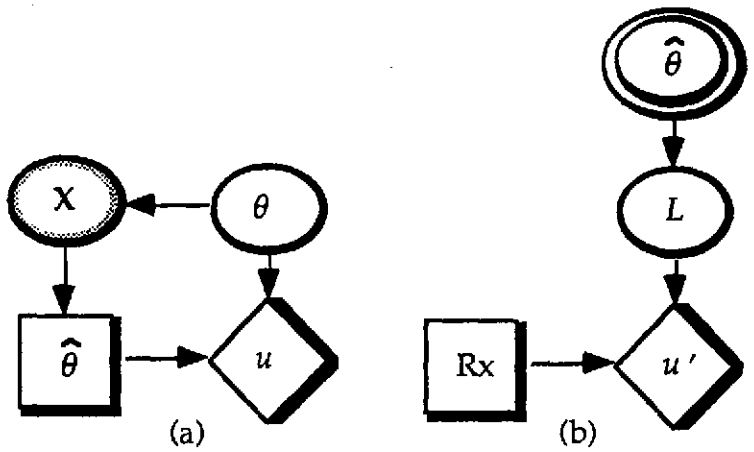


Figure 4.9: Bayesian parameter estimation. (a) The choice of an estimate, $\hat{\theta}$, based on data, $\mathbf{X}$, results from utility maximization of the utility function dependent on that estimate, and the true parameter, θ. The utility function may, or may not, be domain-dependent. (b) The estimate is then used as a *constant* in the domain decision problem. The decision problem in this figure that of treatment choice (Figure 4.4). The utility function is domain-dependent.

value to some upper bound (Figure 4.10a), or it may lie from the right-hand limiting value to some lower bound (Figure 4.10b). The Bayesian convention is to find the narrowest interval around the posterior mean that contains the desired amount of belief (Figure 4.10c); this interval is called the **credible set** (Berger, 1985).

4.3.5.3 Hypothesis Testing

The traditional hypothesis test examines the implications of a parameter of interest taking on a particular value. For instance, in the metoprolol example, we create and test a new parameter, the difference between the drug-induced mortality rates (i.e., the mortality rate due to placebo minus the mortality rate due to metoprolol) (see Section 3.3.2). In the Bayesian paradigm (see Berger (1985)), we incorporate that difference parameter into the statistical model and arrive at a posterior-probability distribution for the difference (Figure 4.11b). The area under the belief curve for values of the difference greater than zero gives the posterior belief in the statement

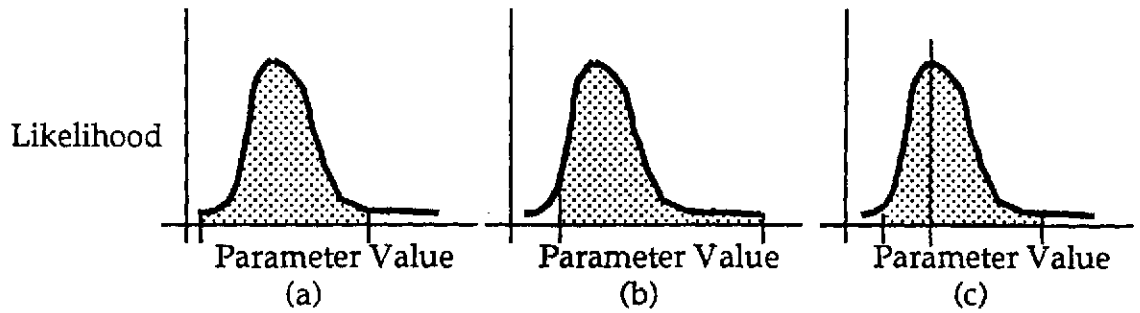


Figure 4.10: Credible sets. These sets refer to an interval over which the belief is of a certain level. In each case, the endpoints satisfy the relationship that $P(a \leq \theta \leq b \mid x) = 1 - \alpha$, where a and b are the endpoints, θ is the parameter of interest, x is the observed data, and α is the small amount of error probability desired. Compared this relationship with the statement implied by a confidence interval: $P\left(a(x) \leq \theta \leq b(x) \mid \theta\right) = 1 - \alpha$, where $a(x)$ and $b(x)$ make clear that the endpoints are functions of the data. Note that this statement is conditioned on knowing the value of the parameter. (a) One-tail credible set, with the parameter's minimum value as one anchor point. (b) One-tail credible set, with the parameter's maximum value as one anchor. (c) The standard credible set, which is the narrowest credible set of the specified belief that contains the mean value of the parameter.

that the mortality rate due to placebo is larger, on average, than the mortality rate due to metoprolol. The larger that area—that belief—the more certain we are about the conclusion about the relative efficacy of the two drugs. Above a certain threshold value of belief, we may choose to accept that conclusion as true. A popular threshold is 95 percent. Clearly, this threshold is arbitrary and community-bound. Regardless of the threshold, this measure is irrelevant for decision making, where the individual posterior-probability distributions are all that are needed, as we saw in Section 4.3.2.

4.4 Strategy Concepts

I first shall show the sequence of steps—the strategy—dictated by the Bayesian approach, and then shall focus on the problem of selecting the appropriate probabilistic

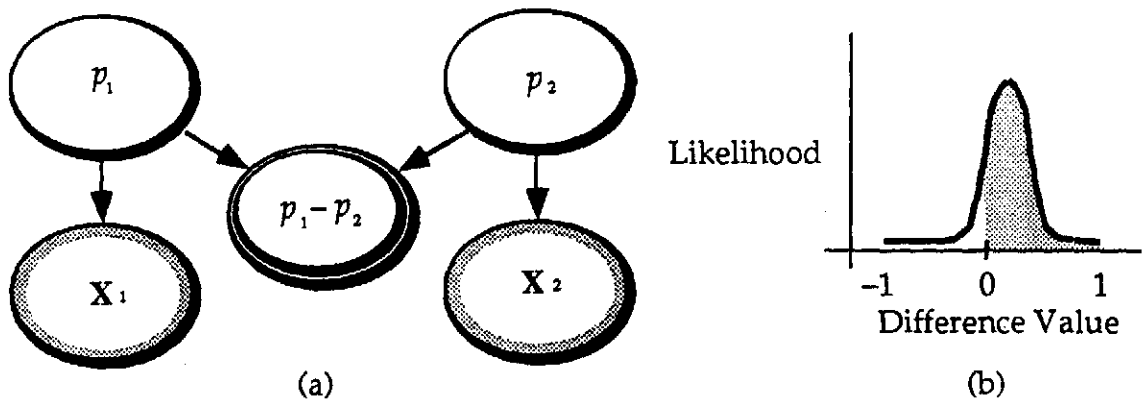


Figure 4.11: Bayesian hypothesis testing. This example shows the test for the hypothesis that the difference between two mortality rates is positive (the Bayesian z-test). (a) The influence diagram for the test. The parameter $\Delta_{\text{ctl,exp}} = \theta_{\text{ctl}} - \theta_{\text{exp}}$ is the difference between the control-treatment and experimental-treatment mortality rates. Compare with Figure 3.6.(b) Belief curve for posterior belief in the difference. If the area under the curve for positive values is greater than a certain threshold, then the reader can conclude that the drug with the smaller posterior mean mortality rate is in fact better than the other drug. The threshold must be set by the reader on external grounds. The area under the curve represents the probabilistic expression, $P(\theta_1 - \theta_2 \geq 0 \mid \mathbf{X})$, rather than the expression used in hypothesis testing, $P(t(\mathbf{X}) \mid \theta_1 - \theta_2 = 0)$, where $t(\cdot)$ is the appropriate statistical function.

model. The general strategy for performing a Bayesian analysis, depicted in Figure 4.12 (repeated from Chapter 1) is as follows:

1. Construct the analysis model relevant to the study using an appropriate methodological formulation of the problem at hand.

2. Assess the necessary prior beliefs from the reader

3. Assess the necessary patient preferences.

4. Include the evidence from the study.

5. Perform probabilistic updating.

6. Perform utility maximization.

7. Examine the posterior distributions.

8. Examine the utilities.

This sequence of steps was shown in Figure 1.6. The only step that we need to justify is the *model-construction* step (step 1).

Proper model selection in Bayesian data analysis is depicted in Figure 4.13a. In this proper analysis, the analyst must specify, for *every* statistical model in the universe, a prior belief that represents the analyst's belief that that model is appropriate to the problem at hand. The space of all parametric models is very large. Specifically, if we label models on the basis of the variables in those models (i.e., on the basis of *structure*), then the space is countably infinite, because we can enumerate the variables in some order (say, alphabetically). Consider, too, that parametric models differ in probabilistic type. Thus, we might construct a Gaussian-based, linear-regression model for the level of an outcome with parameters for age, sex, and prognosis. Or, we might construct a logistic-regression model for the probability of outcome with parameters for ethnicity and cardiovascular status. Note that the two models have no parameters in common, so their prior beliefs must be assessed separately. Even worse, the analyst would also have to specify a prior belief in *every* parameter contained in each of the models.[7] Making all these assessments is clearly impossible. Thus, it is not possible to perform a general Bayesian analysis in finite time. Therefore, analysts must take approaches that are not formally correct, but that may be informed by an understanding of the general analysis.

There are two ways of modifying the proper Bayesian approach. The first is to choose the model and the parameter estimates in some unified way (Figure 4.13b).

[7]The node labeled θ contains a countably infinite list of parameters. The distribution implied by the three nodes, M, θ, and $\mathbf{X}$, $P(\mathbf{x} \mid M, \theta)$, *selects* the parameters from that list that are appropriate to the model M.

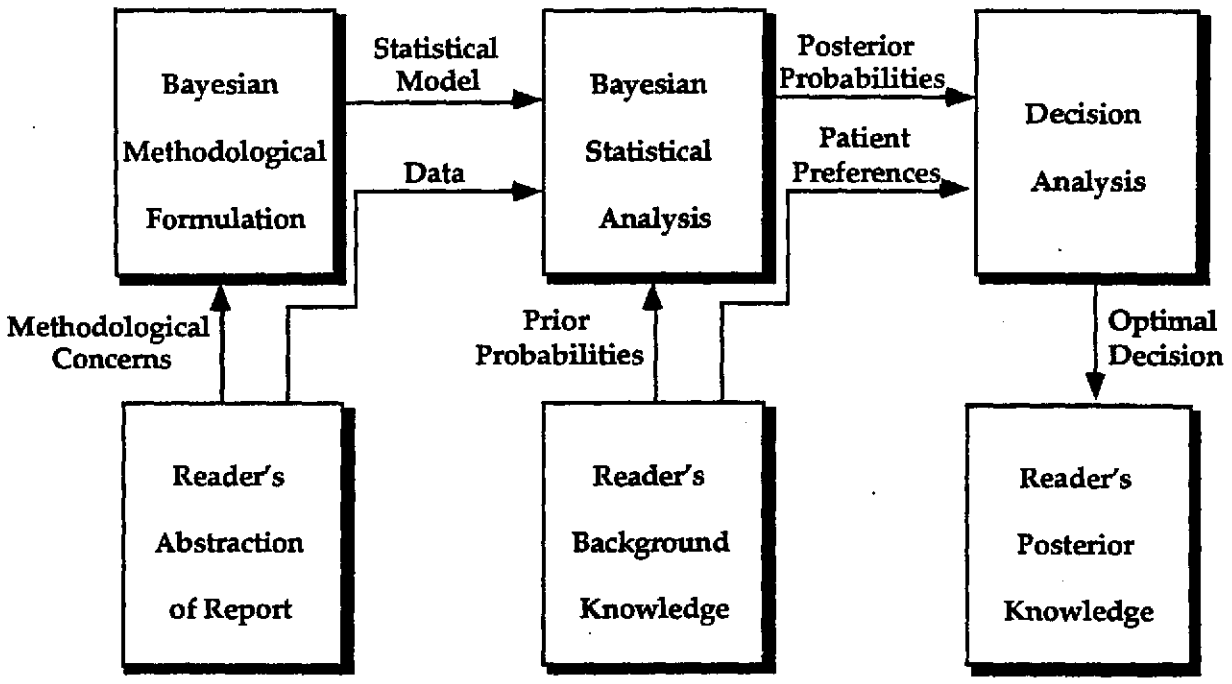


Figure 4.12: The information-flow diagram depicting the Bayesian strategy for using the Bayesian design model, same as Figure 1.6.

Here, the analyst generates *pseudopriors* over models. A pseudoprior is a weight for a model that can be calculated on the basis of attributes of the model, and that, therefore, obviates the need for subjective assessments. The calculations take into account the tradeoff between how well the model *fits* the data at hand and how many parameters are in the model. A tradeoff results because the better a model accounts for the data at hand, the less likely it is to generalize—the model is said to be less *robust*. Furthermore, the greater the number of parameters in the model, the less posterior certainty we have about their values, given the data at hand. The concerns of fit, robustness, and posterior uncertainty would be objectives in the utility function the Bayesian-style analyst would use in choosing a model and its parameter estimates. Statistical tests—here used as heuristics—are helpful in making these choices. The methods used in this strategy are numerically intense,

result in parameter estimates rather than posterior beliefs, and usually do not use a domain-based utility function (Clayton et al., 1986; Linhart and Zucchini, 1986).

The second modification is to choose the model first, and update belief in its parameters later (Figure 4.13c). There are three strategies for implementing this modification. In each of these three strategies, the final analysis performed is conditioned implicitly on the choice of the probabilistic model, as indicated in the figure.

The first strategy is to use domain knowledge to narrow attention to a number of models of interest, and then to assess prior belief in that reduced set of models (Clayton et al., 1986). Self and Cheeseman (1987) calls this strategy *transduction:* The analyst chooses a set of models that share a parameter; the prior belief in that parameter expresses prior belief in different models (Herskovits, 1991).

The second strategy is to choose a model based on principles of model quality, such as simplicity and parsimony. These principles may be both domain- and data-independent; their use is advocated by researchers in the field of abduction (Thagard, 1978).

The third strategy is to construct the model that seems most appropriate to the problem at hand, and to modify it in response to the availability of the data at hand—that is, in response to the metadata of the problem. Thus, we would introduce a noncompliance model only if we knew that there were data referable to this protocol departure. Such data might be specific, such as a listing of the numbers of patients who were noncompliant for the indicated time periods, or they might be nonspecific, such as the identity of the drug, which indicates to a knowledgeable reader the degree of compliance one can expect. Appealing to metadata prevents violation of the principle that data should be counted only once (Wittkowski, 1986), if we were tempted to use the data themselves to select the model *and* to update beliefs in parameters. This strategy is the one I take in this dissertation.

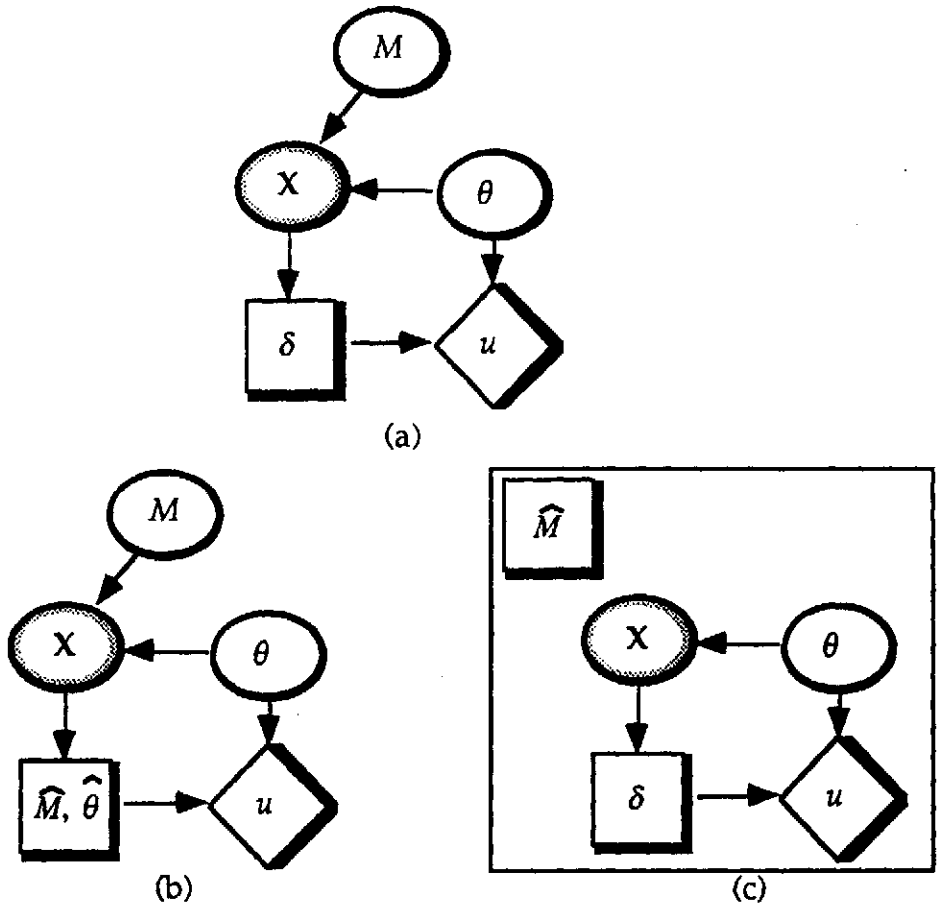


Figure 4.13: Bayesian model selection. (a) The general, proper Bayesian analysis requires the analyst to assess a prior belief in every possible probabilistic models (M) and in every possible parameter (θ). (b) Bayesian-style model selection involves the choice of a particular model ($\widehat{M}$) and a particular estimate for the parameters of that model ($\hat{\theta}$). The utility function determining that choice need not be related to a particular domain. (c) Alternatively, the analyst might select a model on domain grounds, and update belief in the parameters using a domain-dependent utility function.

4.5 The Bayesian Design Model

We can finally view the completed Bayesian design model for the literature problem. It contains the various probabilistic, statistical, and methodological concepts we have discussed in this chapter. The theoretical aspects have been discussed in previous sections; now let us review how the model applies to the metoprolol problem.

We reiterate the example from Section 1.2. Consider a 55-year-old white man who has just had a heart attack and who has been brought into the hospital almost immediately after symptoms began. Besides needing to stabilize his acute cardiovascular status, his physician wants to prevent further deterioration of his general cardiac condition. The doctor knows that metoprolol might improve his cardiac status. This drug has, however, serious known side effects. Should she administer the drug?

The *patient-utility* model (Sections 4.1 and 4.3.2) focuses on life expectancy: The drug associated with the longer life expectancy is preferred. The *outcome* of interest, therefore, is mortality. We make the modeling decision (Section 4.4) to use the patient's probability of death (after he has survived the hospitalization) as the *patient parameter* (Section 4.2.1), which we will assume to be constant over time (constant-hazard model). For the modeling decision of which referent *population* to use (Figure 4.6b), we have at least two choices on the basis of cardiological domain knowledge: middle-aged men and middle-aged adults. If we choose the population of combined sexes, there will be a larger number of studies, each with a large sample size, that we can bring to bear on this problem. This modeling decision exposes the tradeoff between the specificity of the data and the number of data available.[8]

The sample in the *study* (Figure 4.6b) consists of all heart-attack victims from south Sweden in the late 1970s. This characterization represents *selection* from our

[8]This tradeoff is more obvious for women patients, because there are so many fewer studies of women who have had heart attacks. The tradeoff is an example of a larger class of problems, called the *reference-class* problem (Kyburg, 1983).

population on ethnic grounds, but not on the basis of referral, diagnostic purity, or diagnostic-access biases (Sackett, 1979). The *protocol design* is reported to have been that of a double-masked, randomized, clinical trial. There is evidence to support the claim that the protocol was implemented as designed. For instance, the compositions of the metoprolol and placebo groups turned out to be similar with respect to relevant baseline characteristics, corroborating the implementation of randomization (see

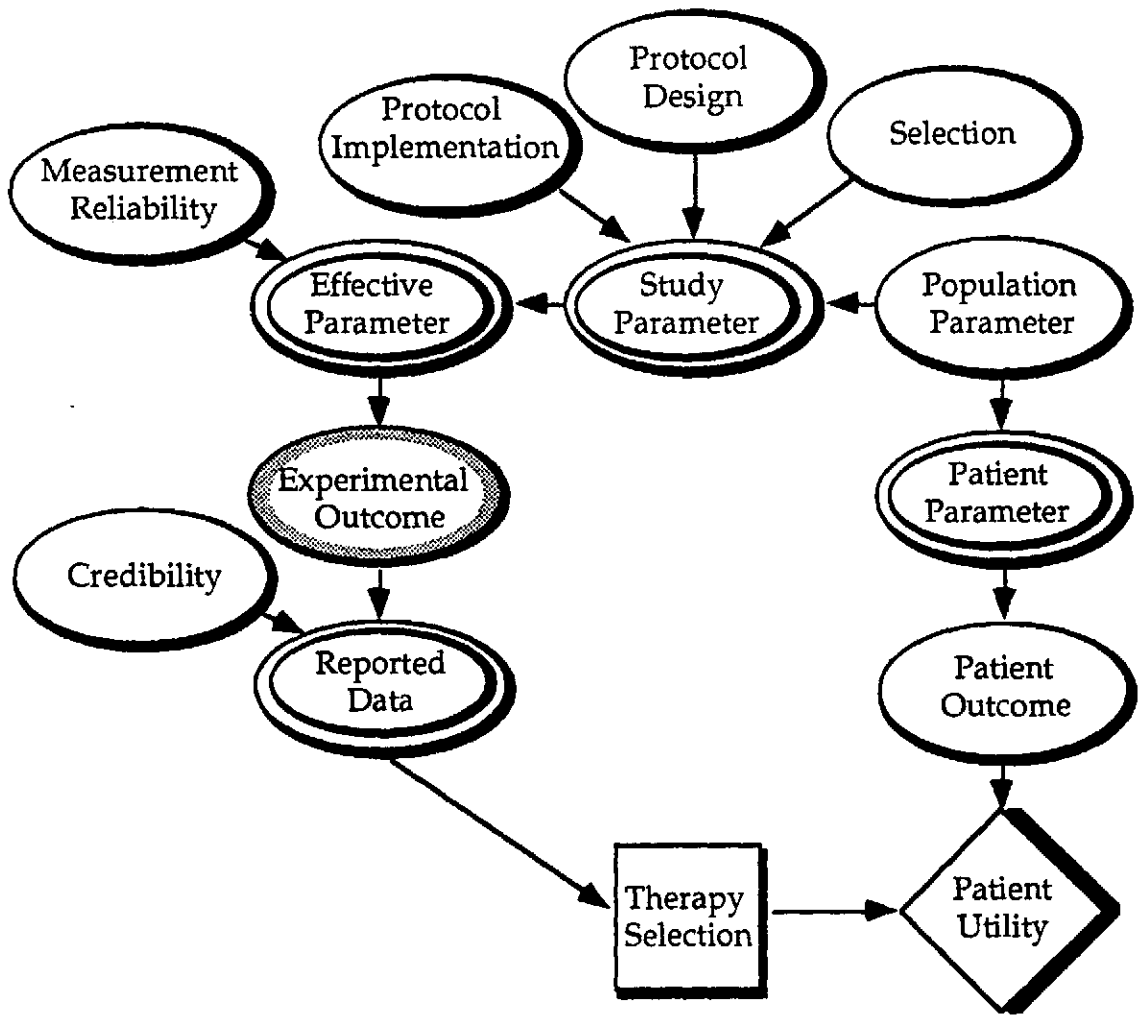


Figure 4.14: Bayesian design model. This framework takes into account the probabilistic, statistical, and methodological concepts discussed in this chapter.

Section 8.3.4 for further discussion about this concern). The numbers of withdrawals from the two groups on the basis of side effects are also similar, suggesting that, if there was some unmasking of care providers such that the treating physicians became aware of the true treatment assignments, the degree of unmasking was the same in both groups.

Protocol departures (Section 4.2.3) are evident in the study, because a number of patients did not receive the treatment to which they were assigned; they *withdrew* from the study. Estimating the actual degree of withdrawal explicitly is important for calculating posterior-probability distributions for the study parameters. This estimation adds bias parameters to be inferred—the probability of withdrawal from metoprolol and the probability of withdrawal from placebo—which, in turn, results in our considering a space of observational models larger than that we would be considering if we did not include the withdrawal bias. The withdrawal-bias parameter in this study models the fact that the study parameter for metoprolol was a result of mixing the treatment group with a third group of patients show received no treatment—that is, a group with the baseline mortality risk (the group of patients withdrawn). (We shall consider this methodological concern in more detail in Section 5.6.2.)

Continuing around Figure 4.14, *measurement reliability* (Section 4.2.3) for mortality studies depends on the sensitivity and specificity of the classification process. These new parameters are P(labeling patient as "dead" | patient is deceased) and P(labeling the patient as "alive" | patient is alive), respectively, both of which depend on patients who have dropped out of the study. The authors inform us that the mortality status of each patient entered into the study was assessed, regardless of subsequent treatment status. Finally, the credentials of the authors are such that we consider the study to have high *credibility*; therefore, we do not need to incorporate this extra debias.

Having specified the model, the physician can now assess her prior beliefs (Section 4.2.1), enter the relevant data, calculate the posterior-probability distributions (Section 4.3.1), assess the relative importance of side-effects losses to mortality gains, find which drug leads to the most utility (Section 4.3.2), and thus, arrive at the appropriate decision.

4.6 Critique of the Bayesian Approach

In this section, I shall discuss how the Bayesian approach satisfies the knowledge-level specifications of Section 2.3.

4.6.1 Intersubjectivity

The system should allow for differences of opinion among readers.

The capability of the Bayesian approach to represent divergent beliefs is the method's obvious strength. The rules of the Bayesian game are straightforward: If two people disagree about the conclusion of a study, then either they disagree about their prior belief, or they disagree about their analyses of the study, or both. If they disagree about their prior belief, then they can discuss the sources of their divergent prior belief. If they disagree about the analysis of the study, then they can dissect their disagreement in terms of the design model of Section 4.5. Bonduelle (1987) describes in detail the use of an influence-diagram–based framework for settling such disagreements.

The Bayesian approach may founder on its reliance on the subjectivity of the assessed prior beliefs. Physicians are not skilled at generating such assessments (Eddy, 1982). There are two strategies for overcoming this problem: to improve the interface and to educate the physician users. In terms of the interface, we might allow the physician to communicate qualitatively with a Bayesian system. Such an interaction

might involve the physician stating that she thinks with a "high" degree of certainty, that a mortality rate is "low." Alternatively, such a system might use a graphical depiction of prior- and posterior-belief curves as the language of interaction with the user. Discussion of educational approaches is beyond the scope of this dissertation.

4.6.2 Objectivity

The system should depend on objective, reproducible, and auditable methods.

The primary danger in the intersubjectivity of the Bayesian approach is that a dishonest reader may work backward from a desired posterior-probability distribution (with its policy implications) to a pdf that she could claim to be her prior belief. The primary defense against this abuse is that the Bayesian approach forces the physician to be explicit about that prior belief. The explicit nature of the representation acts as a tool to audit the physician's behavior. If the physician prefers to take a different approach for a different, but similar, patient, then the physician will have to justify explicitly the changing of her "prior" belief. Thus, the weakness of the approach, in being potentially subverted, can be turned into a strength, in its being objectively audited.

A second strength of the approach is that the probabilistic structure can generate clearly understood, numerical strengths of evidence.[9] A third strength is that multiple analyses may be done, if the analyst is honest (via her prior beliefs) in what she expected to find out before she performed the analysis. A fourth strength is that the *process* of making a decision based on the results of a study is well defined, can incorporate patients' explicit values, and does not require ad hoc heuristics; it is, in fact, objective, although the *content* is subjective.

[9]The strength of evidence for one hypothesis over another is the *likelihood ratio*, $\frac{\ell(\mathbf{X}|H_1)}{\ell(\mathbf{X}|H_2)}$.

4.6.3 Normativity

The system should implement methodologists' knowledge-level prescriptions.

I have shown, in Section 4.2.3, that concerns of methodologists—the desire to ensure high-quality and credible studies—can be represented in the Bayesian framework. It remains to be shown how the details are actually incorporated in an ongoing analysis (see Section 5.8).

The notion of normativity has a more specific meaning in the Bayesian context (Savage, 1972). The concept is usually taken to mean that an appropriate inferential method is one that leads to a conclusion consistent with prior belief and with the data; such a method is said to be *coherent*. Clearly, the Bayesian approach fulfills this criterion, as we saw with Equation 4.2. Equally clear is that any classical-statistical method that violates the Bayesian approach is *in*coherent.

4.6.4 Flexibility

The system should be able to evaluate both the pragmatic effectiveness and the ideal efficacy of tested therapy.

The decision-analytic framework provides a context for reinterpreting and understanding the dispute between analysts favoring efficacy-based interpretations, and those favoring effectiveness-based analyses.

The Bayesian reinterpretation is that the dispute turns on the decision to be made (Sackett and Gent, 1979). The analyst who cares about clinical decision making, under this interpretation, will perform an analysis based on effectiveness, whereas the analyst who concerns herself with discerning physical causality (and with deciding what study should next be performed) will perform an analysis based on efficacy. In the first case, the analyst will care about the effective parameters and their posterior-probability distributions. In the second case, the analyst will examine the posterior

beliefs in the population parameters. The population parameters can be used in the first case, as well, if the analyst has a patient-specific model for how a patient might violate protocol or might differ from the general population of patients. Note that, because the analyst may construct a single model to include both sets of parameters, neither analysis precludes the other.[10]

4.6.5 Adaptability

The system should enable the clinician to express methodological concerns without using statistical language.

Bayesian-formulated measures implement the semantics desired by physicians, such as degrees of belief. Yet, the complexity of a Bayesian analysis is such that it is unrealistic to expect physicians to employ the statistical knowledge necessary to execute the analysis. My conclusion from this potential conflict is that physicians require a semantic layer between them and the full power of a Bayesian statistical analysis. Such a layer permits the physician to express a possibly constrained set of concerns, and creates a Bayesian statistical model reflecting those concerns, protecting the physician from unnecessary statistical details. The structure and function of this layer are the subjects of Chapters 6 and 7, and are two contributions of this dissertation.

4.6.6 Simplicity

A simplified system should help the physician to interpret a single article that she has selected and read; should exclude explicit knowledge about particular statisticians and investigators; should assume that the physician user has the ability to express clearly

[10] If, however, the study has been *designed* with one type of analysis in mind, then the uncertainty brought into the analysis through considering methodological concerns that convert that analysis to one of the other types may overwhelmingly negate any certitude provided by the observed data themselves.

the particular problem at hand; and should support the decision making of a single physician, rather than that of an entire community.

The narrowed specifications can be accommodated in the Bayesian framework, although a computer-based implementation is needed; the development of the specifications for such an implementation is the subject of the next chapter. Note that subjective information, such as the reader's perception of the integrity of the reporting investigators, can be incorporated straightforwardly in the Bayesian framework.

4.7 Previous Systems

As with classical statistics, most of the computer-based systems have been built by statisticians for statisticians. Goel (1988) provides a useful reference list of statistical systems.

Investigators have built a number of graphically oriented systems that allow users to build knowledge-based systems with influence diagrams (Shachter, 1988c; Andersen et al., 1989; Chavez, 1991; Beinlich and Herskovits, 19890). Each of these systems assumes discrete variables in the networks, and is, therefore, not useful for statistical problems. Furthermore, these systems require that the user be conversant in the representational nature and power of the influence diagram, skills that are beyond most clinicians. I shall discuss, in Chapter 7, approaches that enable users to build influence diagrams without understanding the details of their contents.

My approach builds on the Confidence Profile Method (Eddy, 1989; Eddy et al., 1990; Shachter, 1990; Eddy et al., 1991), which enables analysts to combine the results of multiple studies into a coherent diagnostic or treatment policy. The technique, is based on Bayesian probability, and its most recent representation is based on influence diagrams. The method calls for the analyst to assemble an *evidence*

table of information from the different studies, noting the interventions used, the patients studied, the outcomes measured, and the biases involved. The analyst then may construct the influence diagram that integrates the evidence together, or may use adjustment formulae that produce estimates of the parameters of interest. The output of the approach consists of a set of posterior-belief curves; the analyst can infer much information from the relative position of the summary curve with respect to the curves representing the results of the different studies.

We shall find that the CPM and the method developed here supplement each other. The CPM is geared toward policy makers, and, therefore, often advises the use of noninformative prior beliefs. Here, we are targeting individual clinicians, and can, therefore, allow for prior beliefs that truly reflect domain knowledge. The CPM approach is generally used by specialist analysts (Trudeau, 1991), whereas my approach automates the construction of the relevant models to enable wider use of the CPM. We shall examine other differences in Section 7.8.

4.8 Summary

The Bayesian framework satisfies the knowledge-level criteria for solving the literature problem. It does so mainly by preserving many classical probabilistic, statistical, and methodological concepts, and by extending them to include subjective information. The gain is that the overall process becomes auditable and objective. The loss is the increased complexity and the demands on the user to provide just such explicit prior information. My thesis is that a computer-based environment can help to ease the complexity and to fulfill the demands, thereby enabling physicians to use the Bayesian framework to solve the literature problem. How such an environment provides this help is the subject of the next three chapters.

Chapter 5

THOMAS's Design Model

The framework of Chapter 4 can be applied to a large number of contexts in medicine. To show that the framework can be used for solving the literature problem—the point of this dissertation—I shall start by applying it to a limited domain. My working hypothesis is that understanding how to assess methodological concerns in this constrained context will help us to create environments where more complex concerns are involved. In this chapter, I shall discuss the system's design model, summarized in Figure 5.1; it is a specialization of the model shown in Figure 4.14. Our concern shall be to define the limited domain from both the user's and the statistician's points of view, so the domain is well defined either way.

In Section 5.1, I shall define the type of user expected by THOMAS; in Section 5.2, I shall discuss the type of domain to which THOMAS applies. The next two sections narrow the scope of the Bayesian computational context: Section 5.3 discusses THOMAS's utility model, and section 5.4 describes THOMAS's probabilistic model. The following three sections describe THOMAS's use of the three types of parameters introduced in section 4.2.3: Section 5.5 elaborates on the use of population parameters, Section 5.6 shows how THOMAS uses study parameters to deal with departures from protocol, and Section 5.7 describes the use of effective parameters in representing measurement

reliability. In Section 5.8, I discuss how all the models are used to capture notions of credibility. Finally, in Section 5.9, I shall summarize the model. In each section, I shall describe the design model, the limitations of that model, and the ways that the model might be extended.

5.1 Intended User

Because all physicians incorporate new policies into their practice, the literature problem should be universally important to physicians. Yet, not all physicians base their decisions on the clinical research literature (Williamson et al., 1989; Greer, 1988; Hill and Weisman, 1991); fewer still take the time necessary to examine critically the clinical research literature. The audience for a program such as THOMAS comprises these latter individuals. Although studies have documented that such physicians are not adept at statistical and quantitative thinking (Eddy, 1982), medical educators expect physicians to master the qualitative issues manifested in the clinical research literature (Sackett et al., 1991; Haynes et al., 1986). Thus, we shall define our user community as physicians who have an interest in the clinical research literature, and who have a familiarity with the basic concepts of research design and methodology.

This group of intended users could be expanded to include medical students and medical-journal editors. However, the system would have to be modified to take into account educational goals, in the first case, and publishing needs, in the second.

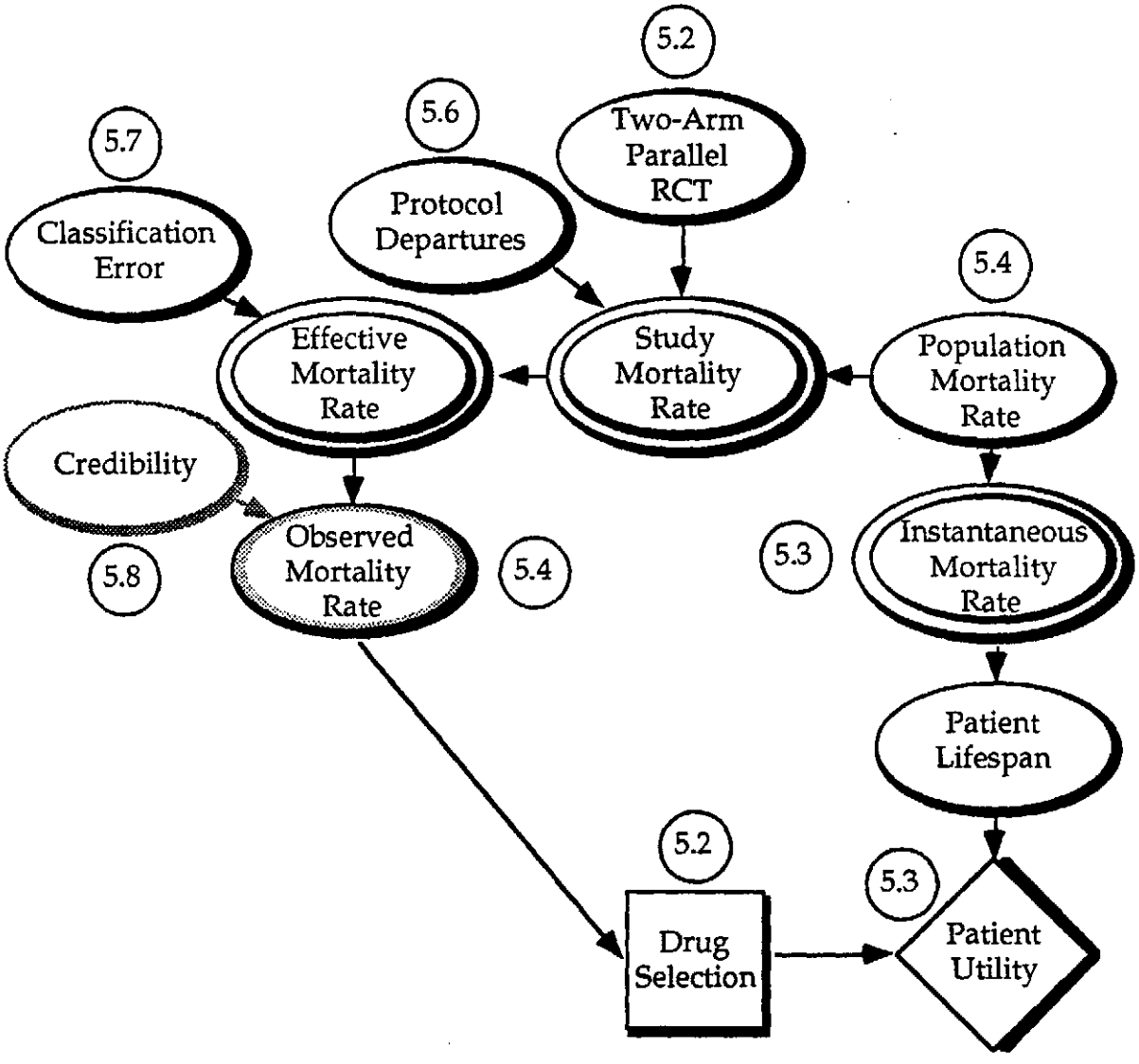


Figure 5.1: THOMAS's design model. This framework is a specialization of that shown in Figure 4.14. Encircled numbers refer to section numbers in this chapter where the indicated entity or relationship is discussed. *Credibility* is only implicit in the model, and, therefore, is represented by a ghost node.

5.2 Restricted Domain

THOMAS operates in two domains: a methodological domain, and a medical domain. THOMAS's *methodological* domain is restricted to a single class of study designs: the *two-arm parallel RCT*, a schematic of which was shown in Figure 1.1. In such studies, patients are assigned to only one of two possible interventions: the control or the experimental. I chose this design because biostatisticians deem it least vulnerable to bias, and academic physicians accept it as the gold-standard method for determining the ideal biological efficacy or the pragmatic effectiveness of therapy. Furthermore, other types of designs can be modeled as modifications of this core design (Eddy et al., 1991).

The *medical* domain consists of patients, diseases, practitioners, therapies, and outcomes. THOMAS can consider any type of disease, class of patients, or practitioner; it does not, however, *reason* about them, as such considerations constitute issues of external validity. The *class of therapy* will be limited to *drugs*. The reason for this restriction is that drug therapy is the type of intervention most amenable to gold-standard testing: The control drug can be made physically almost identical to the experimental drug, maximizing the degree of *masking* possible in the study. *Mortality* is the *outcome* to which physicians and patients pay the greatest attention—we will do so as well. Our concern will be with the lifespan of a patient, symbolically denoted L. As we shall see in Section 5.3, morbidity concerns will also be considered, but not as explicitly is does lifespan.

These restrictions leave THOMAS applicable to 0.3 percent of the medical literature, as indexed by the National Library of Medicine (Meinert et al., 1984). Although this proportion is small, it represents the published data that the academic community finds most influential, as we infer from the recurrent appeal to the results of RCTs or to the demand that proof be offered through the execution of such a study. Although

the statistical components of the model could be extended to other outcomes and therapies, our approach is limited mostly by the system's utility model.

5.3 Utility Model

If a study were to report that the experimental drug leads to a lower mortality rate than does the control drug, then, as clinician readers, we still might not prescribe the experimental drug. There are two reasons for not taking at face value the predicted life expectancy. First is the issue of *uncertainty:* The experimental drug may have led to a better outcome in this study, but it might fare less well in some future comparison. This issue is not relevant in a normative decision-making context, because in choices based on formal decision analysis, only the means of the decision maker's posterior beliefs matters (Howard, 1988).[1] Second is the issue of *treatment cost:* The "better" treatment may have associated with it side effects and financial costs that are not worth even a substantial increase in life expectancy. Here, the issue is one of *preferences,* where mortality gains from the experimental drug, in specific cases, may be perceived as overwhelmed by morbidity losses.

Preferences are captured in *utility models,* which involve balancing potentially conflicting objectives (Klein et al., 1990). A problem in constructing utility models is that of the comensuration of disparate qualities: mixing apples and oranges. For instance, how does a reader balance the objective of minimizing mortality, with the objective of minimizing morbidity when the drug with the lower mortality increases morbidity? One solution is to measure the less important attribute on the same scale as the more important entity. THOMAS takes this approach, using the scale of *life*

[1]Posterior uncertainty is relevant in deciding what *study* to perform next. In such a decision, the investigator should pursue the variable in the decision model with the greatest uncertainty, posterior to the study at hand, where narrowing that uncertainty would result in the greatest increase in utility. This approach is called *control-decision making on the basis of expected value of information* (Howard, 1983).

years.

THOMAS asks the user (see Figure 5.2): How much gain in life span does the patient believe is necessary to balance the treatment morbidity and, therefore, to make taking the drug worthwhile? I call this difference the *pragmatic difference.* If the difference, based on the posterior beliefs, between the life expectancy associated with the experimental drug and that associated with the control drug is less than the pragmatic difference, then the patient should not take the experimental drug. Ideally, THOMAS would use a value that takes into account patient-specific factors (e.g., based on actuarial data for age, sex, and disease, as in the DEALE utility model (Beck et al., 1982)); currently, the system uses the life expectancy calculated only from the analysis of the study at hand.

The utility function that THOMAS uses is

$$u(\delta, L) = L - I_{(\delta=\text{exp})}\Delta, \tag{5.1}$$

where $u(\cdot)$ is utility, δ is drug choice, $\delta = exp$ denotes the experimental drug, L is lifespan, $I_{(\cdot)}$ is the indicator function, and Δ is the pragmatic difference. The posterior utility $\langle u(\delta, L)\rangle = \left\langle L - I_{(\delta=\text{exp})}\Delta \right\rangle = \langle L \rangle - I_{(\delta=\text{exp})}\Delta$; only the life expectancy must be calculated for any given case on the basis of prior knowledge and of the data in the study.

This preference model underlies the set of arcs in Figure 5.1 between the drug-selection and patient-utility nodes and between the patient-lifespan and patient-utility nodes.

An alternative preference model might use quality-adjusted life years (QALYs), which THOMAS would assess by asking the user the relative utility of a year survived without disease while taking the medication to a year of survival with disease without the drug treatment. THOMAS does not, at present, use this model. Adding this extension, however, requires no conceptual changes beyond offering the user the choice

of preference models. Most users would make this choice implicitly by using the utility scale (life years or QALYs) with which they are more comfortable.

The restricted preference model does prevent THOMAS from addressing objectives—or outcomes—beyond mortality. We might think that Equation 5.1 could be used for any time-dependent outcome, replacing lifespan (time to death) with a different time-to-episode variable (e.g., time-to-next-stroke, in a study examining whether an experimental medication reduces the incidence of strokes (Canadian Cooperative Study Group, 1978)). Any such preference model, however, would have to include mortality, as well, in the decision. Thus, the single-objective model of Equation 5.1 would require adaptation for the multi-objective case.

5.4 Probabilistic Models

There are two probabilistic models in THOMAS: one for lifespan, the other for the observed data. In keeping with the strategy of Bayesian model selection that we have chosen (see Section 4.4), THOMAS conditions its analysis on particular probabilistic models.

For lifespan, the system uses the exponential model

$$L \sim \mathcal{E}(\lambda), \tag{5.2}$$

where λ is the instantaneous mortality rate (see page 54 for the symbols used for parametric probability models). This model makes the strong assumption that the instantaneous mortality rate is constant throughout present and future life. The model is often used by biostatisticians for inferences regarding short- and medium-term survival, and serves as the baseline parametric model in survival studies (Miller, 1981) and in mortality-based decision models (Beck et al., 1982).

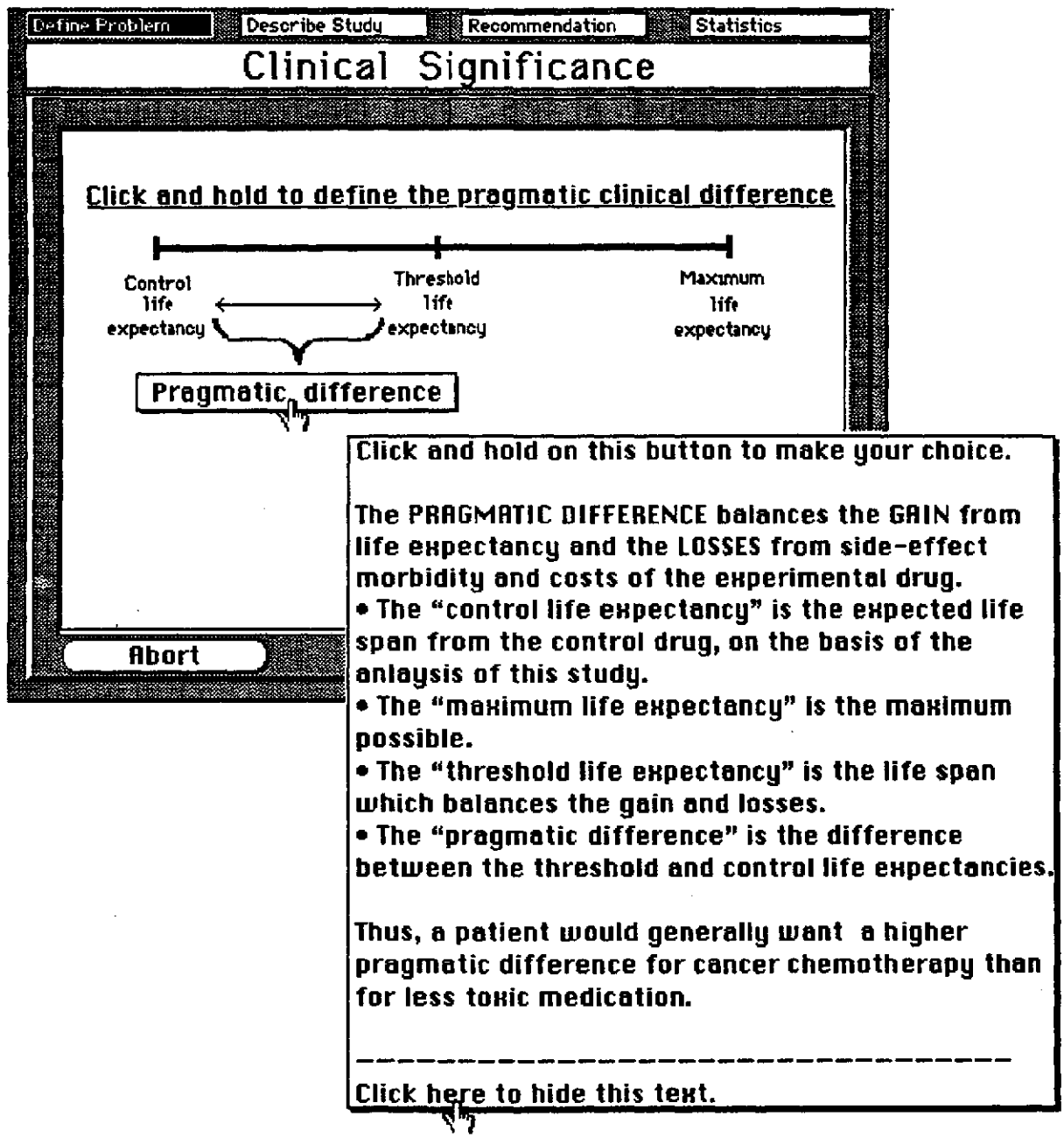


Figure 5.2: The pragmatic difference. (a) This screen shows how THOMAS requests the user's pragmatic difference for the problem at hand. When the button *Pragmatic Difference* is clicked and held, a submenu of possible times is presented; the user can type in a time not listed. (b) This screen inset shows the help text THOMAS supplies to explain the terms used.

A consequence of this choice of probabilistic model is that the life expectancy for an individual patient can be calculated in terms of the parameter:

$$\left\langle L_{\mathrm{Rx}}^{\mathrm{pat}} \right\rangle = \frac{1}{\lambda_{\mathrm{Rx}}^{\mathrm{pat}}}, \tag{5.3}$$

where the superscript *pat* refers to the individual patient, and the subscript *Rx* refers to either of the experimental or the control treatments.

This model underlies the arc in Figure 5.1 between the instantaneous-mortality-rate and patient-mortality-rate nodes.

For the second probabilistic model, discrete observed data are dependent on a parameter in a binomial model, where the parameter is the mortality rate over a specified period of time:

$$x^{\mathrm{obs}} \sim \mathcal{BI}\left(\theta_{\Delta t}\right), \tag{5.4}$$

where x^{obs} is the observed mortality rate (using x to denote data; see Figure 4.8), $\theta_{\Delta t}$ is the mortality rate over a specified time interval, and Δt is that time period, which is also called the *study period* or *observation duration.*

Because lifespan has a particular probability distribution, we can derive exactly the timed mortality rate from the instantaneous mortality rate. Due to our assumption that $L \sim \mathcal{E}(\lambda)$, $P(L \geq \Delta t) = e^{-\lambda \Delta t}$. Now, the timed mortality rate is defined as the probability of surviving just up to time Δt: $\theta_{\Delta t} \equiv P(L \leq \Delta t) = 1 - e^{-\lambda \Delta t}$. Solving for λ gives

$$\lambda = \frac{1}{\Delta t} \ln \frac{1}{1 - \theta_{\Delta t}}. \tag{5.5}$$

Thus, the instantaneous-rate parameter, λ, can be made deterministically dependent on the time-rate parameter, $\theta_{\Delta t}$. This function defines the deterministic relationship between the population-mortality-rate node and instantaneous-mortality-rate node in Figure 5.1.

If we wished to extend the approach taken here and to enable the physician user

to modify these two probabilistic models, the system would be required to have capabilities it does not now possess. First, it would have to determine the appropriate probabilistic model through data-intensive methods (see, for instance Linhart and Zucchini (1986)) such as model-fitting or bootstrapping. Second, it would need access to the entire data set of the study. Third, it would need to apply—and to possess—much more domain knowledge (e.g., that the Weibull distribution is more appropriate in pediatric oncology). On the user side of the interaction, the user would have to be more statistically sophisticated than we assumed she was in Section 5.1. Systems with the first two capabilities have been built by research in AI and statistics (Gale, 1986b; Oldford and Peters, 1988). Although such a system can be modeled within the Bayesian approach I am presenting in the dissertation, two important questions remain: How should prior uncertainty about the propriety of a statistical model be represented within the system? How should the nature of such uncertainty be explained to and assessed from the physician user? There are some solutions to the first question (Geisser and Eddy, 1979; Draper and Guttman, 1986; Clayton et al., 1986), although the prior belief is based on nondomain issues, such as the number of parameters in the model. The second question has not been posed, to my knowledge. These remain fundamental unanswered questions, at present, within Bayesian statistics.

Having defined the various models THOMAS uses, we turn our attention to the parameters within those models, and to the relationships among them.

5.5 Population Parameters

A population parameter characterizes belief in a parameter governing the likelihood of an outcome for a member of a population of patients who have a disorder in common. As we change our focus from the individual patient to that of the population

of patients who have the same condition, we change our focus from the patient's *instantaneous* mortality rate to the population's *timed* mortality rate—for instance, the mortality rate in the year following a myocardial infarction. I shall denote such a mortality rate by $\theta^{\text{POP}}_{\Delta t}$, or simply θ^{POP}.

THOMAS assumes that the individual patient's timed mortality rate θ^{Pt} is the same as the population timed mortality rate θ^{POP}, which is why instantaneous-mortality-rate node in Figure 5.1 is dependent on the population-mortality-rate node. If the user had a belief that the patient and population timed mortality rates were different, she would have to specify a model relating the two parameters, which THOMAS does not currently support.

5.6 Study Parameters

A study parameter governs the likelihood of outcomes of patients enrolled in the study. THOMAS achieves model-construction flexibility by having the system's allow the user to construct different relationships between the population and study parameters. With respect to the Bayesian design model of Figure 4.14, THOMAS does not implement selection models, but, instead, conditions its analysis on a fixed design. Hence, it is the variety of *protocol departures* represented in THOMAS that allows the user to represent important methodological concerns. More than a simple convenience, this availability permits us to build the system on the work of other researchers and to avoid constructing models that might be criticized as subjective.

The mathematical forms of the models THOMAS uses are found in the textbook by Eddy, et al. (1991). In general, these models involve debiasing the primary likelihood of the observed mortality rate given in the study report, as suggested in Section 4.2.3. We shall assume, as recommended by Peto, et al. (1976), that this rate is reported in terms of patients' *assignments* to therapy, regardless of the treatment actually

received by patients in the course of the study. I shall call the group of patients assigned to a therapy the *cohort* of patients assigned to a therapy; the *study mortality rate of patients assigned to therapy* refers to this study group. (If the mortality rates reported are more specific, we will incorporate those data more directly; see Section 6.3.2).

In general, the *observed* mortality rate is dependent on the study mortality rate for the study group. We write $x_{\mathrm{Rx}}^{\mathrm{obs}} \sim \mathcal{BI}\left(\theta_{\mathrm{Rx}}^{\mathrm{study}}\right)$, where Rx refers either to the control or to the experimental treatment. For each protocol departure, we shall model the study-group parameter as a function of population parameters governing component subgroups of the study group itself. We now consider four protocol departures: crossover, withdrawal, noncomplicance, and loss to followup.

5.6.1 Crossover

Patients who *cross over* from one therapy to another are patients who were assigned to one therapy, but in fact received the other. I shall use the term *crossovers* for the patients who switched therapies on the direction of the investigators.[2] Crossing over usually results from side effects of medications.

Assume that a particular set of patients crossed over from the experimental to the control group. Then, the observed mortality rate for the experimental group as a whole reflects a *mixture* of effects: Some of the patients in the experimental group received the control medication, and the remainder received the experimental drug. The degree of mixture depends on how many patients crossed over. The model describes the mixing as it relates to the study parameter for the study group of patients assigned to the experimental treatment. The qualitative relationships among the parameters are shown in Figure 5.3.

[2]This bias should not be confused with a crossover *design*, where patients are purposely treated with both drugs in the course of a study.

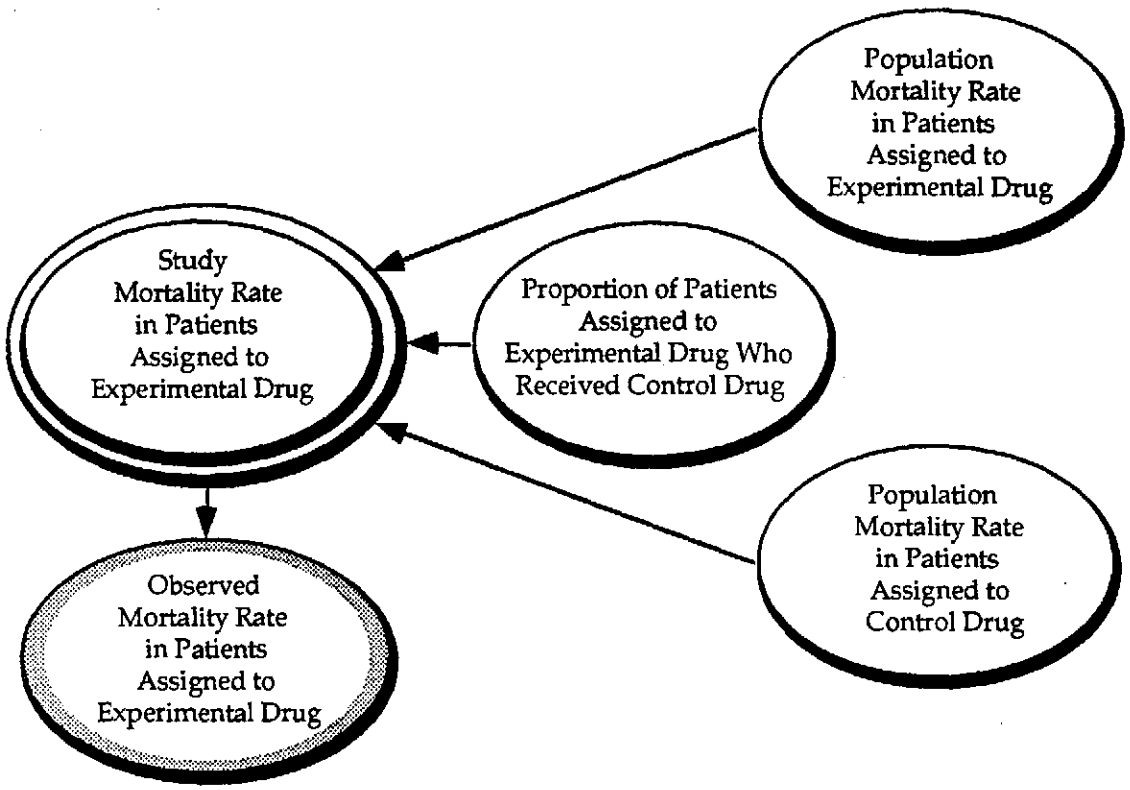


Figure 5.3: THOMAS's crossover model. The study mortality rate in patients assigned to therapy is a mixture of two population parameters, determined by one methodological parameter.

The mathematical model for crossovers is

$$\theta_{\text{exp}}^{\text{study}} = \theta_{\text{exp}}^{\text{pop}} \cdot \alpha_{\overline{\text{exp}\rightarrow\text{ctl}}} + \theta_{\text{ctl}}^{\text{pop}} \cdot \alpha_{\text{exp}\rightarrow\text{ctl}}, \tag{5.6}$$

where $\theta_{\text{exp}}^{\text{study}}$ is the study mortality rate in patients who were assigned to the experimental group; $\theta_{\text{exp}}^{\text{pop}}$ is the population mortality rate in patients assigned to, and who received, the experimental treatment; $\theta_{\text{ctl}}^{\text{pop}}$ is the mortality rate in patients who were assigned to, and who received, the control treatment; $\alpha_{\text{exp}\rightarrow\text{ctl}}$ is the study crossover rate of patients assigned to the experimental drug who received the control drug; and $\alpha_{\overline{\text{exp}\rightarrow\text{ctl}}}$ is the proportion of patients who did *not* crossover; $\alpha_{\text{exp}\rightarrow\text{ctl}} + \alpha_{\overline{\text{exp}\rightarrow\text{ctl}}} = 1$. The subscripts *exp* and *ctl* are interchanged in the model for crossovers from the control to the experimental therapy.

The two crossover-rate parameters ($\alpha_{\text{exp}\to\text{ctl}}$ and $\alpha_{\overline{\text{exp}\to\text{ctl}}}$) are examples of *methodological* parameters, because their values define the likelihood-debiased model (see Section 4.2.3). The values of methodological parameters give a sense of the *quality* of a study. Note that, if $\alpha_{\text{exp}\to\text{ctl}}$ is high (say, close to 0.50), then we would probably think that the investigators were too sloppy in executing the study, whereas if the paper reports a very low value of $\alpha_{\text{exp}\to\text{ctl}}$ (say, 0.001), we would probably conclude that the authors suppressed information from the report. In each case, we are implicitly comparing the reported value with a subjective, internal, prior value. The Bayesian approach asks us to make that value explicit.

There are two assumptions implicit in Equation 5.6. First is the assumption that patients who crossed over did so immediately. Second is the assumption that the mortality rate of patients who crossed over was identical to that of patients who were assigned to, and who received, the crossed-to medication.

5.6.2 Withdrawal

A patient who *withdraws* from a study is one who does not receive the assigned therapy and whose outcome status is known to the study investigators. These patients may have experienced a severe setback from the disease; the study investigators, rather than switch the patient to a different arm, may have decided to let the attending physician care for the patient off the study protocol. Thus, the investigators continue following such patients, and know, at the end of the study, whether the patients survived the study period.

Thus, again, the observed mortality rate is dependent on a study parameter that represents a mixture of mortality rates (see Figure 5.4), as in the crossover model. In this case, however, the admixed rate is the mortality rate of patients who withdrew from the study, which is the mortality rate of patients who have the disease but are not enrolled in the study; this rate is the baseline mortality rate of the patients who

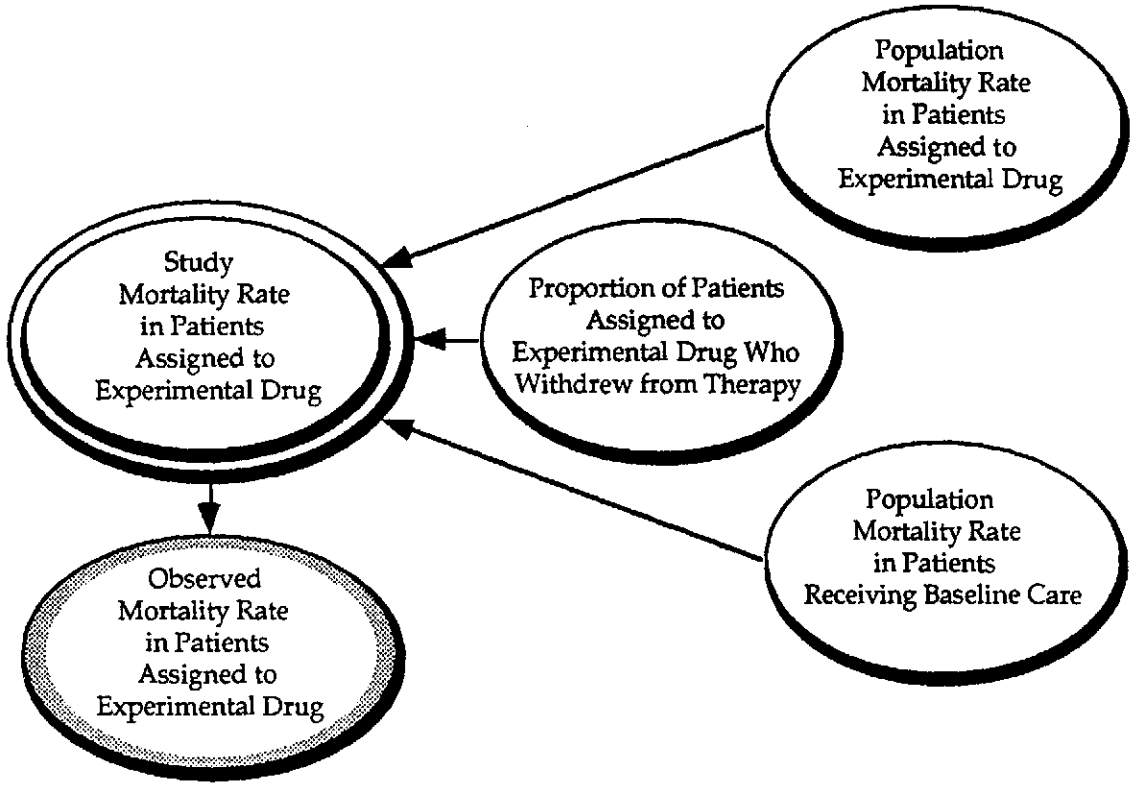


Figure 5.4: THOMAS's withdrawal model. The study mortality rate in patients assigned to therapy is a mixture of two population parameters, determined by one methodological parameter.

withdrew, namely, the mortality rate of patients exposed to standard, or *baseline,* care. If we were assessing a prior belief in this mortality rate, we might choose a rate higher than the mortality rate for the general patient in the study, because patients who withdrew might be sicker than those who did not. Or, we might assume that the rate is the same as that of patients assigned to the control group, if placebo were the control therapy.

The model is

$$\theta_{\text{exp}}^{\text{study}} = \theta_{\text{exp}}^{\text{pop}} \cdot \alpha_{\text{exp, }\overline{\text{withd}}} + \theta_{\text{baseline}}^{\text{pop}} \cdot \alpha_{\text{exp, withd}}, \tag{5.7}$$

where $\theta_{\text{baseline}}^{\text{pop}}$ is the population mortality rate in patients who receive baseline care, $\alpha_{\text{exp, withd}}$ is the study withdrawal rate of patients assigned to the experimental drug,

and $\alpha_{\text{exp},\overline{\text{withd}}}$ is the proportion of patients assigned to the experimental drug who did *not* withdraw; $\alpha_{\text{exp,withd}} + \alpha_{\text{exp},\overline{\text{withd}}} = 1$. For withdrawals from controls, the model is the same, except that we substitute the subscript *ctl* for *exp*; $\theta^{\text{pop}}_{\text{baseline}}$ is still needed. Note that, in either case, assessing a prior belief in $\theta^{\text{pop}}_{\text{baseline}}$ gives the user the chance to express her domain medical knowledge, because such knowledge is necessary in evaluating whether patients who withdraw from a particular therapy are different from those who do not, with respect to factors prognostic for the outcome of interest.

5.6.3 Noncompliance

Patients who do *not comply* with therapy are patients who receive varying exposure to their assigned medication. The study-group mortality rate is a mixture of a potentially complex array of mortality rates. Available models (Lakatos, 1986; Eddy et al., 1991) present the treatment received as a diluted version of the treatment to which a patient was assigned. The degree of dilution introduces another methodological parameter, τ, the proportion of the time noncompliant patients were, in fact, compliant.

The model is

$$\theta^{\text{study}}_{\text{exp}} = \theta^{\text{pop}}_{\text{exp}} \cdot \alpha_{\text{exp},\overline{\text{nc}}} + \left(\theta^{\text{pop}}_{\text{exp}} \cdot \tau_{\text{nc}} + \theta^{\text{pop}}_{\text{baseline}} \cdot \overline{\tau}_{\text{nc}}\right) \cdot \alpha_{\text{exp,nc}} \tag{5.8}$$

where τ_{nc} is the proportion of the time that study patients who were noncompliant were initially compliant with therapy, $\overline{\tau}_{\text{nc}}$ is the proportion of the time they were *non*compliant, and $\tau_{\text{nc}} + \overline{\tau}_{\text{nc}} = 1$. Furthermore, $\alpha_{\text{exp,nc}}$ is the noncompliance rate of study patients assigned to the experimental drug, $\alpha_{\text{exp},\overline{\text{nc}}}$ is the proportion of experimental-group patients who were *not* noncompliant, and $\alpha_{\text{exp,nc}} + \alpha_{\text{exp},\overline{\text{nc}}} = 1$. For noncompliance from the control therapy, the subscript *exp* is changed to *ctl*. These model enables the reader to represent the different types of compliance expected under different treatments; Freedman (1990) considers a number of such models.

Note that this model is an implementation of the directives given by Chalmers and colleagues (1981), quoted on page 45.

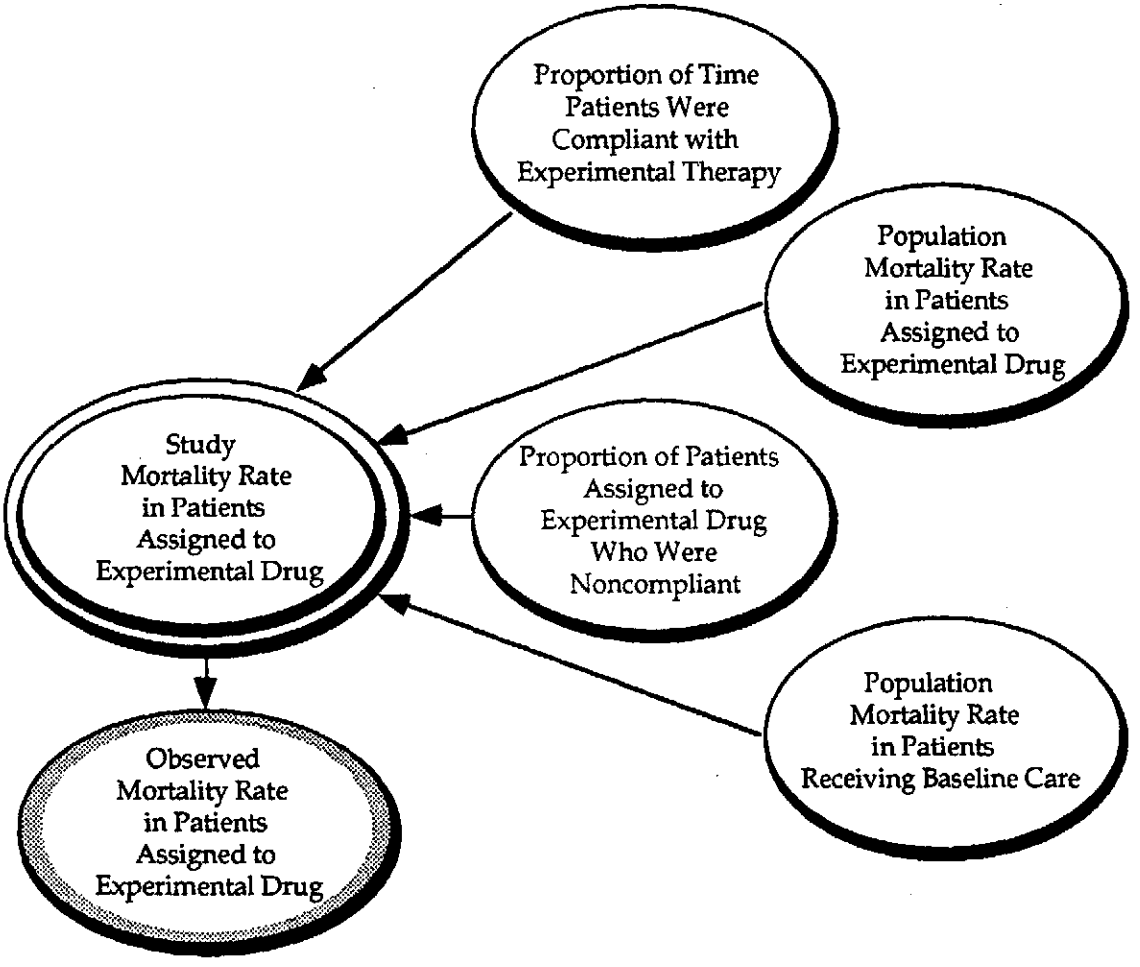


Figure 5.5: THOMAS's noncompliance model. The study mortality rate in patients assigned to therapy is a mixture of two population parameters, requiring two methodological parameters.

5.6.4 Loss to Followup

Patients who are *lost to followup* are patients whose outcome status is not known to the study investigators. The observed mortality rate, therefore, reflects the contribution of the study parameters only of those patients who have remained in the

study (see Figure 5.6). Thus, the study-group mortality rate is a mixture of component rates, as is the withdrawal model, but the evidence is available from only one component.

The model is similar to the withdrawal model, as well:

$$\theta_{\text{exp}}^{\text{study}} = \theta_{\text{exp},\overline{\text{ltfu}}}^{\text{study}} \cdot \alpha_{\text{exp},\overline{\text{ltfu}}} + \theta_{\text{exp,ltfu}}^{\text{study}} \cdot \alpha_{\text{exp,ltfu}}, \tag{5.9}$$

where $\theta_{\text{exp}}^{\text{study}}$ is the study-group mortality rate in patients assigned to the experimental treatment, $\theta_{\text{exp},\overline{\text{ltfu}}}^{\text{study}}$ is the study mortality rate in patients who were *not* lost to followup, $\theta_{\text{exp,ltfu}}^{\text{study}}$ is the study mortality rate in patients who *were* lost to followup, $\alpha_{\text{exp,ltfu}}$ is the proportion of patients who *were* lost to follow up, $\alpha_{\text{exp},\overline{\text{ltfu}}}$ is the proportion of patients *not* lost to followup; $\alpha_{\text{exp,ltfu}} + \alpha_{\text{exp},\overline{\text{ltfu}}} = 1$. By nature of the lack of information about patients lost to followup, any evidence for $\theta_{\text{exp}}^{\text{study}}$ is modeled as dependent on $\theta_{\text{exp},\overline{\text{ltfu}}}^{\text{study}}$: $x_{\text{exp}}^{\text{obs}} \sim \mathcal{BI}\left(\theta_{\text{exp},\overline{\text{ltfu}}}^{\text{study}}\right)$. For losses to followup from the control group, the subscript *exp* is changed to *ctl*.

Note that these models introduce the notion of splitting the study parameter itself into component study parameters, a strategy we will use to implement each of these protocol-departure models; see Section 7.4.1.

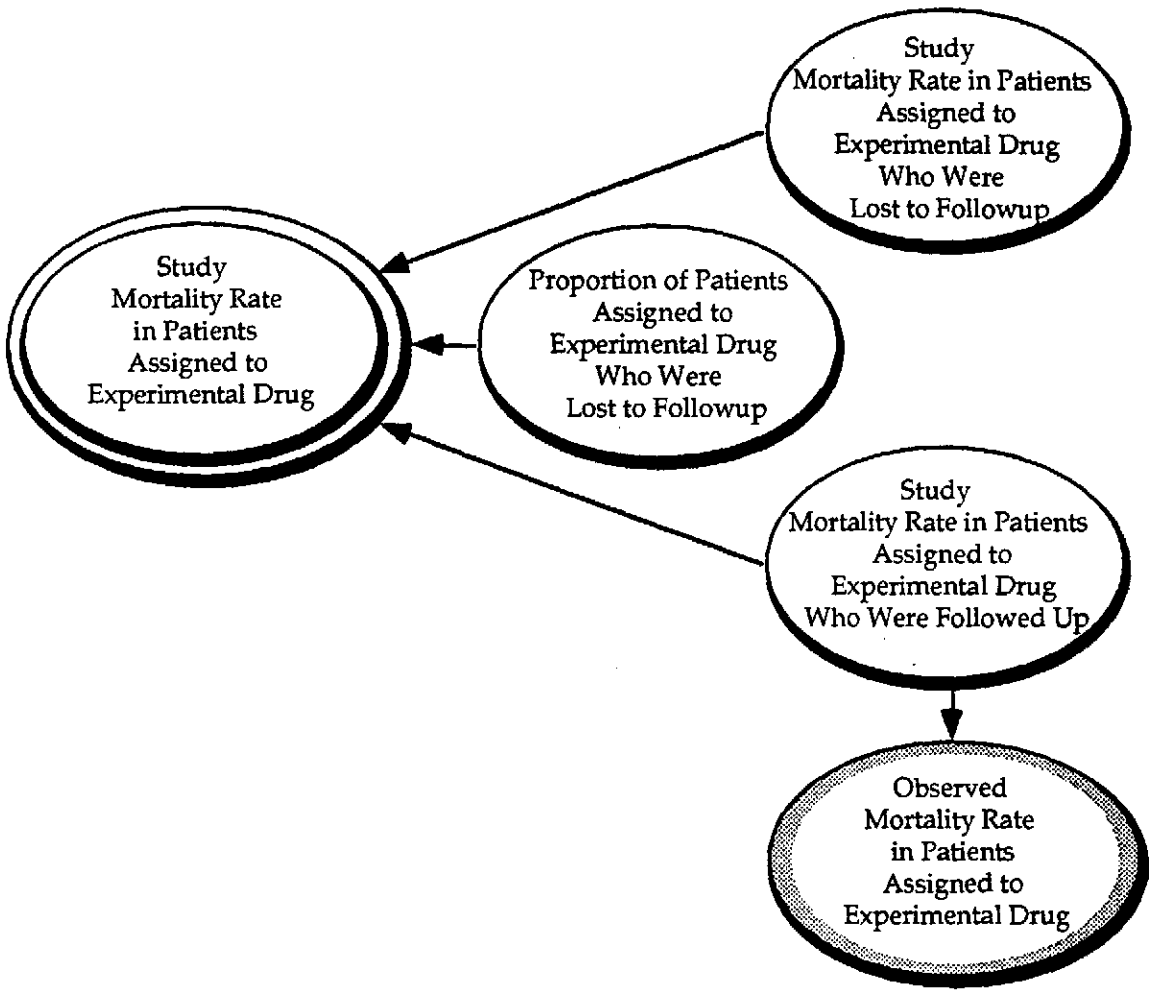


Figure 5.6: THOMAS's loss-to-followup model. The study mortality rate in patients assigned to therapy is a mixture of two other study parameters, determined by one methodological parameter. The observed mortality rate is dependent on the component parameter governing the likelihood of death in patients who were not lost to followup.

5.7 Effective Parameters

An effective parameter governs the likelihoods of the observations made in the study. If the reliability of the measurements of the outcomes made in a study is in question, THOMAS represents the reliability in a model that relates study parameters to effective parameters; observed data become dependent on the effective parameter: $x_{\mathrm{Rx}}^{\mathrm{obs}} \sim \mathcal{BI}\left(\theta_{\mathrm{Rx}}^{\mathrm{effective}}\right)$. The form of the model for measurement reliability depends on the type of the outcome measured. For binary outcomes, such as mortality status, measurement reliability is properly called *classification error,* and is expressed in a *calibration* model in terms of *sensitivity* and *specificity* (see Figure 5.7).

The formula for the effective parameter is

$$\theta_{\mathrm{exp}}^{\mathrm{effective}} = \theta_{\mathrm{exp}}^{\mathrm{study}} \cdot se + \left(1 - \theta_{\mathrm{exp}}^{\mathrm{study}}\right) \cdot (1 - sp)\,, \tag{5.10}$$

where se is the sensitivity for determining mortality status in patients who are dead, and sp is the specificity for determining mortality status in patients who are alive.

5.8 Credibility

A formal model of credibility, as suggested by Figure 4.14, would describe the reported data as an unreliable report of the actual, observed data. For instance, the system would ask the user for the probability that the investigators would report the data they did, if the data were as reported and if they were not, $P(\text{reported } x \mid \text{any } x)$; it would also ask for the probability that the investigators would have reported other data, if the data were as reported and if they were not, $P(\text{other } x \mid \text{any } x)$. These assessments would give the system sensitivities and specificities for the reported data, modeling credibility as an issue of measurement reliability. These assessments are difficult and are not appropriate for this first-phase system.

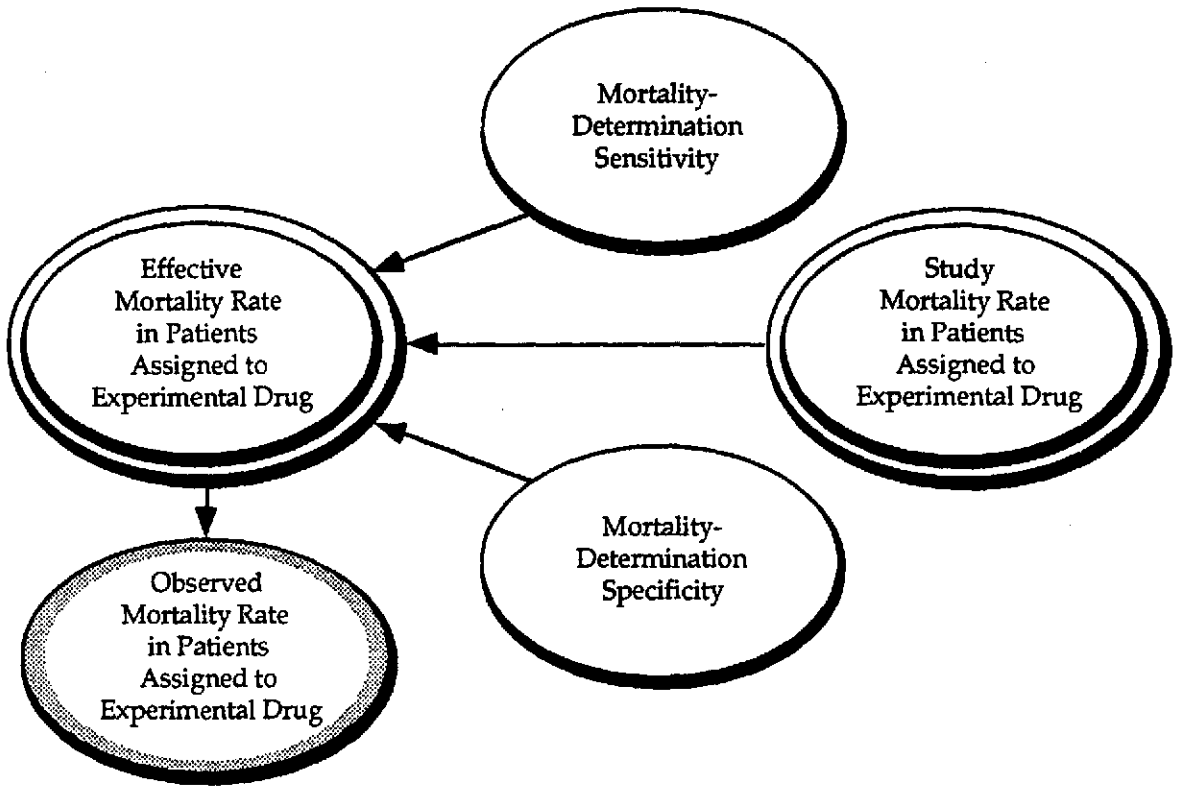


Figure 5.7: THOMAS's classification-error model. The parameter of effective experimental-drug-mortality rate is a function of the study mortality rate, and of the sensitivity and specificity of the mortality-determination procedure.

However, credibility as measurement unreliability is not the only way to interpret the believability of a study report. THOMAS allows the appraisal of believability by assessing prior beliefs in methodological parameters, and through allowing the user to choose which protocol departures she thinks are relevant. Thus, notions of credibility are *distributed* throughout THOMAS's constructed statistical models. This distribution is in contrast to the usual strategy of *unifying* estimates of credibility through the derivation of a single number. This unifying approach is taken by many meta-analysts (Sacks et al., 1987; Reisch et al., 1989), who assess the *quality* of a study by summing checklist scores. This approach is also taken by expert-system builders,

who have used belief networks to assemble a single number (Lehmann, 1988) or have used value theory to compute such a unifying result (Klein et al., 1990). Furthermore, there is no meaningful interpretation of the final number. In each case, there is no justification for multiplying the quality weight (or the credibility probability or the credibility value) by a test score derived from a study report. By contrast, the strategy of distributing at least some aspects of credibility throughout the statistical model results in a process that remains normatively valid.

5.9 Summary

THOMAS specializes the general Bayesian design model as follows:

- User: physician with basic knowledge of methodology and design
- Domain: two-arm parallel RCT comparing two drugs, with the outcome of mortality
- Utility model: threshold model scaled in life years
- Probabilistic models: exponential for lifespan, binomial for observed mortality
- Conjugate models: beta distributions for all rate parameters
- Population parameters: equal to patient parameters
- Study parameters: determined by population parameters through individual models for crossovers, withdrawals, noncompliants, and losses to followup
- Effective parameters: determined by study parameters through a classification-error model for binary outcomes
- Credibility: distributed throughout a statistical model

In the next two chapters, we shall see how THOMAS delivers this design model to the user.

Chapter 6

A Bayesian Interface for the Literature Problem

THOMAS's implementation of its design model is based on Figure 1.6. In this chapter, we shall focus on the components of interaction between the user and the system. The novel component of this computer-based environment is a semantic layer that protects the user from the technical Bayesian framework. I divide the discussion of that layer into two parts. I shall describe the user interface in this chapter in moderate detail, because otherwise it would be difficult to describe what types of information are needed by the system and how the system obtains that information from the user. Similarly, the interface illustrates what types of information are offered by the system and how the system presents that information to the user. The second part—the machine's representation and management of the semantic layer—is presented in the chapter 7.

In Section 6.1, I shall discuss the general principles of building a user interface that reflects a Bayesian context. In Section 6.2, I shall present the control of the interaction between the machine and the user. In Section 6.3, I shall describe how the user communicates her knowledge to the machine: knowledge about clinical significance,

information about the structure of the study, prior knowledge about parameters in the model, and the evidence from the study. I shall also introduce the patient-flow diagram data structure. Finally, Section 6.4 shows how the machine reports back to the user the posterior probabilities, the decision recommendation, the sensitivity of the analysis to different prior beliefs, and the implications of different analyses.

6.1 Interface Principles

The are two basic principles that guide the construction of a user interface within THOMAS's design model. The first is that the interface must use visual metaphors familiar to the intended user. The second is that the interface must be reflect the Bayesian paradigm.

6.1.1 Visual Metaphors

The principle that an interface must use visual metaphors familiar to the intended user applies to any graphical computer environment. System designers believe that following the principle provides greater intuitiveness to the interaction. Yet, the degree of intuition depends on the user for whom for the system is intended; poets would not be expected to find spreadsheets second nature, and physicians cannot be expected to manipulate the parametric statistical models needed to solve the literature problem.

We can, however, expect that the physician user of THOMAS will be familiar with visual metaphors from the research literature itself. A number of such metaphors are available. One graphical device, found in the meta-analytic literature, is the *methodology checklist* (Reisch et al., 1989), where meta-analysts record their assessments of the methodological merit of a study. Another device, used by research authors, is

the *patient-flow diagram* (Hjalmarson et al., 1981), where the investigators communicate the fate of groups of patients during the study. A third device is the *evidence table* (Eddy et al., 1991), where analysts present data about different studies (see Section 4.7). I shall describe my use of these metaphors in Sections 6.2, 6.3.2, and 6.3.4, respectively.

6.1.2 Consistency with the Bayesian Paradigm

The second interface principle derives from our use of a Bayesian design model. The Bayesian interface paradigm is that, *before examining evidence, you must assess relevant prior beliefs.* The paradigm applies to the process of furnishing evidence pertaining to entities of interest, and of constructing the statistical model. The interface issue is how best to obtain the information necessary, in the proper sequence.

In THOMAS, the entities for which evidence is provided are the parameters in the decision model (Figure 4.14). The challenge in applying the Bayesian paradigm in assessing prior belief in, and obtaining evidence for, these parameters arises from the first principle: We want the system to ask for information regarding a parameter without presuming that the user understands the entire statistical infrastructure implied by the request. My primary solution to this problem is for the machine to label parameters with names that the physician user will find intuitive. How the system names parameters will be discussed in Section 7.4.1.

As an example, before the system can accept evidence regarding the mortality of patients exposed to the experimental drug, it must assess from the user her belief about the parameter corresponding to that mortality (the *study mortality rate in patients assigned to the experimental drug*), or about the components of that parameter, in the case of protocol departures (e.g., the *population mortality rate in patients assigned to the experimental drug* and the *population mortality rate in patients assigned to baseline care*). The very notion of a group of patients assigned to a particular drug

implies what Feinstein (1985) has called the underlying *architecture* to the research study. The name of the parameter suggests that parameter's context to the physician user, whose clinical-epidemiological fund of knowledge includes (see Section 5.1) the terms *population, study, observed,* and *assignment.*

The Bayesian paradigm also applies to the process of *constructing* the statistical model pertaining to a study. When a physician indicates that a methodological issue is of concern in analyzing a study, that new concern induces a modification in the statistical model under construction, as was described in Sections 5.6 and 5.7. At times, the modification leads to the creation of new parameters, such as the methodological parameter that describes the degree of crossover protocol departures (see Section 5.6.1). In keeping with the principle, the system first must determine the identity of any new parameters, then must assess the user's prior beliefs in these parameters, and finally must allow the user to provide the evidence. This sequence is a bit different from what physicians often do now: Observe the data, then assess relevant adjustments. My primary solution for implementing the Bayesian sequence is to make constant the order of system requests after the user has informed the machine about the methodological concerns. How the system identifies the relevant new parameters will be discussed in Section 6.3.3.

6.2 Input Sequence

There are three basic subtasks to the task of a consultation for solving the literature problem: formulate the problem, instantiate it, and view the results. In any expert system, there are two strategies for obtaining information from the user: data-driven, where the user directs control of the interchange, implicitly creating a model for the problem, and model-driven (or goal-driven), where the system seeks information to fill an extant model. In constructing THOMAS, I have combined these strategies. On

the one hand, the overall sequence is fixed. The primary reason for such restriction is that, although I assume the user to be knowledgeable about methodological concerns, I cannot assume the user to be aware of all the interactions among those concerns. By forcing the user to advance through a prescribed sequence, I ensure that she does not ignore important interactions (such as, for instance, the dependence of the permitted methodological concerns on the study design). On the other hand, in the subtask of describing the study, I have given the user greater freedom, as I shall show in Section 6.3.2.

The overall sequence is presented to the user as a recursive set of steps to be taken. Thus, for instance, the top-level step, *Consultation,* (see Figure 6.1) comprises five subtasks: define the clinical problem, describe the study, view the results, examine the statistics, and finish the consultation. The problem-definition step, in turn, comprises two subtasks: define the drugs involved and define the meaning of clinical significance. The explanatory semantics inherent in the relationship between one level and the next is that the first level suggests the *why* of the second level, and the second level provides the *how* of the first, to use language familiar to builders of rule-based expert systems (Buchanan and Shortliffe, 1984).

The visual interface for this overall sequence employs the *checklist* metaphor mentioned in Section 6.1.1: The user checks off attributes that apply to the problem at hand (see Figure 6.1). The input graphics vary (See Figure 6.2), depending on which type of selection is required: single choice (out of a fixed set, out of a modifiable set, or in sequence) or multiple choice.

The checklist is the visual manifestation of an implicit *dependency,* or AND-OR, tree. Each screen represents a node in this tree. There are AND nodes, which require that *every* child be satisfied (generally, in an ordered sequence (Figure 6.2c)), and OR nodes, which require that *any* child be satisfied. OR nodes, in turn, may allow for multiple choices (Figure 6.2d), or only single choices (Figures 6.2a and 6.2b).

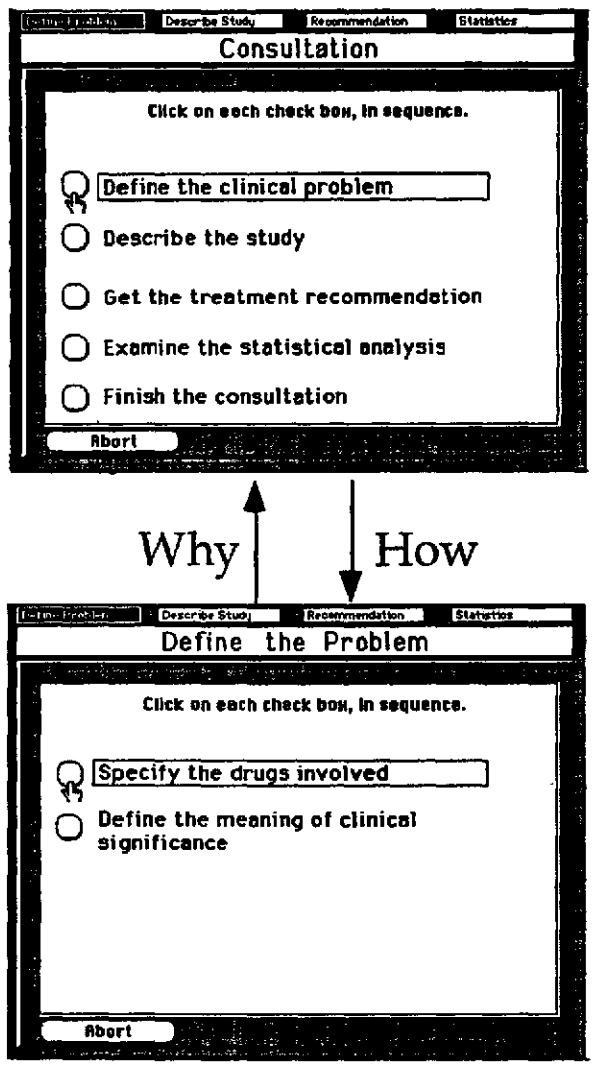


Figure 6.1: Checklist metaphor. The tasks to be performed are listed. As one task is checked off by the user, any subtasks relevent to the task are displayed. This sequence generates a dependency tree for the interface, where the semantics of movement to a lower level (toward the leaves of the tree) is that of *how* a higher level is satisfied, and where the semantics of movement to a higher level (toward the root of the tree) is that of *why* a lower level is requested.

Internal nodes of the tree are, generally, sequential, single-choice AND nodes. Progress through the tree may proceed only when a node and its relevant children have been satisfied. *Satisfaction,* here, means selection. Once a node is satisfied, the machine executes all forward-chaining action dependent on that node. The content of the dependency tree is platform-independent, but its graphic appearance is not. The dependency tree for THOMAS is listed in Table 6.1.

6.3 Input of Content

THOMAS needs four types of technical information: the decision model, the statistical model, the user's prior beliefs in relevant parameters, and the evidence from the study. We shall discuss each in turn.

6.3.1 Decision Model

THOMAS constrains the user's definition of the decision model by assuming the canonical model discussed in Section 5.3. In asking for the pragmatic threshold, the key number in this model (see Figure 5.2), the machine points out that this threshold may be patient- or drug-specific .

6.3.2 Statistical Model

So that it can assess the statistical model from the user, the system requires the user to take several actions. As discussed in Section 5.4, THOMAS currently assumes a single probabilistic model (see Figure 4.13); as shown in Figure 5.1, THOMAS assumes a single protocol design, as well. THOMAS *does* allow flexibility in letting the user establish the protocol implementation and the measurement reliability of the study.

Table 6.1: THOMAS's dependency tree.

Consultation	
I	Define the problem[1]
I.A	Specify the drugs
I.A.1	Specify the experimental drug
I.A.2	Specify the control drug
I.B	Define the clinical significance
I.B.1	Define the pragmatic difference
II	Describe the study
II.A.1	Select the study
II.A.2	Examine the study
II.A.2.a	Label the analysis
I.D.1.a	Choose a label
II.A.2.b	Analyze the study
II.A.2.b.i	Specify the design
II.A.2.b.i.1	Specify the architecture
II.A.2.b.i.1.a	Choose the two-arm RCT
II.A.2.b.i.2	Specify the outcome
II.A.2.b.i.2.a	Choose the mortality outcome
II.A.2.b.ii	Specify the observation duration
II.A.2.b.ii.1	Enter the time
II.A.2.b.ii.2	Enter the units
II.A.2.b.iii	Describe the study execution
III	Get the treatment recommendation
III.A	Specify the use of statistics[2]
III.A.1	Choose effectiveness
III.A.2	Choose efficacy
III.B	Examine the decision
IV	Examine the statistical analysis[2]
IV.A	Compare prior/posterior probabilities
IV.B	Examine probabilities for a parameter
IV.C	Perform a sensitivity analysis
IV.C.1	Alter a prior belief
IV.D	Answer a statistical question
IV.D.1	Choose a question
V	Finish the consultation[2]
V.A	Choose a new user
V.B	Choose the same user, new problem
V.C	Choose the same problem, new analysis
V.D	Quit from THOMAS

[1] Unless otherwise indicated, this and all nodes are AND nodes.

[2] This is an OR node.

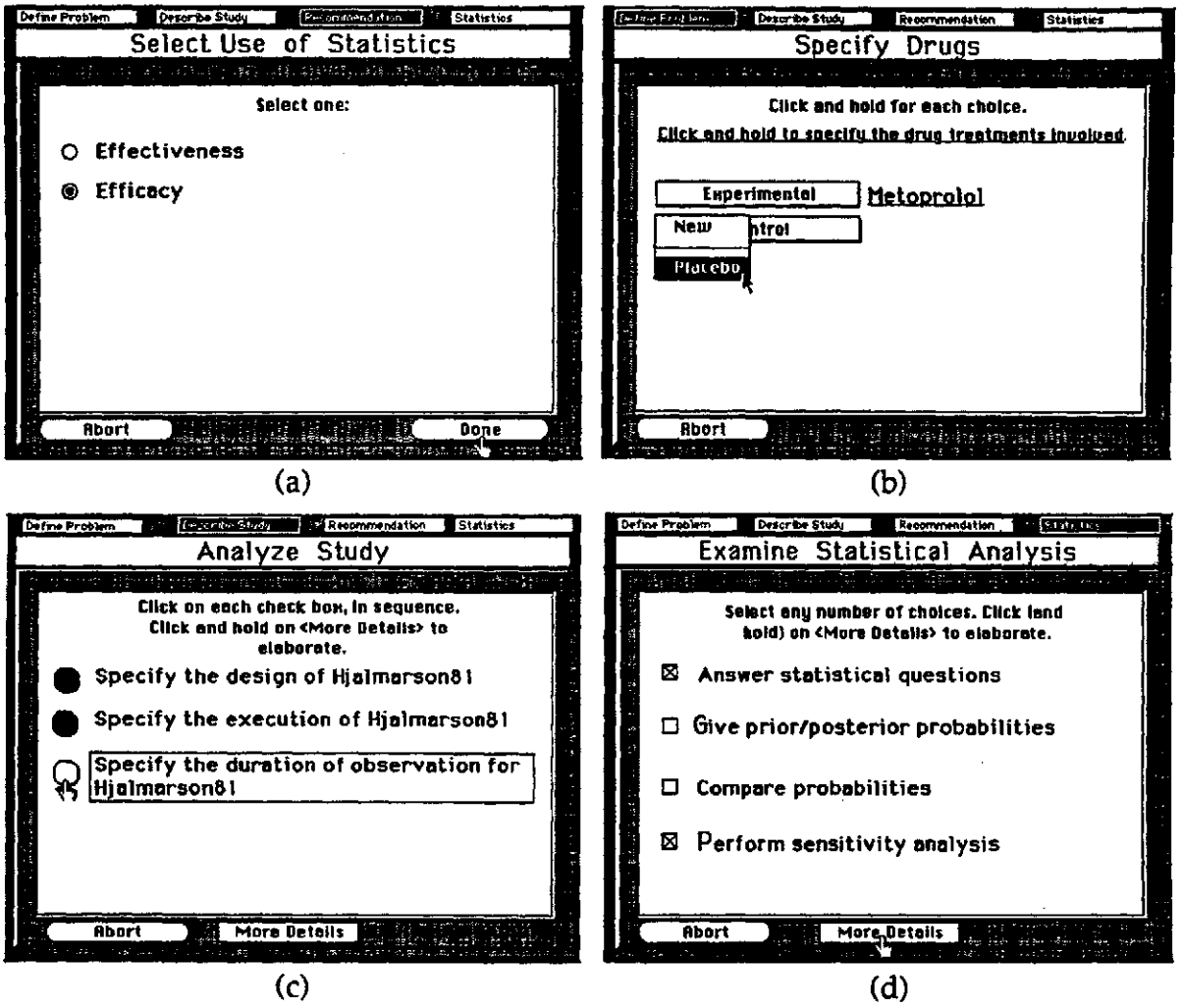


Figure 6.2: Checklist formats. THOMAS uses four interface styles to communicate the type of response needed to satisfy a subtask. (a) Single choice from a fixed set. Selection of one choice automatically deselects any other choice, if one has already been made. (b) Single choice out of a modifiable set. Selection of the choice *New* for the control treatment prompts the system to request the name of the new choice. (c) Single choice in sequence. The system darkens the box corresponding to a choice previously taken; the rectangle directs the user to the next choice to be made. (d) Multiple choice. To proceed with subtasks for the selections, the user must explicitly request *More Details*, at the bottom of the screen.

Although the *protocol design* and the *outcome* of the study (both components of the study architecture (see page 147)) are fixed by THOMAS as the two-arm parallel RCT and mortality, respectively, THOMAS still asks the user to select these choices explicitly, as a way of communicating to the user the "thought" process used by the system, and as a way of marking where the system should be expanded in the future.

In giving input about the *protocol implementation,* the user has the most flexibility, constrained only by the metaphor of the patient-flow diagram (see Figure 1.1 and Section 6.1.1). The diagram employs the cohort as its unit of construction. *A* **cohort** *is a group of patients who share a history. A* **patient-flow diagram** *is a rooted tree of cohorts.* A parent cohort in the tree comprises the patients in each of its children cohorts, and a child cohort comprises patients in the parental cohort whose fate is the same as that of the parental patients, but is different from that of its sibling cohort.[1] The root cohort comprises patients admitted to the study.[2]

For each cohort, the system displays the following information (see Figure 6.3): its functional name, the total number of patients, and the number of patients in the group who experienced the outcome.

The physician uses the patient-flow diagram to pinpoint which methodological issues are of concern. By clicking and holding the mouse pointing device on the name line of a cohort, the user can view a menu of possible protocol departures allowed in that cohort (Figure 6.3). By holding the device on the *Number Died* line, she can view a menu of possible options relating to measurement reliability.

[1]Parental cohorts, in THOMAS, when split, are divided into only two parts.

[2]If patient selection were to be represented in the program, then the root cohort would comprise patients who *potentially* could have been admitted to the study.

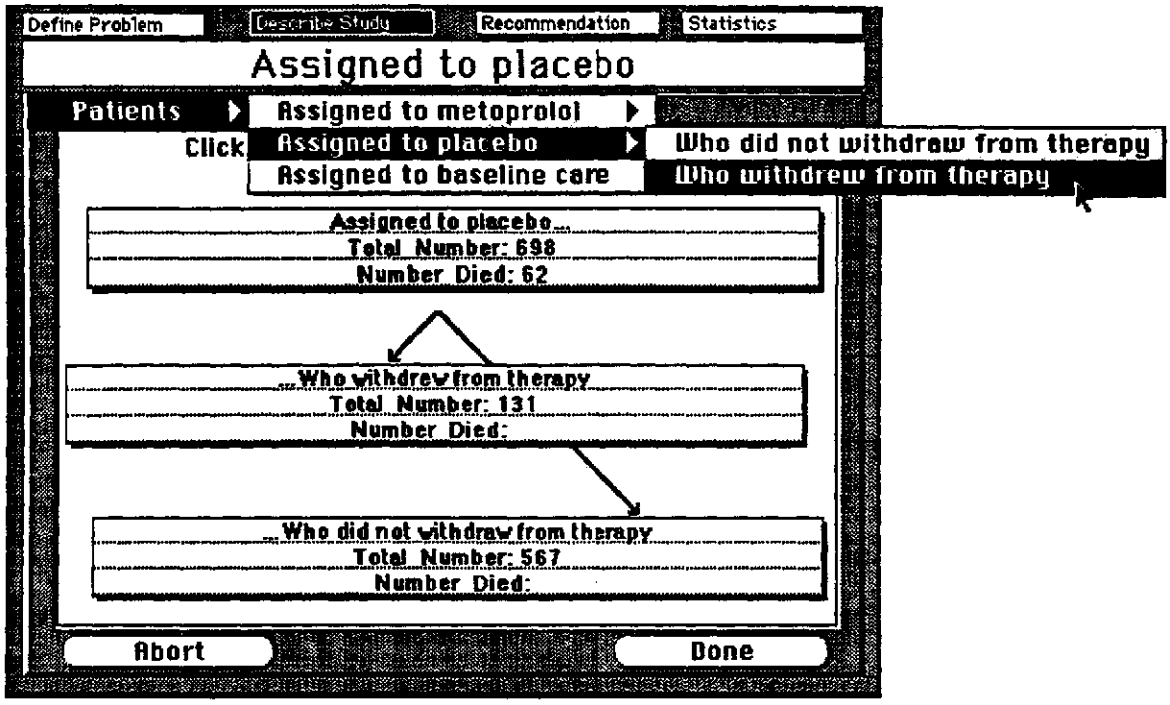


Figure 6.3: The patient-flow diagram. This figure shows part of the patient-flow diagram for the metoprolol example. The three cohorts shown in three boxes refer to the group of patients assigned to placebo, and its components: patients who withdrew from assigned therapy, and patients who did not withdraw from therapy. The box corresponding to each cohort shows the part of the name of the cohort that identifies the group of patients (e.g., the cohort *Patients assigned to placebo who withdrew from therapy* is identified by the name fragment ... *who withdrew from therapy*). The box also has slots for the total number of patients in the cohort and for the number of patients who experienced the outcome of interest. If a number is not known, the slot is left blank. The slot values are entered by the user as evidence (e.g., 131), or are calculated by the machine (e.g., $567 = 698 - 131$).

In addition to viewing the cohorts on this screen, the user can get a synopsis of the entire patient-flow diagram, as shown by the hierarchical graphic at the top of the screen, showing the name of the cohort *Patients assigned to placebo who withdrew from therapy* as a highlighted path through the diagram. (The cohort *Patients assigned to baseline care* is a dummy cohort that allows the system to implement the different protocol-departure models.)

The number of cohorts that can be realistically displayed at a single time is artificially constrained by the version of HyperCard used for this work. More recent versions, with variable-size windows, would usually allow display of the entire patient-flow diagram at one time.

The user adds protocol departures to the statistical model by selecting the departures that took place in the most specific cohort: The system presents a menu of possible departures specific to each cohort (see Figure 6.4). The result of selecting a protocol departure is the creation of two new cohorts: components of the original cohort comprising patients who did—and who did not—experience the protocol departure. (These subcohorts are shown in Figure 6.3.) If the creation of the new cohorts engenders new methodological parameters, then, before allowing the user to enter any data, the system assesses prior opinions about these new parameters. This sequence is in keeping with the Bayesian paradigm, as described in Section 6.1.2.

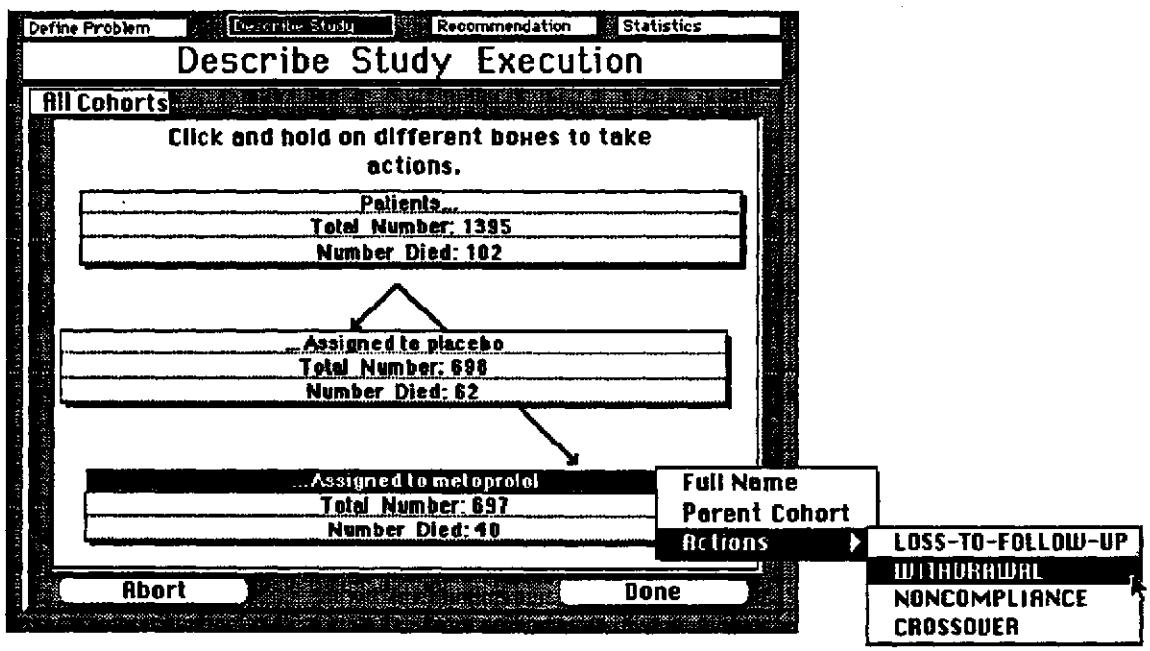


Figure 6.4: Communicating methodological concerns. The system displays potential methodological concerns appropriate to the cohort of *Patients assigned to metoprolol*; in this case, there are four possible concerns. Other options are simply to view the full name of the cohort, and to view the parent cohort of which the present cohort is a component.

6.3.3 Prior Beliefs

The system needs the physician's prior belief in two classes of parameters (see Section 7.3.1): outcome parameters and methodological parameters. Different types of knowledge are communicated by specifying prior beliefs in different types of parameters. *Domain knowledge* is communicated by the user giving prior knowledge about the control and the experimental mortality rates (e.g., how many myocardial infarction patients survive three months). *Study knowledge* is communicated by the user stating that two parameters are equivalent (e.g., the mortality rate in patients who receive baseline care and the mortality rate in patients who receive placebo). *Methodological knowledge* is communicated by the user's belief in methodological parameters (e.g., that 10 percent of patients in a well-done cardiology study may be expected to withdraw from the study). The user's assessment of methodological knowledge implicitly gives the system knowledge about specific authors and about credibility.

THOMAS allows two types of assessment about any of these parameters: numerical and qualitative. In each case, the system translates the user's input into a prior distribution over a parameter, assuming the parameter's uncertainty to have a beta distribution, with parameters α and β.

The style of the *numeric* assessments are based on types of estimates about proportions with which physicians are familiar. One style is the raw *proportion,* where a proportion is assessed as a numerator divided by a denominator (see Figure 6.5). Thus, a physician's prior belief about a mortality rate may be 2 patients having died out of 10 previously observed. This belief is less certain than having observed 20 patients out of 100. The translation into a beta distribution is straightforward; the machine sets α to the reported numerator, and β to the difference between the denominator and the numerator.

The second style of assessing prior belief in a proportion is as a mean with a standard deviation. For instance, the user may report that she believes the mortality

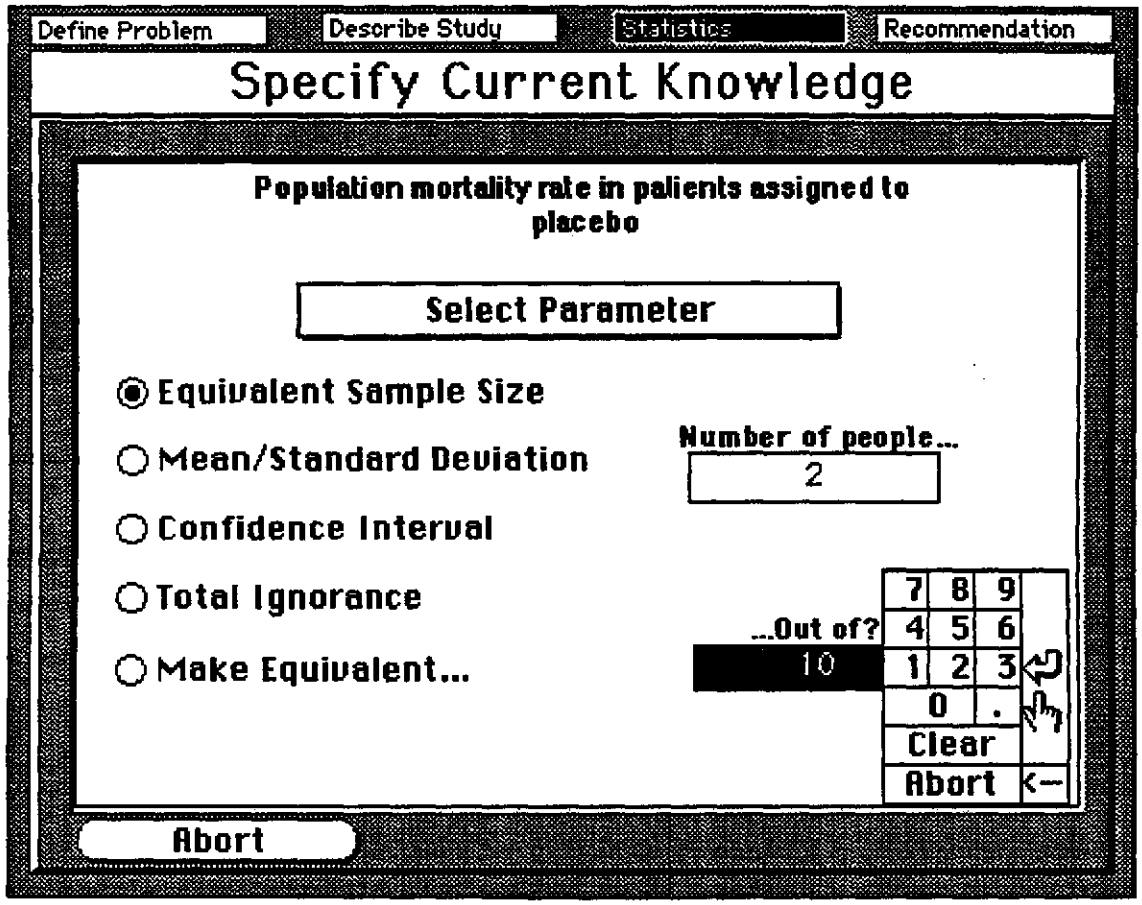


Figure 6.5: Using sample size to assess a proportion.

rate to be 0.2 ± 0.1. The machine computes the beta-distribution parameters from the formulas for mean and variance for the beta distribution, given in Equation 4.1.

The third style is through the user's reporting an interval with an attached level of confidence. The system arrives at the beta-distribution parameters by fitting a normal distribution to the interval, computing the mean and variance of that distribution, and fitting beta-distribution parameters to the calculated mean and variance, again, using the relationships on page 88. For instance, the user may report her belief in a mortality rate to be between 0.1 and 0.3, with 70 percent confidence. The mean would be 0.2, the variance would be nearly 0.1; the resulting α and β are 0.12 and

0.48 (a very uncertain prior).

THOMAS takes two *qualitative* approaches. The first allows the user to stipulate that she has no prior knowledge about the parameter in question. This declaration is called *claiming total ignorance,* and leads to the Bayesian equivalent of performing a classical statistical analysis. The machine creates a prior distribution for the parameter that is $\mathcal{BE}(\frac{1}{2}, \frac{1}{2})$. Many Bayesian statisticians recommend this distribution for a number of reasons, one being that this is the distribution that satisfies a number of conditions: invariance, data-translation, Kullback–Leibler divergence, and maximum entropy (Bernardo, 1979).

The second qualitative approach allows the user to specify that two parameters are *equivalent,* or are the same. In this case, the prior and posterior beliefs for the two parameters are exactly the same, and any evidence for one parameter updates belief in the other. This approach allows the user to communicate such domain knowledge as, for instance, that the mortality rate in patients given baseline care is the same as that of patients treated with placebo—that placebo confers no further benefit to patients. It also allows the user to communicate methodological knowledge; for instance, she can say that the sensitivity of death certificates for detecting death in patients treated with placebo is the same as that in patients treated with the experimental drug.

6.3.4 Evidence

Philosophers of science (Hanson, 1961) have made the scientific community cognizant of the *theory-ladenness of facts:* There are no facts (evidence) without a context within which to place them. A central mission of THOMAS's interface is to provide that context without the user having to understand the formal basis or structure—the statistical model. The patient-flow diagram again performs the functions of focusing a user's attention and of implicitly defining the context. Because study data are

reported in terms of study cohorts, and because the names of the cohorts have semantics clear to the physician, the user should be able to give the system the appropriate data. Figure 6.6 shows such an interaction.

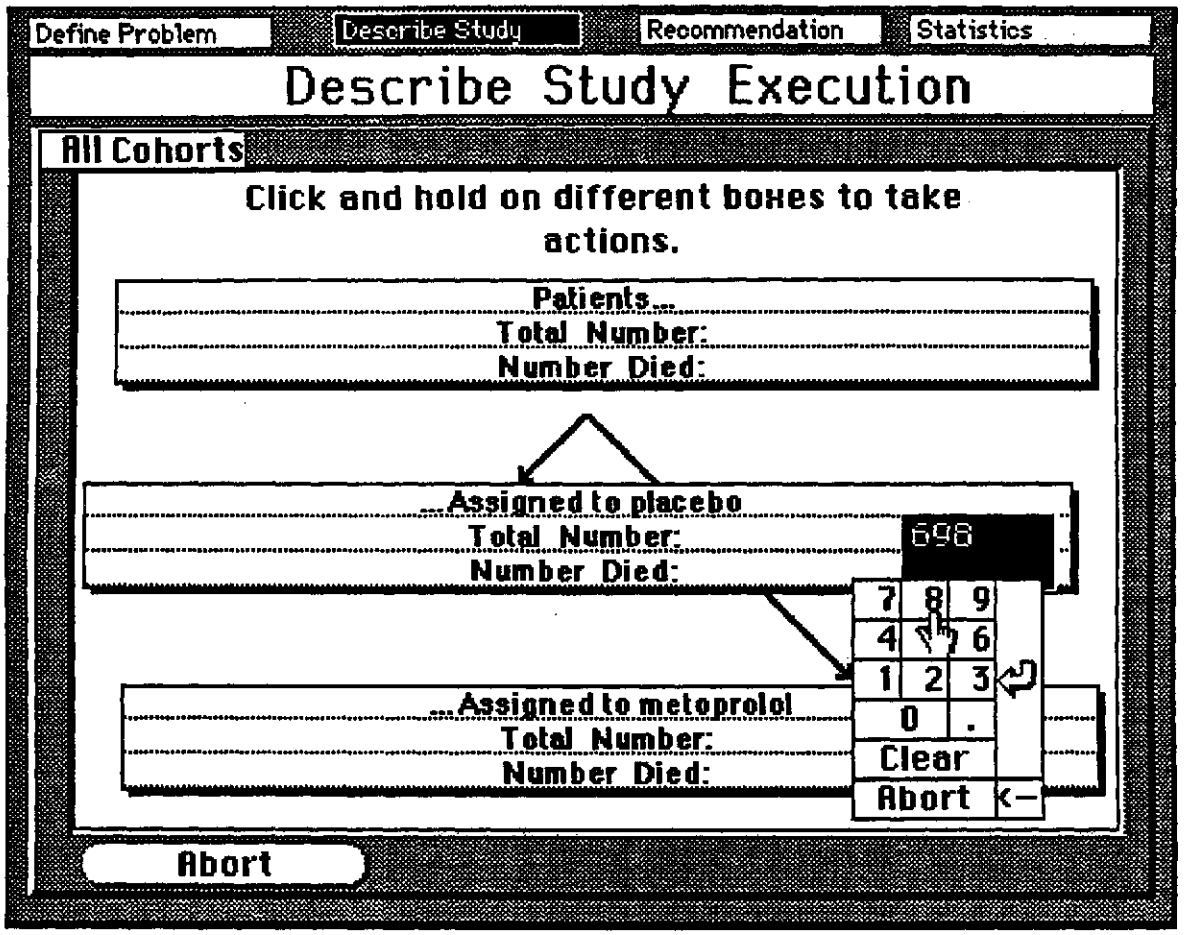


Figure 6.6: Entering evidence into THOMAS. A pop-up keypad interface is anchored to the cohort slot for which the entered number is evidence.

Whenever a datum is input, to ensure that the input datum does not contradict data already input, THOMAS performs constraint propagation through the patient-flow diagram. If there is a contradiction (e.g., a cohort's input total number is not equal to the sum of the already-input total numbers of the cohort's component subcohorts), THOMAS issues an error message and refuses the input datum. The system

also computes sums or differences, if two out of three cohorts (among the parent and two children cohorts) have their data defined; it then stores the results in the appropriate cohort.

6.4 Output Review

There are five different questions to which THOMAS responds:

1. Which drug should the user employ?
2. How has the user's belief in a parameter changed from before to after having viewed the evidence?
3. How do the user's beliefs in different parameters compare?
4. How sensitive are the user's posterior beliefs to different prior beliefs?
5. What effects do different methodological concerns have on the final conclusion?

I shall describe how THOMAS answers each of these questions.

6.4.1 THOMAS's Recommendation

The decision-analytic basis for answering the first question is the utility calculations of Equation 5.1. The system must first choose the parameter on which it will base its calculations, and must then present its conclusions to the user.

The outcome of interest—the patient's lifespan (and, therefore, life expectancy)—can be made dependent on one of two parameters that represent the mortality rates of interest. The *efficacy* choice is concerned with the biological effect of the drug; an efficacy-based analysis, therefore, compares the most debiased mortality-rate parameters, because they represent most closely just this effect. Thus, the efficacy choice

uses the *population* parameters for patients who were assigned to (and who received) the experimental (or control) drug. The *effectiveness* choice is concerned with the effect of the drug in clinical practice; an effectiveness-based analysis, therefore, compares the parameters that reflect all the protocol departures clinicians are likely to encounter, but debiases with respect to the reliability of the assessment instruments. Thus, the effectiveness choice uses the *study* parameters in patients assigned to the respective therapies. The user is given the choice of which analysis she wants (see Figure 6.7a). Thus, THOMAS's hierarchy of parameters enables the system to translate the user-based semantics of effectiveness and efficacy into a meaningful clinical report (see Figures 6.7b and 6.7c).

THOMAS answers this first question by plotting the posterior life expectancies ($\langle L \rangle$) of the two drugs, and comparing the life expectancies, taking into account the pragmatic difference specified by the user. The plot can be labeled, annotated, and saved for later examination.

6.4.2 User's Beliefs

The uncertainties given to THOMAS are the user's prior beliefs in different parameters. The uncertainties THOMAS calculates are the beliefs the user *should* have, given those prior beliefs, given the methodological concerns the user entered, and given the numerical data, assuming the user wishes to remain consistent with Bayesian probability theory. THOMAS's reporting capability allows the user to view the prior and posterior beliefs in parameters, or to have the machine guide the query process.

6.4.2.1 Prior and Posterior Beliefs

Figure 6.8 shows THOMAS's report of prior and posterior probabilities. The report has three components. The first component is the numerical report of the means and standard deviations of the two distributions. The mean reported is actually the mode

of the distribution; the difference between mean and mode is slight in unimodal beta distributions. The second component is the credible interval (see Section 4.3.5.1); in the figure, 95-percent credible intervals are shown. The third component is the graph of the belief.

The interpretation of Figure 6.8, the prior and posterior beliefs in the population mortality rate of patients assigned to placebo in the metoprolol study, is as follows. The prior belief had been set as *total ignorance,* which resulted in a $\mathcal{BE}\left(\frac{1}{2}, \frac{1}{2}\right)$ prior distribution, which has a mean of 0.5 and a standard deviation of 0.707. These latter numbers are shown in the left-hand panel. The 95 percent credible interval for this distribution ranges between 0.0 and 1.0.[3] The graph of the distribution shows a peak at 0.5; it is fairly flat, except at the extremes.[4] The posterior belief, with a mean of 0.0894 and a standard deviation of 0.0108, represents a $\mathcal{BE}\left(62.5, 636.5\right)$ distribution, which is the appropriate posterior for a $\mathcal{BE}\left(\frac{1}{2}, \frac{1}{2}\right)$ prior distribution, updated with the data of 62 deaths and 636 survivors in the placebo group. The posterior credible interval is much narrower than is the prior credible interval. The belief curve is narrower than is the prior belief curve, reflecting smaller uncertainty.

6.4.2.2 Semantically Based Statistical Questions

Further statistical questions can be posed via an additional level of semantics above the raw statistical reports: The system frames comparisons between parameters as answers to statistical questions in which the user probably is interested. Each methodological concern is associated, by the knowledge engineer, with a comparison between parameters specific to that concern. For instance, the very design of a two-arm RCT

[3]The credible interval is, $0 + \epsilon$ to $1 - \epsilon$, where ϵ is a number too small to be printed.

[4]A $\mathcal{BE}\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution is bathtub shaped—flat in the middle, and rising to infinite likelihood at both extremes. For consistency in reporting, THOMAS makes all distributions unimodal.

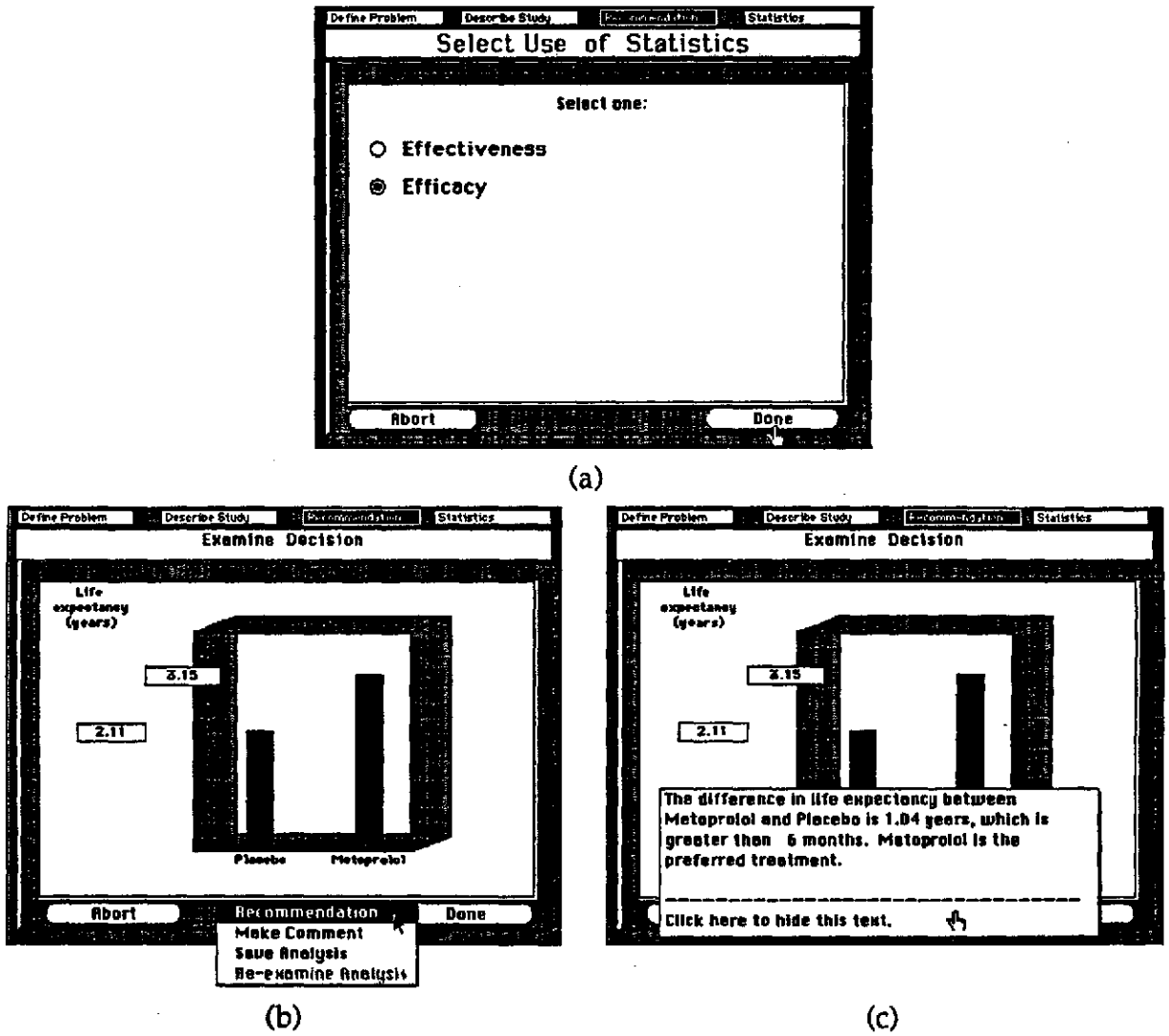


Figure 6.7: The question of drug choice. THOMAS answers this question in three steps. (a) Choice of use of statistics. THOMAS enables the user to answer the question in terms of efficacy or of effectiveness. (b) Plot of posterior life expectancy. The life expectancies are a function of the posterior uncertainty in the parameters implicitly chosen via the choice of use of statistics. (c) Recommendation of treatment. THOMAS's recommendation is a function of the pragmatic difference, previously set by the user (6 months, in this case).

is associated with the following question: Is the experimental-drug mortality rate different from the control-drug mortality rate? This question is attached to a comparison between the population mortality rates in patients assigned the experimental drug and in patients assigned the control drug. When a methodological concern is included by the user, the system creates the comparison's label and selects the parameters to be compared (see Figures 6.9a and 6.9b).

A comparison has two parts. One part is the display of the belief information for each parameter, including the mean, the standard deviation, and the belief curves (see Figure 6.9d). The other part is a display that actually answers the question (see Figure 6.9c). For the second part, the machine creates a new parameter, the identity of which depends on the question and is set by the system builder. In THOMAS, these parameters are always the *difference* between two parameters. Because such a parameter is a function of other parameters, its prior probability can be calculated without further input from the user. Thus, the system can calculate the parameter's posterior belief. The system uses this posterior belief to answer the statistical question by examining the amount of belief adherent to values greater than 0 (i.e., $P(\text{difference} \geq 0 \mid \text{prior beliefs and evidence})$). If a significant amount (e.g., 95 percent) of belief is attached to positive values, then the system can draw a statistical conclusion. For instance, in the mortality-rate comparison, if more then 95 percent of the posterior belief is that the difference is positive, than the control drug has a statistically significantly higher mortality rate, and, hence, the system answers that the experimental drug is statistically better. This comparison is similar in flavor to the z-test for proportions, or to the t-test; see Figure 4.11. However, the answer to the Bayesian comparison does not determine the system's recommendation, as it does in classical statistics.

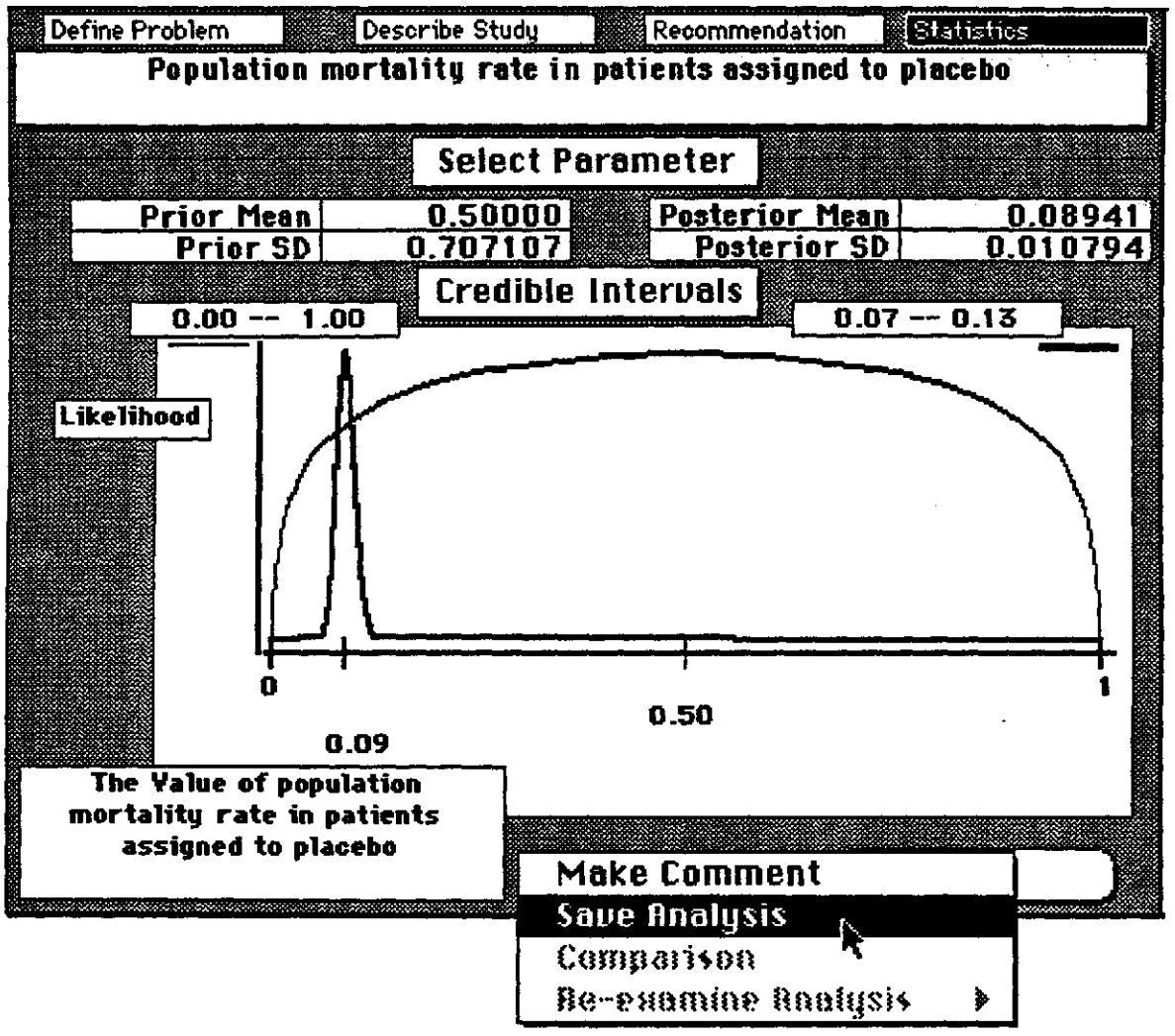


Figure 6.8: The question of prior versus posterior belief. THOMAS displays the prior and posterior beliefs in the population mortality rate in three ways: numbers, credible intervals (Bayesian credible sets), and belief curves. Numeric and credible-set information about the *prior* distribution is on the left, and its belief curve is graphed with the thin line. Numeric and credible-set information about the *posterior* distribution is on the right, and its belief curve is graphed with the thick line. The user is able to make a comment about the report, or to save it for later review.

6.4.3 Sensitivity Analysis

The fourth question allows the user to examine the sensitivity of any conclusion to changes in prior belief. The strategy THOMAS uses in answering this question is to allow the user to make an explicit change in a prior belief in a specific parameter, and then to display the implications of that change in belief.

Figure 6.10 shows a sample interaction between THOMAS and the user in the metoprolol problem. The system allows the user to choose a parameter and to alter her prior belief (see Figure 6.10a; the menu of parameter names displayed by THOMAS is not shown). The physician's entry of prior belief follows the same format as does the entry of initial prior belief (see Section 6.3.3). In Figure 6.10b, the user has employed the equivalent–sample-size method of describing her prior belief in the placebo-population mortality rate; in Figure 6.10c, the user has employed the confidence-interval method for the metoprolol-population mortality rate. When the user indicates she is finished modifying her prior beliefs (not shown), the system performs probabilistic updating as usual, and makes available the same statistical analyses as in the baseline analysis. Figure 6.10d shows a comparison between the prior beliefs in the placebo-population (thin line) and metoprolol-population mortality rates, and Figure 6.10e compares the posterior beliefs in the two parameters. Finally, Figure 6.10f shows the life expectancies based on the new posterior distributions. Comparing this figure to Figure 6.7b, we see that the new prior beliefs do not change the recommendation.

Note that the conclusion regarding the effect of the change in prior belief is made by the user; THOMAS, at present, has no method for comparing the posterior means between analyses, either numerically or graphically.

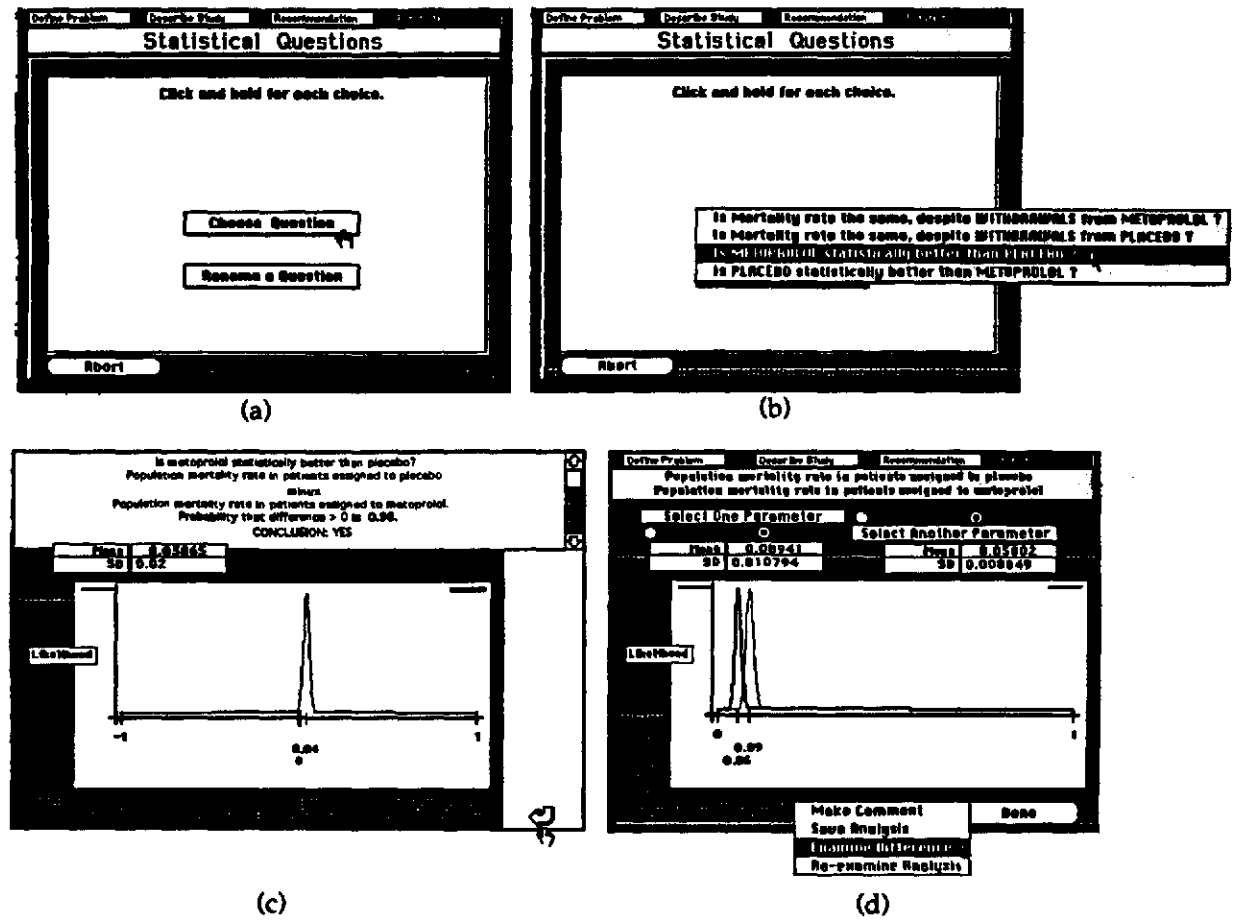


Figure 6.9: The question of statistical conclusion. (a) THOMAS allows the user to pose a statistical question. (b) The user selects a question from a list provided by the system. THOMAS here displays the questions generated by the baseline metoprolol analysis. The user has chosen to ask whether the experimental mortality rate is lower than the control mortality rate. (c) THOMAS answers the question by examining a new parameter, the *difference* between the two population mortality rates. If the probability that the difference is positive is large (over 95 percent), then the system reports the conclusion, *Yes*, the treatments are statistically different. (d) The system also displays a comparison between the two parameters that make up the difference parameter.

6.4.4 Multiple Analyses

Although THOMAS constructs models instead of using belief distributions over models, the user may still believe that no one single model is best. This ambivalence derives from the fact that the more specific the model, the less certainty there will be about estimates, yet the less specific the model, the less domain knowledge is incorporated within the statistical model.[5] To give the user the ability to consider different ways of viewing the same study, THOMAS allows the construction of several models, and the viewing of the implications of a set of models. This process may be called a sensitivity analysis over the model structure.

THOMAS asks the user to label each analysis (see Figures 6.11a through c). If there are analyses already present, the system asks the name of the analysis of which the new analysis is an extension (not shown). Figure 6.11d shows the implied tree of analyses after four iterations. The tree indicates that the analyses taking into account crossovers, withdrawals, and classification error is each a modification of the baseline analysis, whereas the analysis taking into account withdrawals and classification error (*Withdrawals with CE*) is a modification of the analysis taking into account withdrawals.

For analyses that extend previous analyses, the system takes the user directly to the subtask of filling in the patient-flow diagram, with evidence already entered from the foundation analysis made available.

6.5 Conclusion

THOMAS delivers the Bayesian paradigm to the physician user in a number of ways. First, its sequence is based on the Bayesian paradigm of assessing prior belief before viewing evidence. Second, the details it assesses from, and reports back to, the user

[5]This ambivalence is the Bayesian equivalent of the *bias–variance tradeoff* of classical statistics.

are based on a Bayesian view of parameters, evidence, and decisions. The assessments and the recommendation are communicated in as non-Bayesian a language as possible, however, as they use domain terms. Third, the system provides the decision-analytic tool of sensitivity analysis as a way of evaluating the influence of personal beliefs on the values of parameter or of evaluating the influence of different methodological concerns. We now turn our attention to how the system achieves these capabilities.

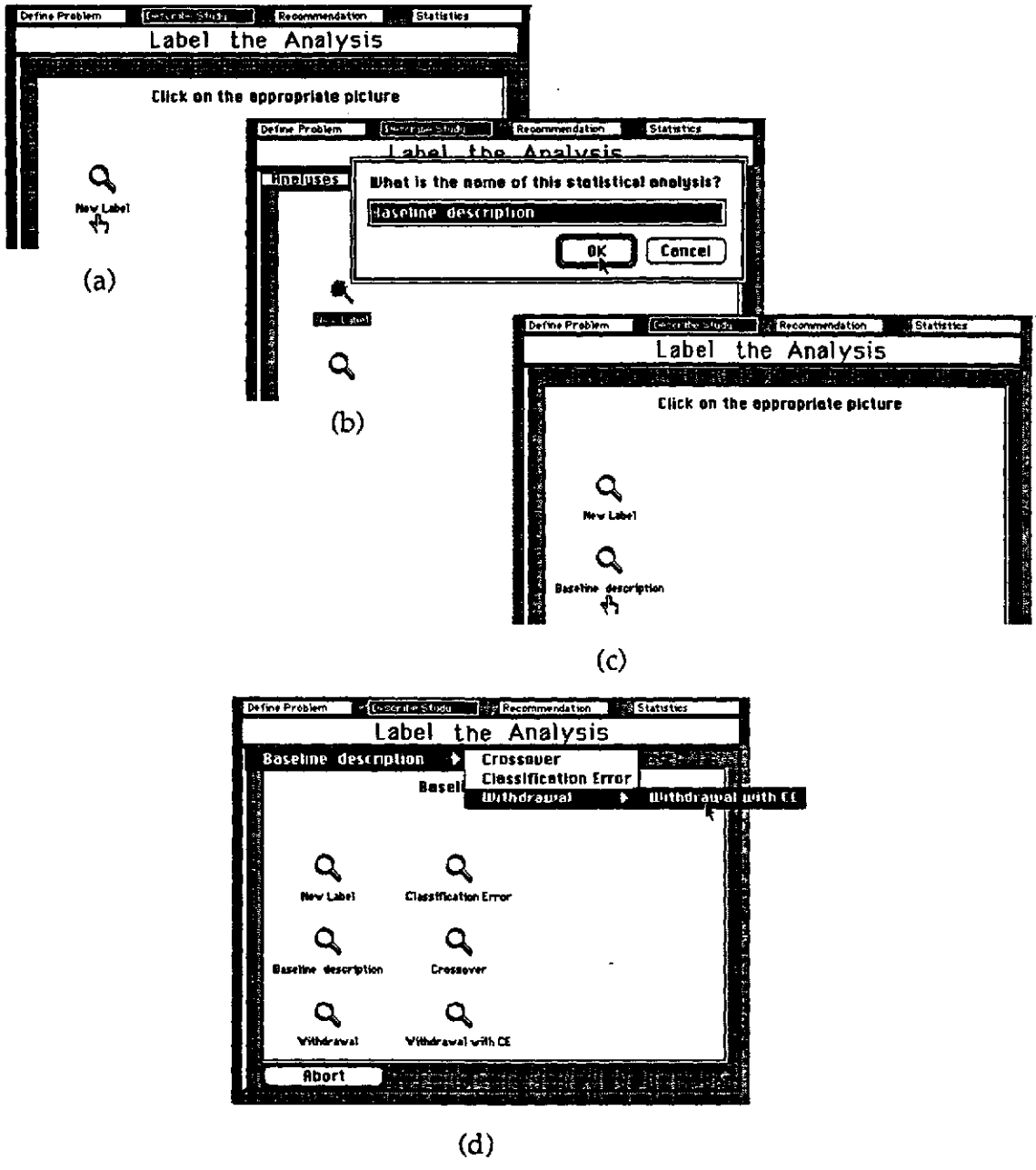


Figure 6.11: The question of effect of methodological concerns. (a) Request for a new analysis. (b) Entry of new analysis name. (c) Request to start the analysis. (d) Analysis tree after several analyses have been entered. Screen displays have been truncated.

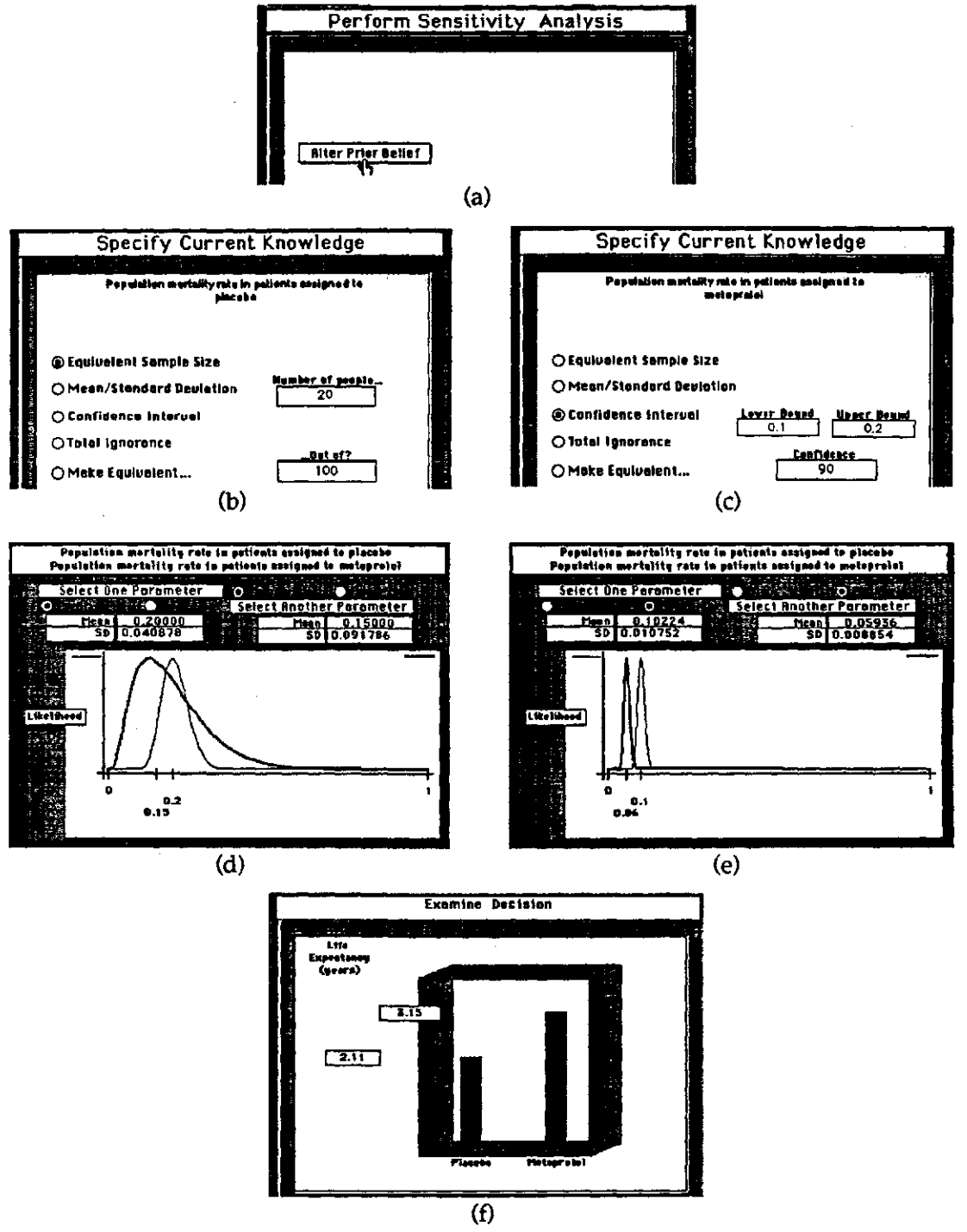


Figure 6.10: The question of sensitivity to prior beliefs. (a) Beginning of dialogue. Choice of parameters names is not shown. (b) Entry of new placebo prior. (c) Entry of new metoprolol prior. (d) Display of prior-belief information. (d) Display of posterior-belief information. (e) Display of recommendation based on new posterior beliefs. Screen displays have been truncated.

Chapter 7

Data Structures and Algorithms

Having seen how the machine and the user interact with each other, we need now to see how the system achieves that interaction. We saw, in Figure 1.6, that this process comprises three components: Bayesian methodological formulation, probabilistic updating, and utility maximization. In this chapter, we shall examine these segments, spending most of our time on the first component: how the system represents, and executes, the process of statistical-model construction.

Because THOMAS represents statistical models as influence diagrams, this problem is an example of the more general problem of assisting a relatively naïve user to construct an influence diagram. Statistical modeling is a process of responding to the *type* and *availability*—the *metadata* (see Sections 3.3.1 and 4.3.3)—of the study. There is structure to this process: There are restrictions to the sequence in which data may be analyzed, and particular metadata demand that specific modeling actions be taken. The system builder[1] can assess directly from domain experts the *rules*[2] that constrain this process, or can derive them from statistical definitions. The challenge

[1] I shall use the term *system builder* to refer to the person who builds a program such as THOMAS; I shall use the term *user* or *physician reader* to refer to a user of a program such as THOMAS.

[2] I use the word *rule* to refer to domain-level strategy heuristics; they need not be implemented as data structures to be processed by a backward- or forward-chaining inference mechanism.

for the system builder is to coordinate these domain rules with the symbolic and algorithmic needs of the process of influence-diagram construction in a way that is accessible to the user.

In Section 7.1, I shall show how the need to adjust conclusions in light of methodological concerns, and how the need for those adjustments to be composed modularly, lead to the metadata-driven approach. This approach uses three data structures: the patient-flow diagram, the metadata-state diagram, and the statistical model. The interactions among these structures are shown in Figure 7.1. I described patient-flow diagrams in Section 6.3.2. I shall discuss the metadata-state diagram in Section 7.2, defining the diagram, explaining the metarules for its construction, and describing its use in THOMAS. In Section 7.3, I shall describe statistical models, emphasizing how the models are structured to allow for automatic processing. Section 7.4 elaborates how these three structures interact to effect the construction process. The process is demonstrated in Section 7.5, where it is applied to the metoprolol example; this example should clarify the use of the data structures and algorithms.

Section 7.6 explains how THOMAS uses the statistical model created in the construction process to update the user's beliefs in the parameters. Section 7.7 shows how THOMAS uses the posterior probabilities so calculated to make its recommendation.

I close the chapter with Section 7.8, where I compare the metadata-driven approach with influence-diagram–construction methods of other investigators.

7.1 Adjustments and Modularity

The notion of constructing statistical models is closely tied to the concept of *adjusting* a statistical conclusion in light of methodological concerns (see Sections 3.3.4 and 4.3.4). As we discussed in Section 3.4, the natural way for physicians to make

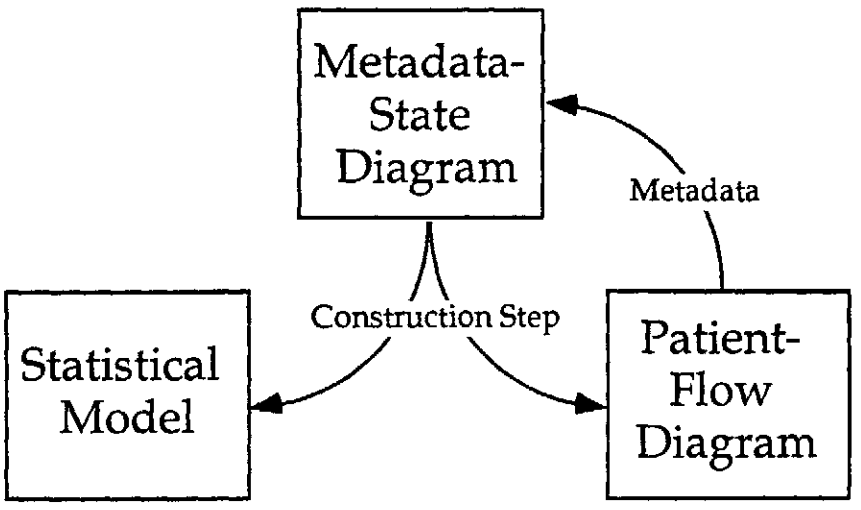


Figure 7.1: Interactions among THOMAS's components. The user's interaction with the patient-flow diagram sends a message regarding metadata to the metadata state diagram. If the message is valid, the metadata-state diagram executes a construction step. The effect of that step on the statistical model is to induce changes in the influence diagram representing the statistical model, such as the creation of new parameters. The effect of that step on the patient-flow diagram is to induce changes in that diagram, such as the creation of new subcohorts.

such adjustments is to modify parameter estimates *after* they been calculated, working from the data to the adjusted model parameters. The appeal of this strategy is the *modularity* of the corrections: the ability of the analyst to consider each arm of a study (or each study of a set of studies) *independently* of the others. If such independence were to hold true, then adjustment formulae for one arm (or study) would apply, formally, to any other. An implication of this notion of independence is that the overall adjustment due to a set of methodological concerns would simply be the union of adjustments due to the individual concerns making up that set. We shall find that, as in rule-based expert systems (Heckerman and Horvitz, 1988), adjustment in the data-driven direction leads to incorrect conclusions. The nature of the unit of modularity, therefore, must be reformulated.

The problem with data-driven adjustments is that the process violates two knowledge-level criteria established in Section 2.3 for THOMAS: objectivity and normativity. In Section 3.5.1, we saw that these adjustments may not be auditable, leading to a nonobjective process. As we shall now show, from the Bayesian perspective of normativity (see Section 4.6.3), the strategy may not be *coherent*, because it may lead to the calculation of a posterior belief that is inconsistent with the prior belief and the data. Inconsistencies arise because methodological concerns, when included in the statistical model, make the arms (or studies) *de*pendent on each other. To remove the inconsistencies from within the classical framework, the analyst would have to obtain information that is difficult, if not impossible, to assess from the investigators or from the user. From within the influence-diagram–based framework, the solution to this difficulty is to separate the assessment process from the adjustment process; the modularity is situated in the assessment process, whereas the adjustments emerge from the calculation of probabilistic updates.

The following example should clarify the potential incoherence and lack of modularity of the data-driven–adjustment process. Consider a study where outcome classification had been assessed in a way that is potentially unreliable; the analyst would be interested in assessing the sensitivity and specificity of the assessment process. The two methodological parameters—sensitivity and specificity—are common to the two treatment arms; this sharing of parameters makes the adjustments for the evidence with*in* each arm dependent on the outcome parameter in the *other* arm. We shall see that the dependencies introduced are difficult to assess.

Figure 7.2 displays this example. There are two study arms, the control and the experimental, where the entities of interest are the mortality rates in the two arms. Each arm has outcomes—deaths—that provide evidence for the respective mortality rates. From the simple model in Figure 7.2a, the adjustment model can be constructed in the direction shown in Figure 7.2b. No new assessments are needed, and the reversal

leads to formulae for deriving the mortality rates from the observations, in the data-driven direction and independently in each arm. As the notion of modularity implies, the formulae for the control- and for the experimental-arm mortality are the same.

Figure 7.2c shows the original model of Figure 7.2a with the methodological parameters of sensitivity and specificity added. To derive the appropriate adjustment formulae, the analyst has two choices. One choice is to assess new dependencies, such as the dependency of the sensitivity on the specificity, and the dependency of the mortality rates on the methodological parameters (Figure 7.2d). These dependencies, however, are difficult to determine, and run against the experience of physicians. The other choice is to assess the a priori dependence of one mortality on the other (Figure 7.2e). But this dependence violates the desideratum of assessing evidence for the arms independently of each other. In either case, the adjustment formula for one arm is *not* the same as the formula for the other one.[3]

There are two solutions to this problem. One is to teach physicians how to make the needed assessments. This approach would force physicians to view their medical experience in ways that are clearly not natural to them.

The second is to separate the assessment from the calculation: Allow the physician reader to structure the analysis in the order with which she is comfortable (the data-driven direction), but to make the system construct the actual model in the model-driven direction. The model-driven direction remains important, because it defines what information is to be assessed from the user. The *metadata-driven* approach works in this way: The user supplies metadata in the data-driven direction, which are used by the system to produce the statistical model in the model-driven direction. The modified statistical model then has sites for evidence that has semantics of the data-driven direction, and that, therefore, can be assessed from the user.

[3]Because, as I showed in Chapter 3, influence diagrams can be used to represent classical models, the argument I have just presented—using an apparently Bayesian structure (the influence diagram)—is not biased against the classical approach.

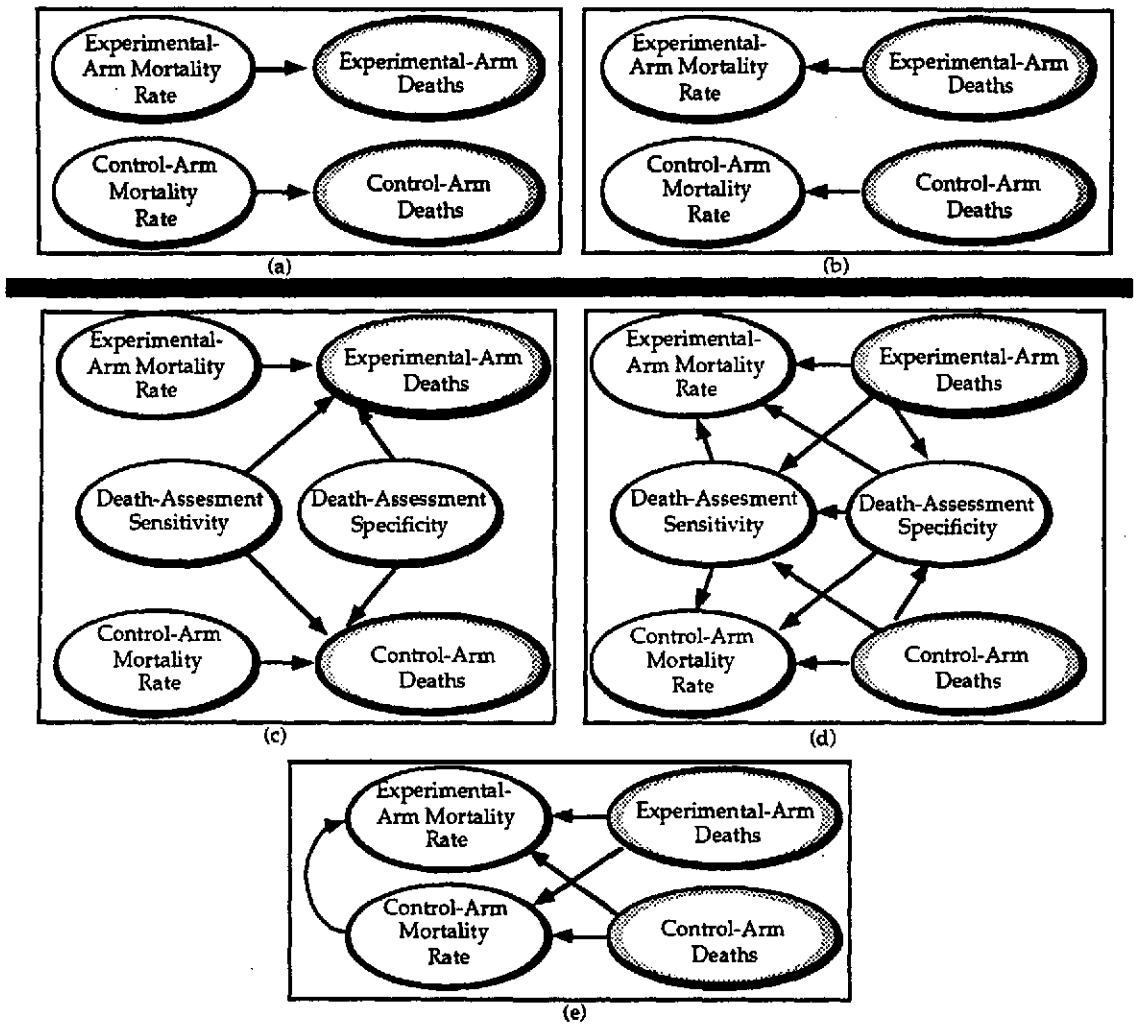


Figure 7.2: Dependence in adjustments. Arcs show the direction of assessment. (a) A simple two-arm model for the relationship between the entity of interest (mortality rate) and evidence (deaths) in two arms of a study. (b) The simple model, with dependencies reversed, as in data-driven adjustment models. The two arms remain independent of each other. (c) The simple model modified by the introduction of methodological parameters (sensitivity and specificity) in common to the two arms; the arms are now dependent on each other. (d) The previous model with the dependencies reversed. New dependencies must be assessed. (e) The previous model with the methodological parameters averaged out. The dependencies of the two arms are made explicit; compare with Fig. 7.2b.

7.2 The Metadata-State Diagram

The task of coordinating metadata-driven statistical-model construction is given to the *metadata-state diagram.* The model-driven direction of statistical-model construction is implemented in the actions—construction steps—taken by the system in response to the user's concerns. *A* **construction step** *for the statistical model is an action that takes a formally valid influence diagram and produces another formally valid influence diagram, where the new influence diagram now has components representing the concern of interest.* The final statistical model, therefore, has a history, consisting of a sequence of construction steps. Because there are restrictions on—and preconditions to—taking individual construction steps, the specific rules for the construction process induce an order on the space of histories of statistical models.

The metadata-state diagram reflects this order. The *states* in the diagram refer to states of the analyst's knowledge about a group of patients—the metadata—and the directed *arcs* refer to the construction steps permitted in response to metadata possibly available at a particular state of knowledge.[4] The semantics of each state are those of the the path from the root state to the state—the state's history. Thus, the data-driven direction is established through the metadata permitted at each state.

Figure 7.3 shows a simplified version of the diagram that governs THOMAS's interaction with the user. The displayed diagram implements several *rules.*[5] Arc 1 embodies the rule that *protocol departures may be considered only after a patient's assignment status is known.* This rule concerns the *semantics* of protocol departure, a specific *type* of concern. Note that the knowledge about the assignment status belongs to the analyst (reader), in keeping with the definition of a state in this diagram. Arc 2 (more precisely, a corresponding absence) symbolizes the rule that *outcome*

[4]Loops are allowed; e.g., the ability to modify errors in the evidence requires the implied recursion.

[5]The use of rules in building a Bayesian system is permitted because of the Bayesian difficulty with model construction; see Section 4.4.

evidence may not be given for patients who are known to have been lost to followup. This rule concerns the implications of a particular methodological concern. Arc 3 represents the rule that *classification error may be modeled only if outcome evidence has been provided.* This rule concerns the *availability* of information. These rules all concern metadata of the study, rather than the primary data.

System builders have the choice of making these rules explicit, or making the state diagram induced by the rules explicit, or both. In choosing to make the rules explicit, as in a rule- or blackboard-based system, the system builder leaves implicit the interactions among the rules. These interactions are made visible in the state diagram. However, in choosing to make the metadata-state diagram explicit, the system builder may lose the explanatory power resulting from the modular semantics of rules, and risks an exponential explosion in the size of the diagram. When both rules and diagram are used, the rules can act as an interface for the system builder: The system builder enters rules, and the machine modifies its metadata-state diagram. Different choices among these approaches lead to different control algorithms. In this dissertation, I have chosen the explicit-diagram approach, because of the exploratory nature of the work.

7.2.1 Metarules

The *structure* of the metadata-state diagram is domain-dependent. For the literature problem, that structure is based on the sequence of events that subjects may experience in the course of a study. Diagrams of potential patient histories, which are potential patient-flow diagrams, are found in clinical epidemiology texts (Feinstein, 1985) or in expert-system interfaces (Musen, 1989); an example is shown in Figure 1.1. Because the metadata-state diagram is implicitly a structure of rules, rules for generating the diagram are *metarules.* Thus, the metarule for the structuring the metadata-state diagram for the literature problem is the following.

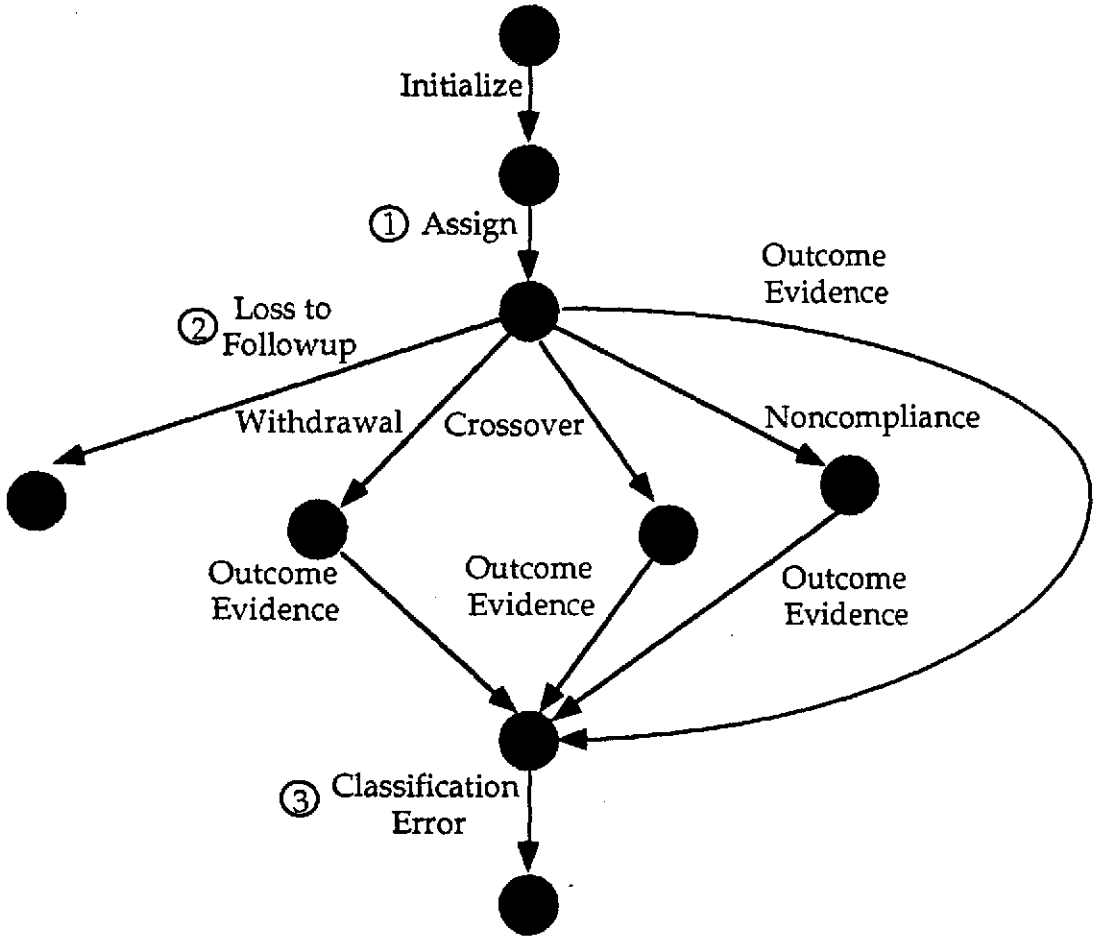


Figure 7.3: A metadata-state diagram. This diagram presents a simplified version of the metadata-state diagram used by THOMAS. Each arc is labeled with the name of a metadatum that is valid at the indicated state. The numbered arcs are discussed in the text. The direction of an arc assumes that time increases in the same direction as the arc.

Metarule 7.1 *Order states in the metadata-state diagram according to the temporal sequence of patients' experience in a study.*□

Metarule 7.1 expresses the notion of using the potential patient-flow diagram to construct the metadata-state diagram. The *use* of the metadata-state diagram in analyzing a study is guided by one metarule.

Metarule 7.2 *Assign evidence to the most specific cohort possible.*□

Metarule 7.2 is the Bayesian analyst's response to the reference-class problem (Kyburg, 1983). The metarule keeps the model as simple as possible, and it results in the smallest posterior uncertainty in a parameter: The further away evidence is from its appropriate parameter, the more likely intervening concerns will dilute that certainty. The metarule is implemented implicitly in THOMAS's user interface (Section 6.3.4).

The use of the metadata-state diagram to sequence a *series* of analyses of the same study (see Section 6.4.4) is guided by the principle of not bringing in irrelevant information.

Metarule 7.3 *If (1) no specific prior knowledge is available, or (2) no specific evidence is available, or (3) the type of distinction is not important to the problem at hand, then do not model the concern (i.e., do not invoke a modeling rule); enter concerns for which you have the most evidence before concerns for which you have less evidence.*□

The main justification for Metarule 7.3 is that it minimizes the analyst's time spent in the analysis. In terms of the usefulness of the results—also important to the analyst—if the conditions of the metarule are met, but the user does proceed to model the concern, then the posterior belief in many parameters may be diffuse, no

matter how much specific information is in place in other parts of the model; a set of such diffuse beliefs is not useful. This diffuseness becomes even worse when there are more basic parameters than there are strong prior beliefs or data. I currently leave the metarule implicit in the options that the system makes available to the user.

7.2.2 Use of the Metadata-State Diagram

The metadata-state diagram mediates between the data-driven direction, via metadata input, and the model-driven direction, via model construction of statistical-model construction. The top-level controlling loop of the algorithm works as follows: The patient-flow diagram translates the user's metadata directive into a machine-usable format that includes the metadatum and a target cohort. Unless the directive signals termination of the modeling process, the system examines the metadata-state diagram to determine whether the directive is permitted, by inspecting the arcs emanating from the state in the diagram referred to by the target cohort. If the directive is permitted, the system then executes the construction step indicated by the metadatum, indirectly modifying the patient-flow diagram and the statistical model. If new parentless parameters are created by the construction step, the user is asked to assess prior beliefs about those parameters.

7.3 The Statistical Model

The third component of the metadata-driven approach is the statistical model, represented as an influence diagram. The statistical model is a specially structured influence diagram, whose structure provides a lexicon linking user-based semantics to statistical-model components. Thus, the statistical model involves *types* of variables, as well as *types* of larger components that allow the statistical model to represent

methodological concerns. The larger components—within which variables reside—are *levels,* whose semantics are based on the approach for likelihood debiasing given in Section 4.2.3 and on the design model for the program (see Figures 4.14 and 5.1). There are four levels of variables in THOMAS's hierarchy: the population, the study, the effective, and the patients levels. The topology of the hierarchy is shown in Figure 7.4.

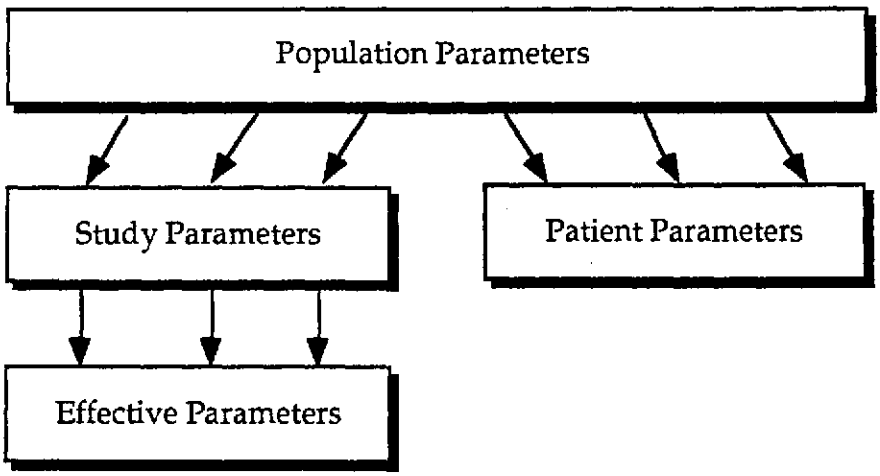


Figure 7.4: The levels of THOMAS. Variables are defined within only one of the levels: population, study, effective, or patient.

As with any hierarchy, the purpose of separating nodes into levels is to limit the interactions possible between nodes, which makes construction and assessment easier. Nodes in one level of the statistical model may have parents only in their own level, or in the level above them, and may have children only in their own level, or in the level below them; nodes two layers away are irrelevant if values of the nodes in the intervening layers are known. This limitation exploits the notion of conditional independence inherent in influence diagrams.

Within levels, nodes and arcs are typed, allowing further semantic identification and interpretation. The type of a *node* dictates that node's semantics to the user, its possible values, the form of its probability-distribution function or deterministic

function, and its level in the hierarchy of the influence diagram. The type of an *arc* depends on its location: between levels (*interlevel* arcs) or within levels (*intralevel* arcs). *Intralevel* arcs are divided into arcs connecting nodes of the same type (*istotypal* arcs), and those connecting nodes of different types (*heterotypal* arcs). By virtue of the acyclic nature of influence diagrams, there is no need for *up*ward arcs between levels.

Table 7.1 lists the classes and types of nodes and arcs used in THOMAS. We shall now examine the components of this structured statistical model.

7.3.1 Types of Nodes

As suggested by THOMAS's design model, THOMAS uses two types of variables: the *outcome* type and the *parameter* type. An outcome-type variable reflects an outcome of interest in a study, such as lifespan. A parameter-type variable parameterizes the uncertainty in an outcome-type variable. The statistical model is composed of parameters, whereas the decision model—linked to the statistical model at the patient level—contains outcome nodes as well.

The possible values of an *outcome* node depend on the outcome itself. For lifespan, the possible values are all positive numbers. The probability-distribution function (pdf; see page 53) associated with an outcome node depends on the outcome type and on the probability model assumed by the knowledge engineer.

The possible values for a *parameter* node depend on the pdf for which the variable is a parameter. For an instantaneous-rate parameter of an exponential distribution, these values are all numbers between 0 and 1. The parametric pdf for a parameter node may be chosen on the basis of its possible values. For THOMAS, I assume a beta distribution for parameters whose possible values are bounded, a log-normal distribution for parameters whose possible values are bounded on one only side, and a normal distribution for parameters whose possible values are unbounded (Shachter,

Table 7.1: Types used in statistical-model construction in THOMAS.

Type Class	Type Used in THOMAS
Level	Population
	Study
	Effective
	Patient
Node	Outcome
	e.g., lifespan
	Parameter:
	Outcome parameter
	e.g., mortality rate
	Methodological parameter
	e.g., crossover rate (a subset *parameter)*
Arc	Interlevel:
	Population → Study: Selection bias
	Study → Effective: Measurement reliability
	Isotypal:
	Population parameter → Population parameter: Domain knowledge
	e.g., baseline mortality rate = placebo mortality rate
	Study parameter → Study parameter: Protocol departure
	e.g., mixture model: dependence of an outcome parameter on component outcome parameters
	Heterotypal:
	Methodological parameter → Outcome parameter: Protocol departure
	e.g., mixture model: dependence of an outcome parameter on a subset parameter

1988a).

Parameters come in two subtypes: outcome and methodological. *Outcome parameters* parameterize beliefs in domain-level concerns (outcomes). Outcome parameters are further divided into subtypes that are named by the different levels. Thus, there are population outcome parameters, study outcome parameters, effective outcome parameters, and patient outcome parameters. This type classification provides the user access to parameters, through the semantics of their levels.

Methodological parameters are divided into subtypes as well, depending on the methodological concerns represented in the system. A *subset parameter* signifies the proportions of patients in a cohort who did, or did not, experience a protocol departure. Other subtypes depend on the particular methodological concern. For instance, modeling the methodological concern of noncompliance requires a special parameter to represent the proportion of time a patient remains compliant with therapy.

Finally, nodes have types defined within the language of influence diagrams: deterministic parameters, chance parameters, and basic parameters (chance parameters without parents).

7.3.2 Types of Arcs

The type classification of arcs, that gives semantics to individual arcs and to groups of arcs in THOMAS, gives the system much of its ability to communicate meaningfully with the user. The key semantics are those of *difference:* An arc between parameters allows the system builder or the knowledge engineer to account for *differences* between subjects, whether these differences are due to baseline status, to treatment status, or to biases encountered.

The semantics of *interlevel* arcs depend on the levels connected by the arc. An arc between the population level and the study level enables the system builder and the user to represent ways in which study parameters *differ* from population parameters,

which are ways in which study patients are different from population patients. Such differences derive from selection bias. Thus, the semantics of an interlevel arc between the population and study levels are those of selection bias (see Table 7.1). An arc between the study level and the effective level enables the system builder and the user to represent ways in which effective results differ from those measurable, ideally, in the study patients. Interlevel arcs between the study and effective levels, therefore, have the semantics of measurement reliability.

Isotypal arcs furnish one location for the system builder or the user to represent domain knowledge. For instance, the *population mortality rate in patients treated with placebo* and the *population mortality rate of patients given baseline care* are two parameters in the same level. The domain belief that each results in the same lifespan is represented by an arc between nodes representing the parameters, with the identity function as the target node's determinisitic function. Isotypal arcs are used also in methodological models. For instance, an arc between the node representing the outcome parameter *study mortality rate in patients assigned to metoprolol* and the node representing the outcome parameter *study mortality rate in patients assigned to metoprolol who received placebo instead* is an arc between nodes of the same type, but denotes the fact that the latter's value is dependent on the former's. The actual form of the dependency relies on the methodological model involved.

Heterotypal arcs participate in methodological models, as well. Thus, an arc between a subset-parameter node and an outcome-parameter node is part of the model for representing crossovers.

7.4 The Construction Steps

To coordinate the actions taken in traversing the arcs of the metadata-state diagram during the construction steps, THOMAS depends on one more set of structural relationships: the relationships among the three major data structures (Figure 7.1) and the components of those data structures, shown in Figure 7.5. The consultation is instantiated by the user during a consultation with THOMAS, and comprises a definition of the problem and study, as well as two major data structures—a patient-flow diagram and a statistical model. The problem definition comprises a number of treatments (e.g., control and experimental) and a number of outcomes (though THOMAS only deals with mortality). The patient-flow diagram is a tree of cohorts. Each cohort refers to the treatment received by patients in that cohort and to outcomes assessed in the study. A cohort refers to other cohorts below it in the patient-flow diagram, and to population, study, and effective parameters. Each cohort also refers to a state in the third major data structure, the metadata-state diagram. Parameters refer back to their owning cohort and to other parameters in the statistical model. The statistical model refers to parameters, and to the history of methodological concerns that modify the statistical model in the course of THOMAS's interaction with the user. Finally, the metadata-state diagram, created by THOMAS's system builder, refers to states and to transitions. The central coordinating object is the diagram is the cohort, the object manipulated by the user. By referring to the cohort, the system can locate the different elements it needs for its processing.

Specific construction steps are specialization of the generic construction-step invoked by the top-level controlling loop (see Section 7.2.2). There are three types of construction steps in THOMAS: the inclusion of methodological concerns, the assessment of prior belief, and the assessment of evidence. We shall examine each type.

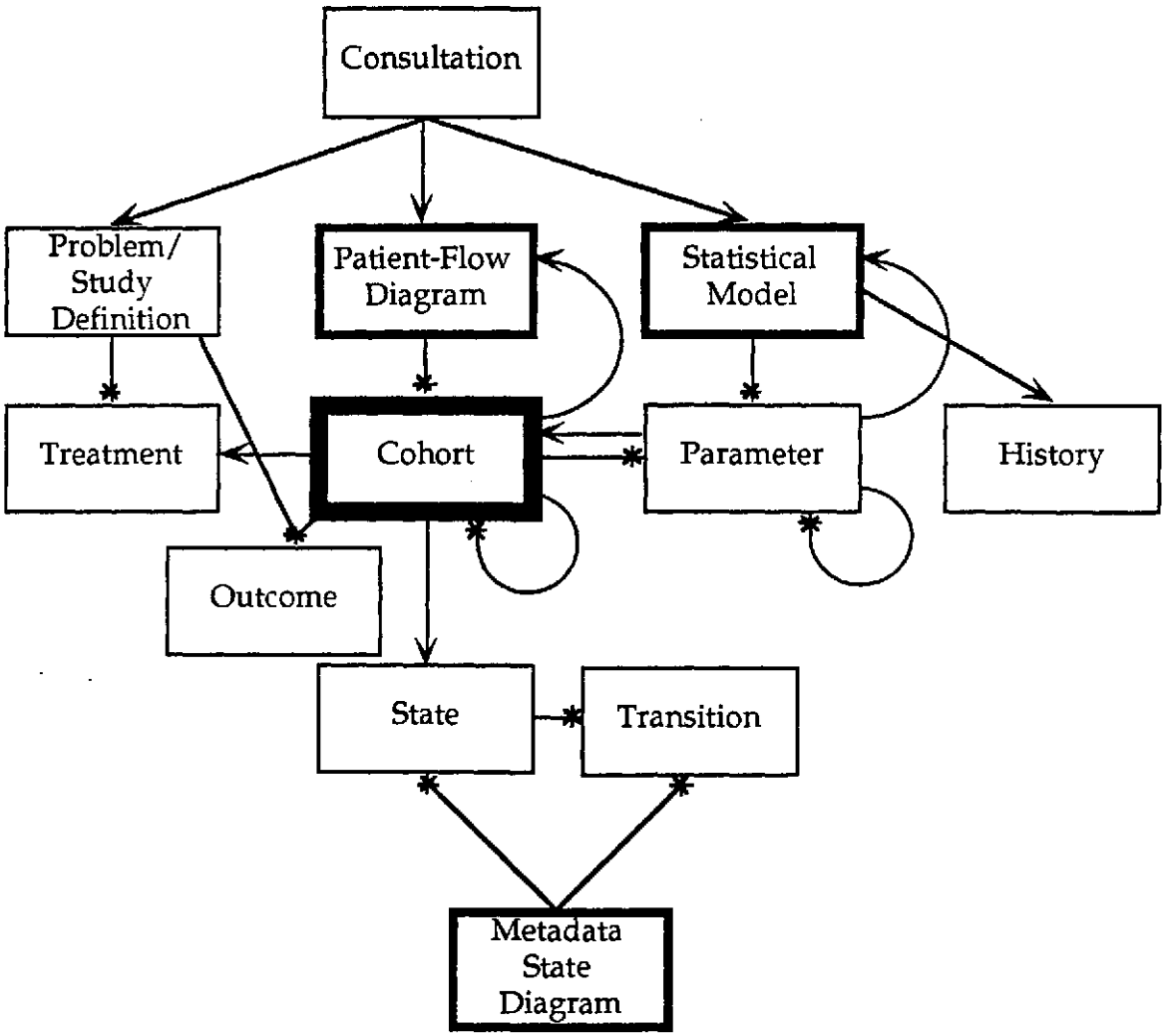


Figure 7.5: THOMAS's data structures. The relationships among the different objects are modeled analogously to an *entity–relationship diagram* (Chen, 1976), where the boxes indicate objects, the arc $\rightarrow$ indicates reference (e.g., a cohort refers to a single patient-flow diagram), and the arc $\longrightarrow\!\!\!\!*$ indicates multiple ownership (e.g., a cohort owns many parameters). The cohort entity is outlined the most distinctively, because THOMAS uses it as the point of reference with the user. The three major components of Figure 7.1 are outlined by medium-intensity borders.

7.4.1 Protocol Departures

The user adds protocol departures to the study description in the sequence comfortable to her, but following Metarule 7.3. As she does so, the system recursively splits the cohorts of the patient-flow diagram into components whose semantics depend on the specific protocol departure. The root cohort of the patient-flow diagram is the cohort of patients admitted to the study. There is always a subcohort which experienced the protocol departure (the *Yes* cohort) and one that did not (the *No* cohort).

When the system creates a subcohort, it *names* the new cohort on the basis of its parental cohort and of the protocol departure invoked. Each protocol departure is associated with a name fragment. Because the fragment may depend on dynamically collected information (such as the name of the control drug), the fragment is not static text, but is a function of other components of the consultation. The new cohort's name is a concatenation of the name of its parental cohort (e.g., *Patients who were assigned to metoprolol*) and the fragment of the protocol departure (e.g.,*who received placebo instead.*)

Modification of the statistical-model involves three steps: the creation of new outcome and subset parameters, the addition of parameters specific to the protocol departure, and the addition of domain knowledge to the statistical model.

New outcome parameters are created in the population, study, and effective levels; the patient level is never altered. The new parameters are named on the basis of the type of the parameter (e.g., *population mortality rate*), a prepositional connector (generally *in*), and the name of the new cohort. To represent the relationship between the new parameters and the parameters of the parental stratum, THOMAS creates methodological *subset* parameters; one subset parameter is made the deterministic complement of the other ($\alpha_{\text{no}} = 1 - \alpha_{\text{yes}}$), so only one new basic parameter is created (α_{yes}) at each level. For both outcome and subset parameters, the effective parameters

are initially identically dependent on the study parameters,[6] and the study parameters are initially identically dependent on the population parameters.

The linkages between the new outcome parameters and the parental parameters differ at the three levels. Because the purpose of the effective layer is to allow representation of errors of measurement, *no* connection is made between the new and old effective parameters.

The purpose of the study level is precisely to allow representation of errors that arise in the course of the study. Here, the old parameter is made a deterministic mixing function of the two outcome and the two subset parameters ($\theta_{\text{old}}^{\text{study}} = \alpha_{\text{yes}}^{\text{study}}\theta_{\text{yes}}^{\text{study}} + \alpha_{\text{no}}^{\text{study}}\theta_{\text{no}}^{\text{study}}$).[7] If the old parameter had had a previous dependency (e.g., on its corresponding parameter at the population level), that dependency is severed before the new dependencies are put in place. The avoidance of upward arcs and the hierarchical relationship between the old and new parameters prevents this process from producing cycles in the statistical model.

The main purpose of the population level is to represent domain knowledge. Initially, the new population parameters are left isolated.

After the initial creation of outcome and subset parameters at each level, extra parameters are added as needed by the protocol departure. For instance, for noncompliance, a parameter must be added at each level representing the proportion of time noncompliant patients were initially compliant.

Finally, domain knowledge is added, primarily through modification of the population level. For instance, because the cohort of patients who do *not* experience the protocol departure continues to receive the assigned drug, the population parameter corresponding to this cohort is made identically dependent on the population parameter corresponding to the cohort of patients assigned to the drug; $\theta_{\text{no}}^{\text{pop}} = \theta_{\text{assigned}}^{\text{pop}}$. If

[6] That is, an effective parameter is made deterministically dependent on a study parameter, and the deterministic function is the identity function.

[7] θ refers to an outcome parameter.

the system builder or the user wishes to represent the notion that patients who do not depart from protocol are, in general, less severely affected than patients who do, the relationship between θ^{pop}_{no} and $\theta^{pop}_{assigned}$ should be altered to reflect that domain knowledge; THOMAS does not currently include such a model.

7.4.2 Measurement Reliability

THOMAS uses classification error as its model of measurement reliability for rates. The patient-flow diagram need not be modified structurally. Modification of the statistical model takes place through changing the interlevel arc between the study and the effective levels. Specifically, the effective outcome parameter, which is initially identically dependent on the study outcome parameter, becomes dependent on the basis of the calibration function: $\theta^{effective}_{outcome} = \theta^{study}_{outcome} \cdot se + (1 - \theta^{study}_{outcome}) \cdot (1 - sp)$. Thus, two new parameters are created: the sensitivity (se) and the specificity (sp). The same outcome in different cohorts can have different sensitivities and specificities, unless the domain knowledge that they are the same is added to the model.

7.4.3 Prior Beliefs

As in any influence diagram, the assessment of prior belief is limited to nodes that have no parents—basic parameters. Only population parameters and methodological parameters may be basic parameters, because patient parameters and study parameters are always dependent on population parameters, and effective parameters are always dependent on study parameters, by definition of THOMAS's levels. After taking a construction step, the system locates new basic parameters by searching for all basic-parameter ancestors of the effective parameter of the target cohort whose prior belief is unknown. Parameters thus found are presented to the user for assessment (see Section 6.3.3).

Subset parameters should be assessed differently. Prior belief specified in a population subset parameter should have the population-referred semantics, whereas prior belief specified in a study subset parameter should refer to belief about the particular investigators or the particular study setup. THOMAS does not currently use this scheme.

7.4.4 Evidence

THOMAS allows evidence only for effective parameters. Making study data update belief in the effective parameters enables the system to add in models of measurement reliability at any point during the interaction with the user, and preserves the semantics of the effective layer.

Before incorporating study data into the statistical model, the system ensures that the new data are consistent with any previously entered information. *Consistency* here means that the number of patients in subcohorts must add up to the number of patients in the cohort of which the subcohorts are components. The consistency check is made by constraint propagation through the patient-flow diagram.

7.5 Example of Metadata-Driven Construction

To clarify how the data structures and algorithms work together, we shall inspect the metoprolol problem in greater detail, by examining the system's response to the user's initiatives; these initiatives are sent by the interface as LISP forms to the top-level processor. We shall assume that the scene has already been set by the user: She has defined the pragmatic difference, the identity of the control and experimental drugs, the name of the experimental design, and the identity of the outcome of interest. At this point, the patient-flow diagram consists of a single cohort, *Patients*. This cohort points to the first state of the metadata-state diagram. We shall now view a series of

metadata directives and their effects.[8]

(thomas-command '(initialize))

The function thomas-command signals a directive to the metadata-state diagram, and the list '(initialize) is a metadatum that refers to a transition that, as its construction step, initializes the statistical model. The initial model appears as in Figure 7.6: Only the baseline population and patient parameters are created, with the latter identical to the former, and the four levels are laid out. The topology of the hierarchy follows Figure 7.4.

(thomas-command '(assign to placebo))
(thomas-command '(assign to metoprolol))

As Figure 7.3 shows, assign is a transition permitted after initialize. The assignment construction step creates a cohort for the indicated therapy in the patient-flow diagram, and creates a population, a study, and an effective outcome parameter for the therapy. These parameters are identically deterministically dependent on one another in a chain, so the study and effective parameters are initially assumed to be equivalent to the ancestor population parameter. This state of affairs is depicted in Figure 7.7. When the system returns to the user, it locates the new basic parameters—the placebo and the metoprolol mortality rates—and requests prior knowledge about these parameters (not shown).

(thomas-command '(number for patients assigned to placebo is 698))
(thomas-command '(number for patients assigned to metoprolol is 697))

[8]The directives are taken from an actual session with THOMAS, where the directives are generated by the system's interaction with the user; hence their LISP format.

The metadata directive **number**—not shown in Figure 7.3—is one form of evidence available from the study. *Patients who were assigned to placebo* and *Patients who were assigned to metoprolol* are the target cohorts for the directives.

```
(thomas-command '(withdrawal from metoprolol))
```

This metadata directive is less specific than that for **number**, and the system has to find any and all cohorts that might be relevant. Usually, only one cohort is relevant; in this case, the single relevant cohort comprises those patients who were assigned to metoprolol and who received the drug.

The effect of the withdrawal construction step on the patient-flow diagram is the creation of two new cohorts. In this case, the cohorts represent *Patients who were assigned to metoprolol who continued to receive metoprolol* (the *No* cohort) and *patients assigned to metoprolol who withdrew from therapy* (the *Yes* cohort).

The effect of the construction step on the statistical model is shown in Figure 7.8. This figure shows the mixture model created at the study level (the arcs labeled 1), the encoded domain knowledge that patients who withdraw from therapy receive baseline care (arc 2), and the user's prior belief that the *population mortality rate in patients who withdraw* is the same as the *population mortality rate in patients who are assigned to baseline care* (arc 3). Note that population parameters are created for each column within the study-level hierarchy. These population parameters allow for different domain assumptions by different system builders or for different domain assumptions in different protocol departures. For instance, the belief that a patient who withdraws from the assigned-placebo

cohort has more severe illness than does the average placebo-assigned patient would introduce a nonidentity function between the nodes θ^{pop}_{ctl} and $\theta^{pop}_{ctl,withd}$. Modeling the subset parameter $\alpha_{ctl,withd}$ as having components in each level allows the system builder to differentiate withdrawal rates observed in the study from those expected in general.

The result of this construction step, for the statistical model, then, is that the direction of dependencies among the study parameters is bottom-up—from specific subcohorts to the most general—whereas the direction of dependencies among the population parameters is top-down; both are data-driven directions. The direction from population parameters to effective parameters is model-driven. The data- and model-driven directions are quite literally orthogonal to each other.

When the system returns to the user, it first seeks new parentless nodes; in this case, the *proportion of patients who withdrew* is such a node. THOMAS therefore asks the user (not shown) for her prior belief about this parameter, pointing out that such a prior belief gives the system the user's sense of the credibility of the researchers, as well as of the attrition expected of patients in studies such as the one under review. The parameter (the *population mortality rate in patients assigned to baseline care*) also has had no prior belief specified, and has now been incorporated into the model. The result of the user's choice of making that parameter equivalent to the *population mortality rate in patients assigned to placebo* is indicated by the arc labeled 3 in Figure 7.8.

```
(thomas-command '(number for patients assigned to metoprolol who withdrew
    is 40))
```

This datum is the only evidence regarding patients who withdrew, because the study did not give more specific information regarding outcomes

in this group.

```
(thomas-command '(evidence for patients assigned to metoprolol
    outcome mortality type count numbers 12))
```

This metadata directive is the transition referred to as *outcome evidence* in Figure 7.3. Note that the evidence is, in fact, a count of patients, because the corresponding outcome parameter is a binomial success rate. The numerical evidence incorporated into the system, in this case, is 12/698; that is, the system must combine information from the number and evidence directives.

```
(thomas-command '(classification error for mortality
    in patients assigned to metoprolol))
```

The effect of this directive on the statistical model is shown in Figure 7.9.

We can make three of points about the final statistical model. First, the methodological concerns are clearly visible as well-defined structures within the model. The mixture model for withdrawals and the calibration model for classification error stand out. Furthermore, domain knowledge and even prior knowledge can be gleaned from the structure. Second, the structure of the changes that occur to the statistical model changes at the different levels. At the population level, the changes to the statistical model reflect mostly domain and prior knowledge. The internal structure of the population level is from general to specific: from patient assignment to specific cohort histories. At the study level, the changes reflect protocol departures, and the internal structure follows a specific-to-general direction. At the effective level, the changes reflect measurement reliability, and the internal structure is dictated by the outcomes of the study. Third, notions of credibility are distributed throughout the model: in

determining prior-probability distributions for the basic parameters in the population level, in assessing prior belief in subset parameters in the study level, and in appraising prior belief in the classification-error parameters. Credibility assessments are involved, as well, in the user's choice of concerns to model.

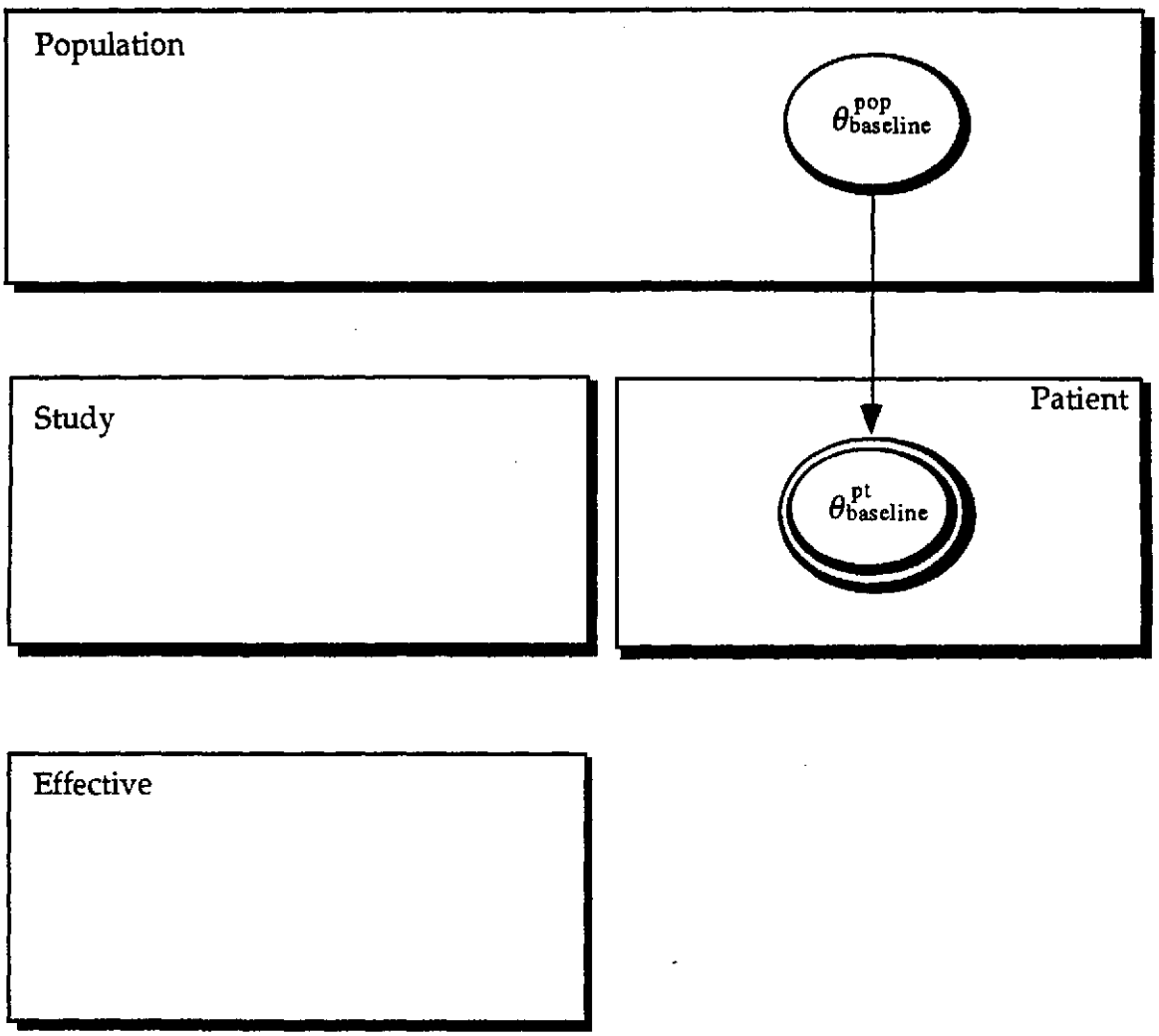


Figure 7.6: THOMAS's initial statistical model. The symbol θ refers to outcome parameters; the superscripts refer to the level of a parameter, and the subscripts refer to the history. Thus, $\theta^{pt}_{baseline}$ is the *patient mortality rate in patients who receive baseline care*. The patient baseline node will not be shown in subsequent figures, because of space considerations.

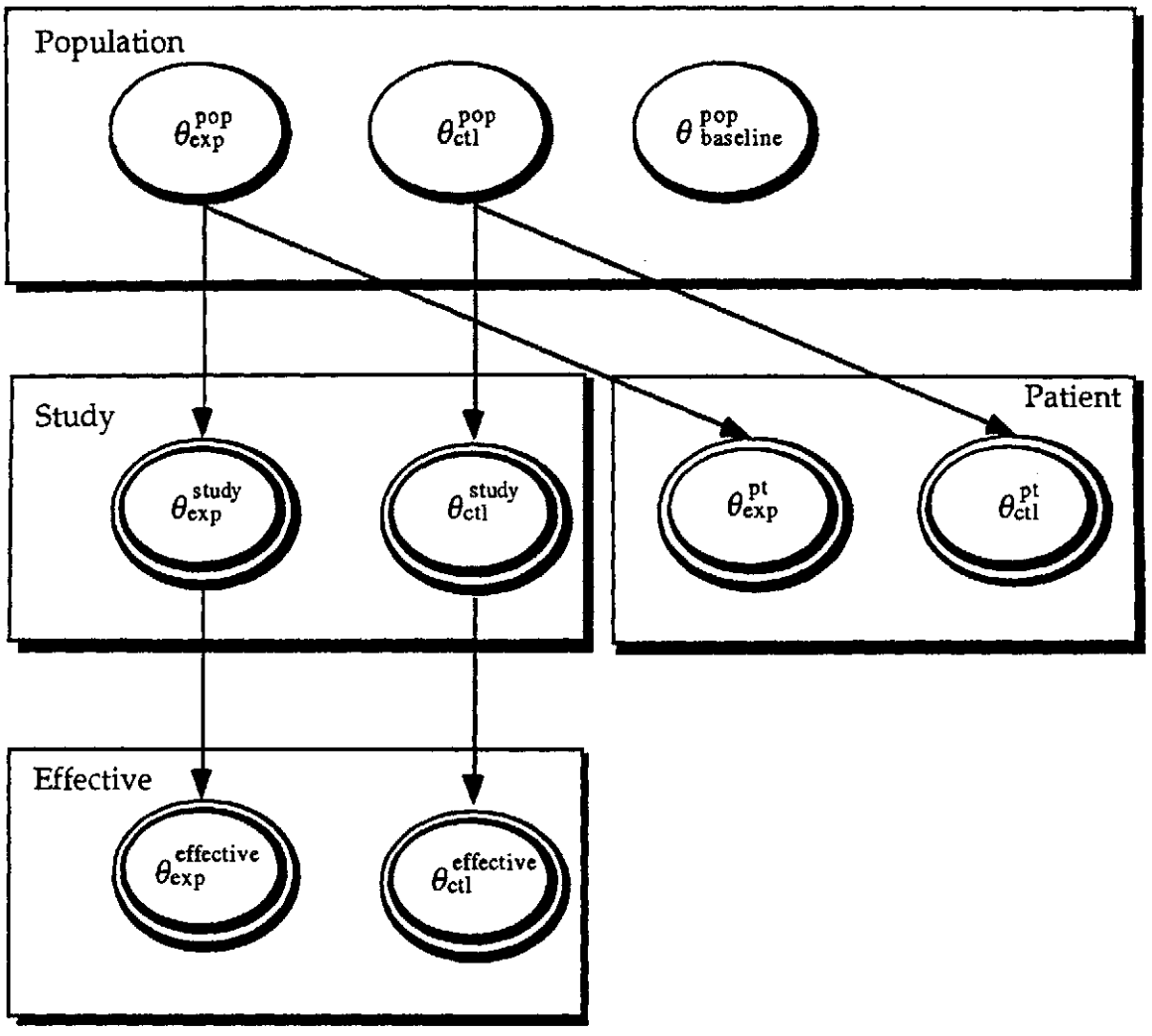


Figure 7.7: THOMAS's statistical model after assignment. Outcome parameters are created for each of the hierarchy's levels, and for each of the two treatments (experimental and control). All functional dependencies are set initially to *identity*, so evidence directly updates population parameters, even though evidence must be dependent on effective parameters (see Section 7.4.4). The patient level will not be shown in subsequent figures, because of space considerations; it is not modified further.

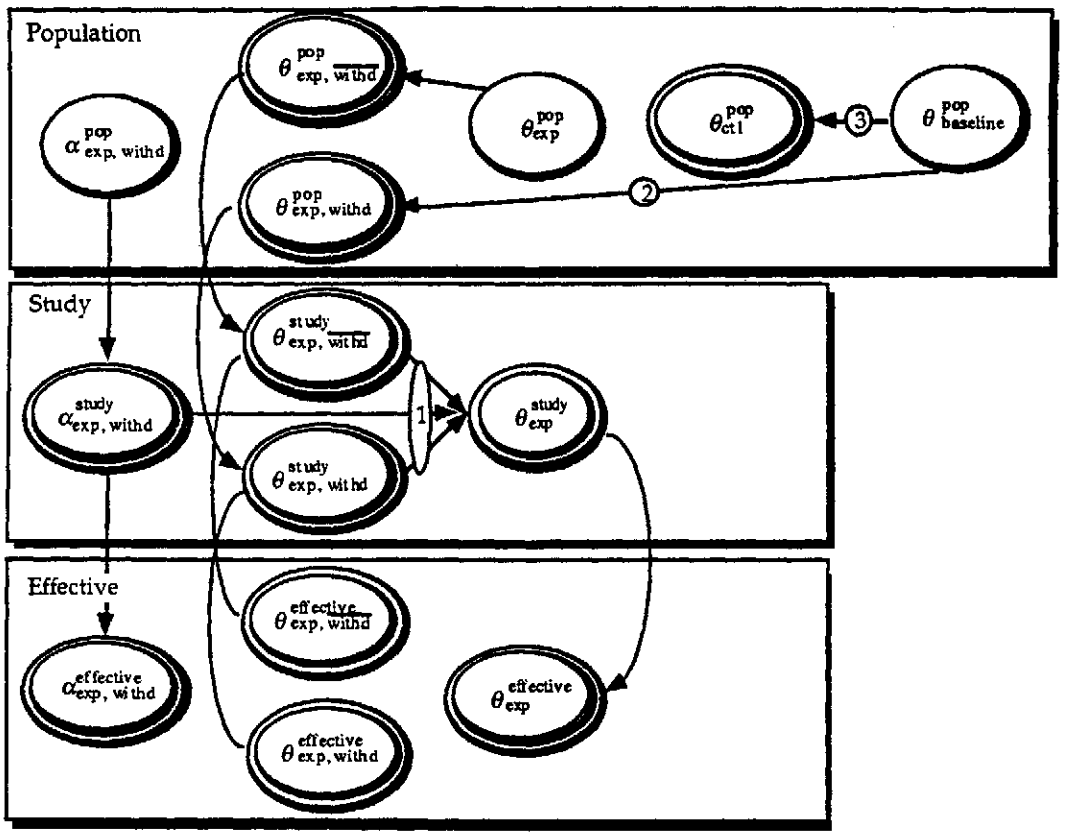


Figure 7.8: Inclusion of withdrawals in THOMAS. Because not all patients assigned to the experimental treatment received the therapy, evidence for the effective parameter does not update the belief in the population parameter directly, as it did in the initial model. The creation of component population and study parameters implements the indirection. The numbered arcs are discussed in the text. The symbol α refers to a subset parameter; withd and $\overline{\text{withd}}$ indicate *withdrew* and *did not withdraw,* respectively. Thus, $\theta^{pop}_{exp,\,\overline{withd}}$ refers to the population outcome parameter in patients assigned to the experimental drug who did not withdraw from therapy.

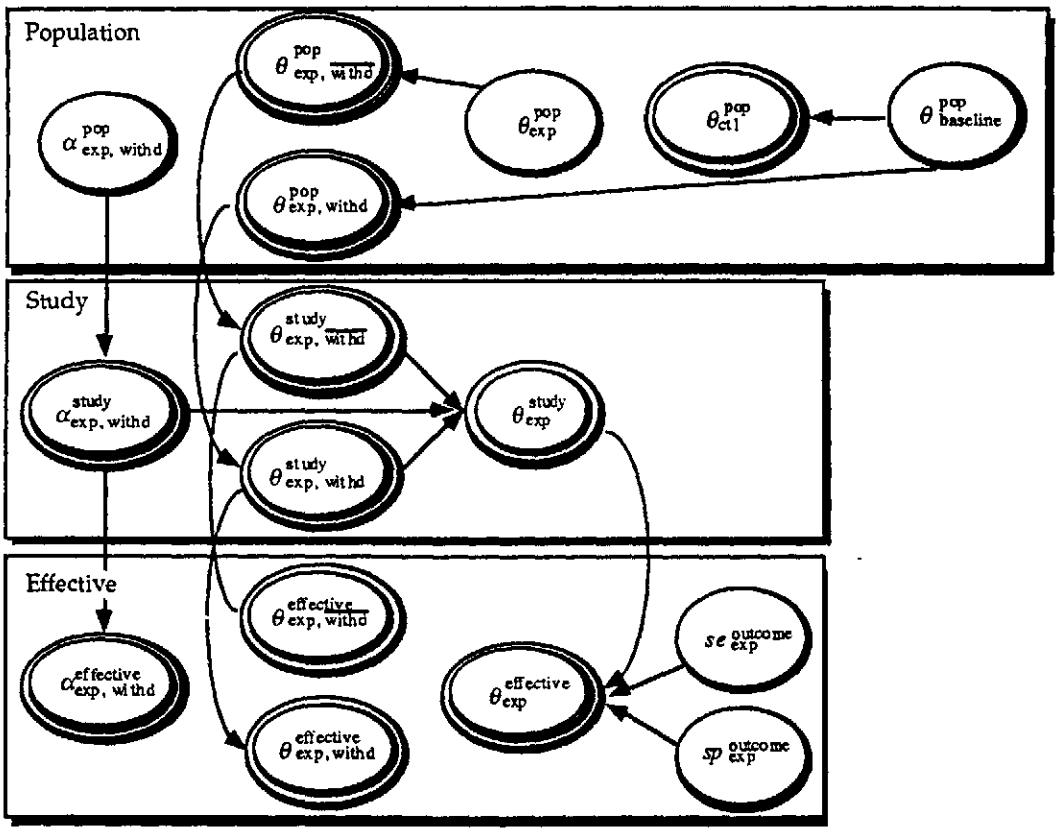


Figure 7.9: Inclusion of classification error in THOMAS. The model of Figure 7.8 is modified only in the effective layer. The symbol *se* refers to a sensitivity parameter; *sp* refers to a specificity parameter. These methodological parameters are indexed by the outcome to which they refer (as the superscript) and by the cohort to which they refer (as the subscript).

7.6 Probabilistic Updating

Once the statistical model is constructed, THOMAS proceeds with probabilistic updating. This process entails computing the posterior probability distributions of all parameters in the statistical model. THOMAS uses Shachter's posterior-mode analysis (Shachter, 1990), a modification of the approach used by Berndt et al. (1974). This approach involves estimating the posterior mean of the multivariate joint distribution of all parameters by the mode of a multivariate normal distribution that approximates the true posterior joint distribution.

The algorithm approximates all distributions as normal distributions. It then searches for the posterior mode via a modified Newton–Raphson steepest ascent in a multivariate space spanned by the basic parameters. Assumptions in the algorithm used are that the basic parameters are marginally independent and that the pieces of evidence are each conditionally independent of one another, given a parameter. To find the posterior mode, the algorithm searches in a direction in multivariate, basic-parameter space given by

$$\mathbf{d}(\mathbf{b}) = -\widehat{\mathbf{H}}^{-1}(\mathbf{b})\nabla_b \mathbf{a}(\mathbf{b}) \tag{7.1}$$

where $\mathbf{b}$ is a vector, of length b, of basic-parameter values, $\mathbf{d}$ is the direction vector (also of length b), $\mathbf{a}(\mathbf{b})$ is the vector, of length a, of values of all parameters, basic and deterministic, $\nabla_b \mathbf{a}(\mathbf{b})$ is the gradient of that vector with respect to the basic parameters, and $\widehat{\mathbf{H}}^{-1}(\mathbf{b})$ is the inverse of an approximation for the Hessian matrix of the posterior joint probability function (of basic parameters). The approximation for the Hessian, of dimension $b \times b$, is

$$\widehat{\mathbf{H}}(\mathbf{b}) = (\nabla_b \mathbf{a}(\mathbf{b}))^{\mathrm{T}}\, \mathbf{K}\, \nabla_b \mathbf{a}(\mathbf{b}), \tag{7.2}$$

where K, the Cramer matrix, of dimension $a \times a$, is

$$\mathbf{K} = \begin{bmatrix} \nabla^2 \mathrm{LP}(\mathbf{b}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \sum_{j=1}^{n} \nabla_{\mathbf{a}} \mathrm{LL}_j(\mathrm{X}_j \mid \mathbf{a}(\mathbf{b})) \left(\nabla_{\mathbf{a}} \mathrm{LL}_j(\mathrm{X}_j \mid \mathbf{a}(\mathbf{b}))\right)^{\mathrm{T}} \tag{7.3}$$

where $\mathrm{LP}(\mathbf{b})$ is the log prior probability of the basic parameters (a b-to-1 function), $\nabla^2 \mathrm{LP}(\mathbf{b})$ is the Hessian of the prior (and whose dimensions are $b \times b$), and the $\mathbf{0}$s are matrices of zeroes used to fill in the dimensions of that Hessian to those of $\mathbf{K}$. The index, n, is the number of pieces of evidence, X_j is the jth piece of evidence, and $\mathrm{LL}_j(\mathrm{X}_j \mid \mathbf{a}(\mathbf{b}))$ is the primary log-likelihood function for each datum (an a-to-1 function); ∇_a represents the gradient with respect to all the parameters.

The search concludes when a set of values of the basic parameters—$\hat{\mathbf{b}}$—maximizes the posterior log likelihood. The posterior mean of the true posterior distribution is then estimated as the value of all the parameters at the calculated maximum,

$$\mathrm{E}(\mathbf{a}(\mathbf{b}) \mid \mathrm{X}) \approx \mathbf{a}\left(\hat{\mathbf{b}}\right), \tag{7.4}$$

where X is the vector of evidence. The variance of the true posterior distribution is estimated as

$$\mathrm{Var}(\mathbf{a}(\mathbf{b}) \mid \mathrm{X}) \approx -\nabla_a \left(\mathbf{a}\left(\hat{\mathbf{b}}\right) \mid \mathrm{X}\right) \hat{\mathrm{H}}^{-1}\left(\hat{\mathbf{b}}\right) \left(\nabla_a \left(\mathbf{a}\left(\hat{\mathbf{b}}\right) \mid \mathrm{X}\right)\right)^{\mathrm{T}}. \tag{7.5}$$

Because, in THOMAS, many of the deterministic parameters are identical to other parameters, THOMAS first strips away those redundant parameters before proceeding with the search, and attaches to the nonredundant parameters any evidence that was dependent on a stripped-away parameter. When the search is completed, the posterior means and variances of the relevant parameters are copied back, along with the original evidence, into the stripped-away deterministic parameters.

Once the posterior means and variances have been found, the parameters for the distributions of the basic parameters can be found. For instance, if a rate is beta

distributed, its α- and β-parameter values can be calculated from the calculated mean m and variance s^2 from the following equations (compare with Equations 4.1):

$$\begin{aligned} \nu &= \tfrac{m(1-m)}{s^2} - 1, \\ \alpha &= \nu m, \\ \beta &= \nu(1-m). \end{aligned} \tag{7.6}$$

If the algorithm fails to converge within a reasonable time limit (or iteration limit), it returns the last point visited in multivariate space. However, the *semantics* of such a failure are that the model is underspecified. Such underspecification arises when there are parameters with diffuse priors and minimal evidence. These conditions are likely to arise, for instance, in classification-error models, where the sensitivity and specificity are low and there is no evidence from the study. Metarule 7.3 works against this eventuality.

7.7 Utility Maximization

Finally, the system arrives at its recommendation by maximizing utility. The general equation for utility maximization was given in Equation 4.4. THOMAS takes advantage of the fact that, for the limited model currently used by the system, one level of expectation computation can be avoided. The utility equation, Equation 5.1, specialized to the case of the exponential distribution for lifespan, is

$$\delta^* = \max_{\delta \in \mathcal{A}}{}^{-1} \int_0^\infty \int_0^1 \left(L - I_{(\delta=\exp)}\Delta\right) P(L \mid \lambda)\, P(\lambda)\, d\lambda\, dL, \tag{7.7}$$

where the $\mathcal{A}$ is the set of decision alternatives (experimental or control) and Δ is the pragmatic difference. If we switch the order of the integrands, then, by virtue of the

expectation of an exponentially distributed variable,

$$\begin{aligned} \delta^* &= \max_{\delta \in \mathcal{A}}{}^{-1} \int_0^1 P(\lambda) \int_0^\infty \left(L - I_{(\delta=\text{exp})}\Delta\right) P(L|\lambda)\, dL\, d\lambda \\ &= \max_{\delta \in \mathcal{A}}{}^{-1} \int_0^1 P(\lambda) \left(\tfrac{1}{\lambda} - I_{(\delta=\text{exp})}\Delta\right) \mathrm{d}\lambda. \end{aligned} \tag{7.8}$$

The chosen alternative is

$$\delta^* = \max{}^{-1} \begin{cases} \left\langle \frac{1}{\lambda} \right\rangle & \text{for control;} \\ \left\langle \frac{1}{\lambda} \right\rangle - \Delta & \text{for experimental.} \end{cases} \tag{7.9}$$

Now, the posterior-probability distribution that is calculated from the probabilistic-updating step is the beta distribution for θ—the timed mortality rate. Therefore, to calculate the utilities under each therapy, the system needs to calculate $-\Delta t \left\langle \frac{1}{\ln(1-\theta)} \right\rangle$. There is no closed form for this calculation; THOMAS calculates the expectation using numerical integration.

7.8 Comments

In addressing the problem of helping relatively naïve users to create influence diagrams, we may come to the following conclusion on the basis of the work presented in this chapter: Such aid is highly domain-dependent, so the sequence and strategy must be crafted carefully, with the domain in mind. The domain of randomized clinical trials is large, encompassing many areas that would be called "domains" in and of themselves (e.g., cardiology), but it remains a particular domain nonetheless.

Other researchers have come to similar conclusions. Goldman (1990) has developed a language for constructing influence diagrams that represent natural-language stories. His construction language is more general than the one described here, in that a user is able to create any arbitrary influence diagram using the primitives of the language. Much of the language is specific, being limited to a circumscribed vocabulary. Yet the language deals, as well, with the domain-dependent problem of relating

symbols—subgraphs of an influence diagram—to a particular class of tasks—natural-language understanding. In particular, he uses a "skeletal belief network" (p. 51) that is a hierarchical and typed influence diagram, as is THOMAS's statistical model. Goldman's levels refer to tasks in natural-language understanding: *plan explanations of plans, plan explanation of words, determine type of word denotation,* and *determine word-type of words.* Furthermore, the response to the user is guided by rules that are triggered by metadata about inputs and that make metadata-type conclusions. For instance, if the input phrase is a prepositional phrase then the system is instructed to establish arcs between nodes representing specific classes of word tokens (Goldman, 1990, p. 90).

Goldman's approach is *model-driven* in the sense that the system has a strong sense of what types of information to expect and how they interact. Other investigators use a model-driven approach as well in automating influence-diagram construction.

Breese's (1987) approach is similar to mine in that the final influence diagram is built from components provided by the system builder. Because his approach is general, Breese requires the user to have more knowledge of influence diagrams.

Another sense of model-driven construction is where the system builder provides the system with an initial large model which the system shrinks during run time. Wellman (1988) and Holtzman (Holtzman, 1989) take this approach in building medical-therapy advisory systems.

There are numerous examples of implicitly typed influence diagrams, like the one that THOMAS uses, in the AI and uncertainty literature. QMR-DT (Shwe et al., to appear) has nodes of the types *disease* and *finding,* with arcs of the type *has.* Pathfinder (Heckerman, 1990) has nodes of the type *distinguished* and *feature,* with arcs (or sets of arcs) of type *subset independence* and *hypothesis-specific independence.* Agosta (Agosta, 1988) has nodes of types corresponding to the hierarchy of objects in vision systems, where the different levels have the semantics of *abstraction* of a

percept, as opposed to THOMAS's semantics of difference between cohorts.

I discussed the Confidence Profile Method of Eddy and colleagues (1991) in Section 4.7. The software distributed with CPM the book—FAST*PRO—allows a user to assemble an evidence table and to derive posterior belief curves. As currently implemented, influence diagrams are not used. Rather, the system uses normal-distribution approximations that allow the program to use adjustment formulae for each arm and to perform the calculations in the evidence-to-parameter direction. As an approximation, this method is useful; in large models, however, it may be unpredictable, as we discussed in Section 7.1.

Because model selection or construction is difficult to effect in the proper Bayesian manner (Section 4.4), the AI approaches used by these investigators are all appropriate. The system designer must simply be careful that an approach is not inconsistent with the Bayesian paradigm.

Chapter 8

Current and Future Status

The claim of this dissertation is that decision analysis and the Bayesian paradigm can form the foundation of a computer-based environment to aid physicians in making clinical decisions on the basis of scientific data from the clinical research literature. THOMAS was created to test this claim. In this chapter, I shall describe, in Section 8.1, the experience physician readers have had in using THOMAS. Then I shall compare, in Section 8.2, the current version of THOMAS to the specifications of decision analysis and the Bayesian paradigm. In Section 8.3, I shall suggest future directions in expanding THOMAS, showing how the framework developed in this dissertation can be used to represent methodological concepts that have, heretofore, not received formal attention.

8.1 Usability

I have evaluated the usability of THOMAS in a semiformal manner, judging the ability of individual users to interact with the system. Three clinicians—a general internist, a general pediatrician, and a pediatric hematologist-oncologist, the first two of whom were familiar with decision theory—and a statistician used THOMAS in the intended

way. Each reader was asked to choose an article reporting an RCT comparing two drugs, with the endpoint of mortality.[1] He perused the article before a session with THOMAS.[2] I ran each session, taking command of the physical interaction with the system, explaining concepts (when THOMAS had no help text for the concept), audiotaping the session, and taking notes. The users read the study reports carefully in response to interrogatives from the program, which directed the analysis. At the end of the session, the users were asked for their reactions to the system, and, particularly, for their criticisms of the concepts and implementation. Each session lasted approximately 45 minutes, most of which was spent in reading the article and in discussing the concepts.

The task was conceptually familiar to each user. The subjects understood the checklist interface ("This looks pretty nice"), but found that navigating it was, at times, confusing. The users found the patient-flow diagram, on the other hand (and as predicted), to be self-evident.

Each user reacted to the request for a decision model with surprise, but reported that the request to be the most useful—and difficult—part of the interaction. All subjects understood the notion of the pragmatic difference, although each found it difficult to make the global assessment, preferring to make the judgments piecemeal, in terms of component objectives.

All users felt comfortable specifying prior beliefs, and recognized the difference between making the assumption of "total ignorance" and specifying such belief. Two users would have preferred to make the experimental mortality rate dependent on the control mortality rate and a relative risk, instead of making the two mortality rates

[1]The reports analyzed were the metoprolol study by Hjalmarson, et al. (1981), a study testing the efficacy of sclerotherapy in patients with esophageal varices (Veteran Affairs, 1991), and a study comparing aspirin and sulfinpyrazone in the prevention of stroke (Canadian Cooperative Study Group, 1978). Note that the second study evaluated a nondrug treatment and that the third study evaluated an outcome other than mortality. These articles were used at the users' requests.

[2]The program was implemented on a Macintosh IIfx with 8 MB of RAM for these sessions.

independent.

The users thought that the system's use of life expectancy as the basis for its recommendation was intuitive. It was much easier to understand than the Bayesian graphical equivalent of the p value (see page 108).

When asked specifically what deficits they perceived in the program, the two users knowledgeable in decision theory stated that they thought that the system would be too difficult for use by clinicians who had not had theoretical training in the decision sciences. The single "naïve" user, on the other hand, spotted all the statistical assumptions and limitations of the program. She noted problems such as the questionable propriety of the constant-hazard probability model for lifespan. She also found frustrating the inability to use the system to specify the role of patients' pre-existing conditions in possibly confounding the conclusion (see Section 8.3.3).[3] Thus, the experience with the naïve user suggests that an academic clinician has precisely the level of methodological knowledge necessary to use THOMAS, and that THOMAS is written at the proper level of abstraction.

The statistician found the process statistically sound, given the constraining assumptions.

We should note that this evaluation is limited in that it demonstrates that the system *could* function as intended. Assessing the system's use by a wide range of clinicians, or determining its potential impact on clinical care, must await future work.

8.2 Satisfaction of Specifications

The internal validity of THOMAS depends on the system's meeting all the specifications we have set for the program. There are four sets of specifications: (1) the

[3]She was also frustrated by the lack of data in the study report to furnish the details of such a model, had it been available.

claims made in the thesis statement (Section 1.5); (2) , the claims of THOMAS's proposed behaviors (Section 1.6.1); (3) the high-level desiderata for solving the literature problem based on a knowledge-level analysis of the problem (Section 2.3); and (4) the specifications implied by the strategy inherent in the decision-analytic approach (Section 4.4). Points 3 and 4 were evaluated in Section 4.6. In Sections 8.2.1 and 8.2.2, we shall briefly examine points 1 and 2.

Internal validity also demands demonstration that the system can produce a rich array of meaningful statistical models. Such a demonstration shall be made in Section 8.2.3.

8.2.1 Thesis Statement

We recall the thesis statement: Decision analysis and the Bayesian paradigm can form the basis of a computer-based environment to aid physicians making clinical decisions on the basis of scientific data from the clinical research literature. The system's use of a utility model to make recommendations satisfies the conditions of the *clinical focus of decision making* and the *applicability of decision analysis.* THOMAS's interface being designed for the physician user, satisfies the stipulation of the *physician reader as the target user.* The nature of the checklist, the patient-flow diagram, and the knowledge assumed of the user all depend on the physician as user. THOMAS is a *dynamic computer-based environment*, and assumes that *published scientific data are the primary source.* Finally, the system *extends the Bayesian paradigm* by having the user express her prior belief about model structure, not just parameter values. Thus, the system fulfills the desiderata.

8.2.2 Behavior-Based Desiderata

Section 1.6.1 presented more specific design goals. The checklist implements the requirement *to analyze a study in a structured way.* The ability to examine a series of analyses (Section 6.4.4) fulfills the requirement *to examine a study in multiple ways* and *to examine the sensitivity of any posterior belief or decision to different prior beliefs.* The use of prior belief (Section 6.3.3) allows the system *to incorporate domain knowledge into an analysis.* The construction steps of the metadata-state diagram (Section 7.4) enable the system *to incorporate methodological knowledge into an analysis.* THOMAS's probability plots (Section 6.4.2) allow the user *to examine the change in belief in any parameter* and *to compare the beliefs in any two parameters,* whereas the plots of life expectancy (Section 6.4.1) enable to system *to determine the optimal therapy.*

8.2.3 Variety of Models

To demonstrate THOMAS's ability to build, assess, and evaluate a wide range of models, I performed a modified sensitivity analysis along a number of dimensions. The model dimensions examined are prior belief, certainty of the data, single protocol departure, two protocol departures, and classification error. Table 8.1 summarizes the descriptions of the models. Model *Baseline* includes the raw data from the metoprolol study and the withdrawals from both treatment groups. The next three models vary some of the values for evidence or prior belief. Model *RealPrior* uses for the population parameters a prior belief stated by two of the users in the usability evaluation. The model *LowCert* examines the effect of fewer data than were in the metoprolol study (an imagined total of 75 patients, rather than 1595 patients). Model *LowWD* examines the effect of low withdrawal rates.

The next four models evaluate the effect of enlarging the *structure* of model *Baseline*. Model *GoodCE* adds classification error for patients in the metoprolol cohort, with a prior belief in the sensitivity of 0.94 and a prior belief in the specificity of 0.94. Model *BadCE* has a prior belief in the sensitivity of 0.85, lower than in model *GoodCE*. Model *WDNC* adds noncompliance to both treatment arms. Model *WDNCCE* adds classification error for metoprolol deaths to model *WDNC*, with a prior belief in the sensitivity of 0.94.

Table 8.1: Description of the sensitivity-analysis models.

Model Label	Prior[1] Belief	Mortality Rate[2]	Cohort Size[3]	Protocol Departure[4]	Classification Error[5]
Baseline	0.5, 0.5	0.05	697	0.2	—[6]
RealPrior	10, 115	0.05	697	0.2	—
LowCert	0.5, 0.5	0.05	100	0.2	—
LowWD	0.5, 0.5	0.05	697	0.05	—
GoodCE	0.5, 0.5	0.05	697	0.2	0.94,0.94
BadCE	0.5, 0.5	0.05	697	0.2	0.85,0.94
WDNC	0.5, 0.5	0.05	697	0.2, 0.2	—
WDNCCE	0.5, 0.5	0.05	697	0.2, 0.2	0.94,0.94

[1] Belief in the parameters of the prior beta distribution for the two mortality rates.
[2] Observed mortality rate in patients assigned to metoprolol.
[3] Number of patients assigned to metoprolol.
[4] Observed rate of withdrawal from both metoprolol and placebo cohorts.
[5] Prior belief in sensitivity and specificity, respectively.
[6] Not applicable.

Table 8.2 summarizes THOMAS's structuring of these models and its performance in evaluating them. The basic parameters are the marginally independent variables in each model. The identity parameters are the number of parameters identically deterministic to some other parameter in the model and reflect the overhead of model construction in THOMAS. The ratio of the number of identity parameters to the total number of parameters gives a sense of that overhead. The listed performance figures

indicate how long the probabilistic-update step takes.

We can draw several conclusions on the basis of the information in Table 8.2. First, the model-construction overhead is high, with up to one-half of the structure not contributing to the calculations or to the system's recommendation. However, the large number of parameters currently superfluous gives us a sense of the number of methodological concerns that could be included in the model, if the user so wished. Second, the performance is a function of both the number of parameters in the model, and the degree of posterior certainty in the model. The degradation of performance with increase in model size is obvious; the relationship to posterior certainty is less apparent. If the slower performance were due just to the increase in the number of parameters, then the degradation from model *Baseline* to model *GoodCE* would be similar to the degradation from model *WDNC* to *WDNCCE*. Yet we find that model *WDNCCE* has a disproportional increase in execution time. The slower performance is due to the broadening of the posterior mode of the posterior joint distribution in the latter model; the broader the mode, the longer the posterior-mode-analysis algorithm takes to converge to a single answer.

Table 8.3 lists the posterior beliefs calculated by THOMAS for each model. We can draw some conclusions from these values.

1. The posterior mean for the mortality rate in patients treated with metoprolol is intermediate between the observed mortality rate in the metoprolol cohort and the mortality rate in patients treated with placebo. This behavior was codified by Rennels (1986) as a heuristic. We find, in contrast, that THOMAS implicitly *derives* this heuristic from its knowledge of statistical theory.

2. Because the posterior beliefs for model *Baseline* and model *RealPrior* are the same, we see that adding a prior belief that is less certain than the data does not necessarily affect the posterior beliefs.

Table 8.2: Sensitivity-analysis results.

Model Label	Basic[1]	Deter-ministic	Identity	Total	Overhead[2]	Performance[3] (sec)
Baseline	5	3	17	25	0.68	26.08
RealPrior	5	3	17	25	0.68	27.66
LowCert	5	3	17	25	0.68	23.48
LowWD	5	3	17	25	0.68	24.65
GoodCE	7	4	16	27	0.59	37.23
BadCE	7	4	16	27	0.59	40.6
WDNC	9	10	24	43	0.56	140.95
WDNCCE	11	12	22	45	0.49	224.5

[1] Number of basic parameters.

[2] Ratio of number of identity parameters to total number of parameters.

[3] Tested on a Macintosh IIsi with 5MB RAM, a math coprocessor, and System 6.0.7.

3. We would expect that less certain evidence, as in model *LowCert*, would lead to less certain conclusions; we see, however, that the posterior beliefs, in model *LowCert* in the mortality rates each has a *lower* variance. The reason for this conflict with intuition is that the posterior means for model *LowCert* are lower than in model *Baseline*. The more extreme rate parameters are, the smaller their variances will be, simply because we *know* that they cannot take on lower values (when close to zero).

4. Model *LowWD* gives a more certain posterior mean to the two mortality rates of interest. This behavior is what we should expect: The more certain we are that patients actually received the medication to which they were assigned, the more certain we should be about the inference about the medication's effects.

5. The improvement in uncertainty in model *GoodCE* over model *Baseline* has the same reason as in model *LowCert*: The posterior mean for metoprolol (the cohort potentially misclassified) is closer to zero than in model *Baseline*.

Table 8.3: Sensitivity-analysis posterior values.

Model Label	Evidence: Metoprolol Mean	Evidence: Metoprolol (Var[2])	Evidence: Placebo Mean	Evidence: Placebo (Var)	Posterior Belief[1]: θ^{pop}_{met} Mean	θ^{pop}_{met} (Var)	θ^{pop}_{plac} Mean	θ^{pop}_{plac} (Var)
Baseline[3]	0.057	(0.7776)	0.089	(2.32)	0.0449	(10.5)	0.0833	(11.4)
RealPrior[4]	0.057	(0.7776)	0.089	(2.32)	0.0449	(10.5)	00833	(11.4)
LowCert	0.020	(3.92)	0.040	(15.4)	0.0211	(4.57)	0.0438	(8.51)
LowWD	0.057	(0.7776)	0.089	(2.32)	0.047	(4.08)	0.0784	(4.57)
GoodCE	0.057	(0.7776)	0.089	(2.32)	0.0143	(2.79)	0.0981	(3.97)
BadCE	0.057	(0.7776)	0.089	(2.32)	0.0945	(1.54)	0.102	(3.03)
WDNC	0.057	(0.7776)	0.089	(2.32)	0.0584	(19.2)	0.113	(33.3)
WDNCCE	0.057	(0.7776)	0.089	(2.32)	0.0177	(4.33)	0.141	(48.6)

[1] The parameter symbols follow the convention in this dissertation: θ^{pop}_{met} is the population parameter for patients treated with metoprolol.
[2] Variance $\times 10^4$.
[3] The prior belief in each of the parameters θ^{pop}_{met} and θ^{pop}_{plac} in all models (except in model *RealPrior*) is a mean of 0.50, with a variance of 1250×10^4, a very uncertain belief.
[4] The prior belief in each of the parameters θ^{pop}_{met} and θ^{pop}_{plac} in this model is 0.10 (6.89), a belief whose mean is closer to the posterior and more certain than that of the other models.

6. The even greater improvement in model *BadCE* is unclear.

7. The degradation in posterior certainty due to the incorporation of a second protocol departure indicates that the model's uncertainty is much more sensitive to protocol departures than it is to classification error. This sensitivity is a result of the system's interpolation of mortality rates for the unobserved subcohorts.

8. The further dissipation of certainty in θ^{pop}_{met} is the result of the classification error applying to only the metoprolol cohort.

Thus, THOMAS can build and evaluate a variety of statistical models. An explanation facility for the system's output would be helpful, and is an area for future research.

8.3 Representational Richness

Throughout this dissertation, I have alluded to the ability of the framework to represent methodological issues beyond those of protocol departures and measurement reliability (see, for instance, the introduction to Chapter 5). In this section, I shall show how issues such as correlated prior belief, baseline characteristics, and randomization. The implementation of these issues is left to future work.

8.3.1 Correlated Prior Belief

THOMAS's current model assumes that the control-cohort and experimental-cohort mortality rates are independent. Yet, often, an analyst or a reader believes that the two rates are related to each other—for instance, through a relative risk. This state of affairs is depicted in Figure 8.1. Note that the model adds structure to the relationship between the outcome parameters at the population level. This model allows the user to express how much she expects the experimental drug to improve mortality as compared to the control drug, regardless of the actual value of the latter rate, rather than to specify each expected mortality rate separately.

8.3.2 Component Effects

A second set of models allows the user to specify that observed outcomes are the result of component effects, much as in models for analysis of variance (ANOVA) and for linear regression. In this section, we examine the use of this approach.

Study observations always reflect interventions beyond the treatments of interest alone; Figure 8.2 presents a model for this state of affairs. In this model, the study parameters are each the sum of two component parameters: the population parameter of interest and the nuisance study parameter that reflects the contribution of ancillary factors in the particular study. In the case of rate outcomes, it is inappropriate to

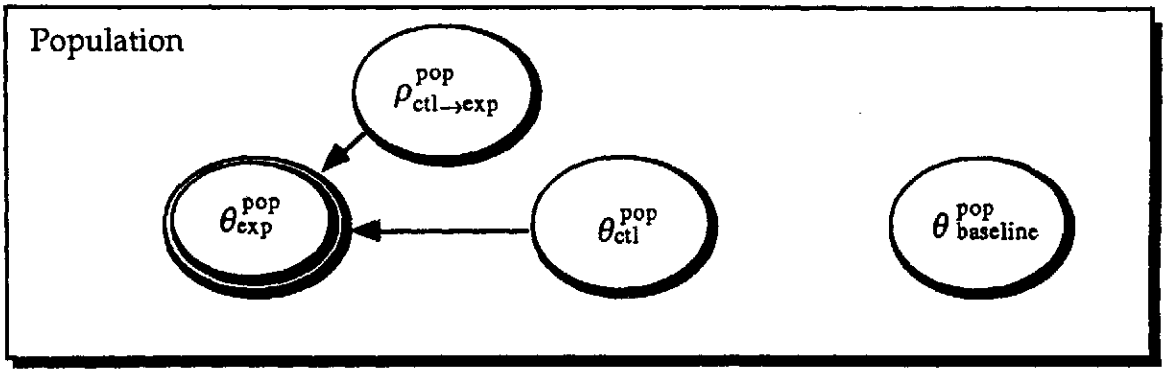


Figure 8.1: Correlated prior beliefs. This influence diagram depicts the belief that the experimental mortality rate is equal to the product of the control mortality rate and a relative risk: $\theta^{pop}_{exp} = \rho^{pop}_{ctl \rightarrow exp} \theta^{pop}_{ctl}$

model a rate as the sum of other rates, because the sum may be greater than 1. Thus, statisticians routinely transform rates into log odds: $\omega = \ln \frac{\theta}{1-\theta}$. Log odds range over the entire number line and have the semantics of *component risks.* We can obtain a rate from log odds by the formula $\theta = \frac{e^{\omega}}{1+e^{\omega}}$. Thus, an analyst can specify a prior belief for either entity, and derive a distribution for the other. I shall use log odds in the ensuing models.

Therefore, in the case of rate outcomes, the component model is $\omega^{study}_{exp} = \omega^{pop}_{exp} + \omega^{study}_{ancillary}$, or, $\omega^{pop}_{exp} = \omega^{study}_{exp} - \omega^{study}_{ancillary}$; that is, the population parameter of interest is the excess risk the experimental drug has over the risk embodied in the details of the study execution. This risk is the quantity most investigators presume to be invariant across studies, and therefore belongs in the population level. For a specific patient, the excess risk is added to the baseline parameter for that patient, much as in a log-odds model.

In the case of placebo studies, ω^{pop}_{ctl} would be considered to offer no benefit beyond the care received in the course of the study,[4] so it would not be included in the model.

[4]The placebo effect is a well-known phenomenon, but this effect is identical to that of ancillary treatment given *in the course of a known study.* This effect is common to both treatment cohorts.

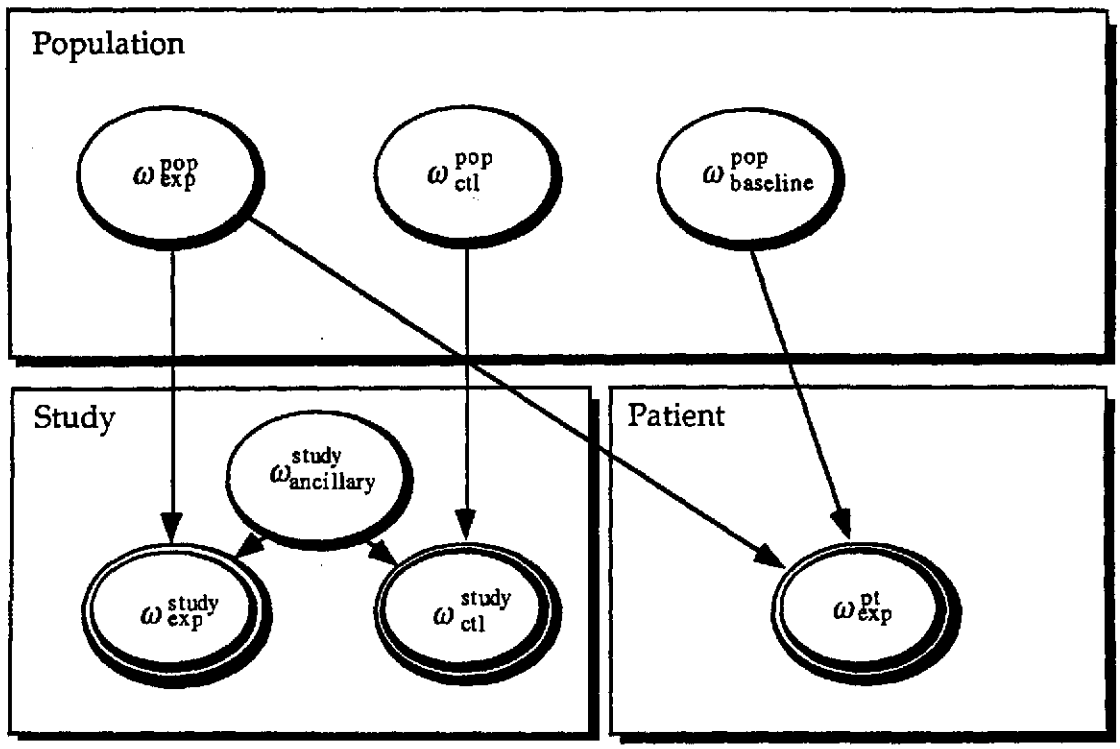


Figure 8.2: Basic component model. The study parameters are sums of a within-study component ($\omega^{study}_{ancillary}$) and a population (between-study) component (ω^{pop}_{exp} or ω^{pop}_{ctl}). The patient parameter is the sum of the baseline and the treatment-specific components.

Therefore, if evidence were available for both ω^{study}_{exp} and ω^{study}_{ctl}, there would be enough evidence to update belief in the population experimental parameter.

If the control treatment were thought to have a benefit beyond the ancillary therapy, the analyst would face the prospect of having three parameters, but only two sources of data (see Metarule 7.3). Such a situation results in uncertain posterior means for all three parameters. To arrive at a moderately certain conclusion for the experimental-treatment effect, the analyst would require either more evidence, or more prior information; otherwise, the posterior belief in ω^{pop}_{exp} would be uncertain. Alternatively, if a second study of the two drugs were available, there *would* be enough

evidence to supply meaningful posterior beliefs on the population parameters. This state of affairs is shown in Figure 8.3.

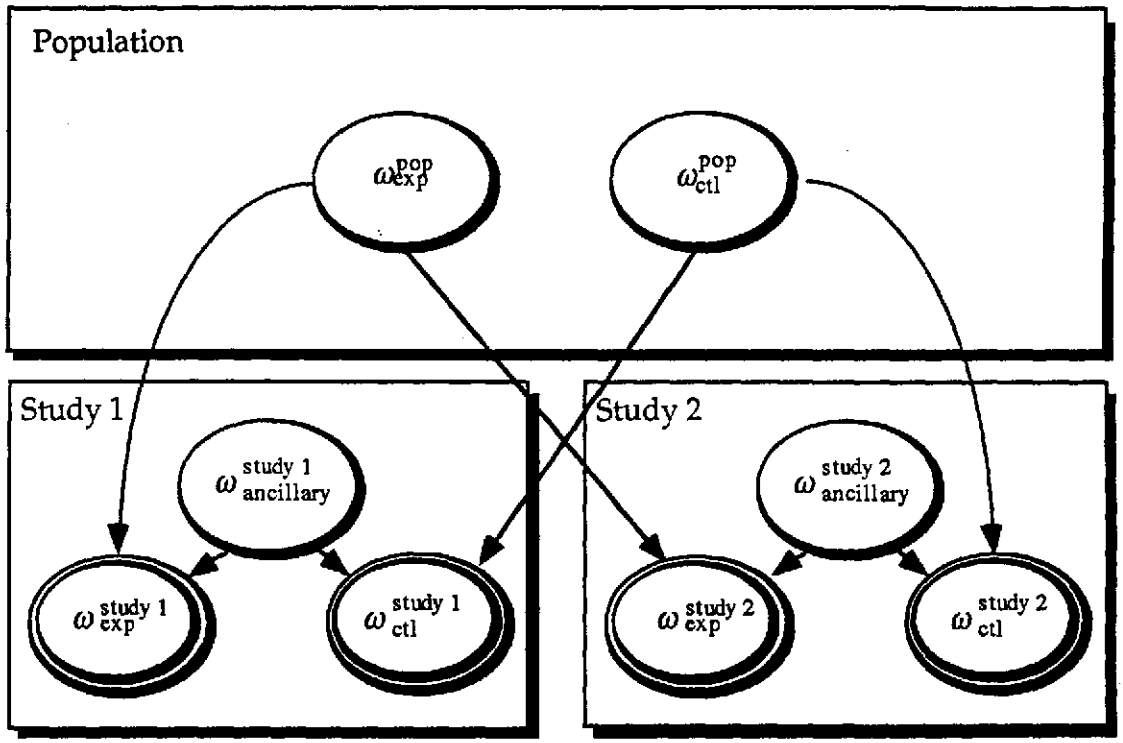


Figure 8.3: The use of two studies. If the study parameters are modeled as sums of population parameters and within-study effects, then, with two studies, the number of basic parameters is equal to the potential number of pieces of evidence. The population baseline parameter is omitted in the interests of space.

In any of these models, the population baseline parameter can play a number of roles. As in THOMAS's models, it may stand in for the default natural history of the disease. In this case, the population parameters supply the addends or diminuends that modify the natural course of an illness. Or, we can model $\omega^{study}_{ancillary}$ as explicitly dependent on the baseline population effect. In this case, any difference between the

ancillary and baseline parameters has the semantics of the effect of study participation.

8.3.3 Baseline Characteristics

Investigators and clinicians often are interested in knowing which patient characteristics predict the patient's course of health or the patient's response to therapy. In either case, we model the study outcome as dependent on parameters representing patient characteristics. The form is the same as in modeling the effect of ancillary therapy. For instance, the contribution of disease severity in predicting prognosis may be to *add* a constant amount to the mortality rate. However, the semantics of these models is not the same as those of ancillary-therapy models, because ancillary therapy represents what the medical community is doing to the patients, not what the patients bring to the study. Figure 8.4 gives an example of modeling severity. A characteristic, such as severity, is called a *covariate*. The model shown applies to dichotomous covariates, and introduces a new parameter, γ, which I shall call a *gamma* parameter. The $\gamma_{\text{cohort}}^{\text{covariate}}$ parameters are defined as *the proportion of patients within the* cohort *who have the stated* covariate. The algebraic model is $\omega_{\text{exp}}^{\text{study}} = \omega_{\text{exp}}^{\text{pop}} + \gamma_{\text{exp}}^{\text{severity}} \omega_{\text{severity}}^{\text{study}}$. The gamma parameters do *not* have the same semantics as coefficients in a regression model; rather, they attenuate the ω parameters, which do have those semantics. If the analyst wishes to represent a series of covariates, he may use a like number of gamma and log-odd parameters.

A major difficulty with baseline-characteristic models is that they require individual data on each patient; these analyses are akin to logistic regression. Thus, studies, as currently reported, would be inadequate for the task of providing an analyst–reader with the information needed to make a rational decision in this context. However, with such data, a number of important methodological concerns can be represented.

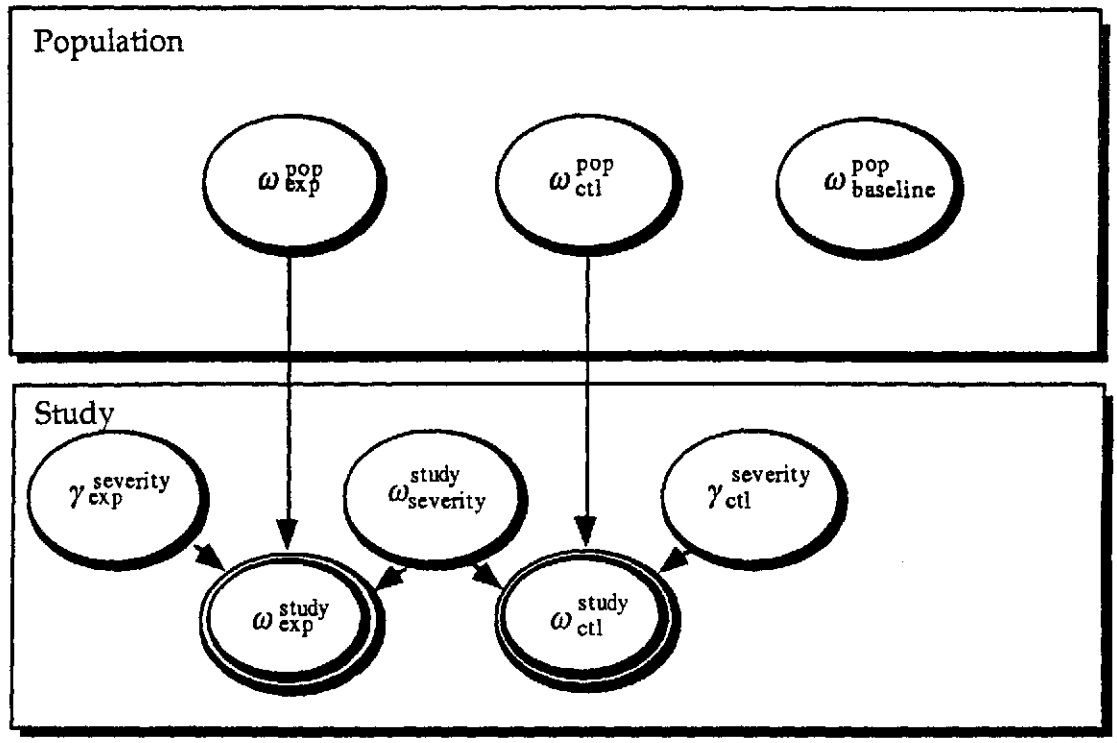


Figure 8.4: Modeling baseline characteristics. The $\gamma^{covariate}_{cohort}$ parameters are defined as the proportion of patients within the *cohort* who have the stated *covariate*.

In particular, we can represent the issues of *asymmetric cohorts*. These issues comprise problems of randomization, of masking, and of asymmetric protocol departure. I shall describe only the use of asymmetric cohorts in modeling randomization.

8.3.4 Randomization

The goals of randomization are to prevent cheating, so that the investigator does not know the prospective assigned treatment before recruiting a patient into the study, and to make the treatment groups comparable, so that differences in outcomes can be attributed to the differences in the treatments, and not to differences in group

composition.[5] There are numerous sources of information that the reader can use to assess the effectiveness of the randomization and the equivalence of group composition, including the method of randomization, the identity of the investigators, and the table of baseline characteristics describing the treatment groups (usually "Table 1" in reports of RCTs). Our model of randomization must incorporate all these components.

Figure 8.5 shows one possible model: it demonstrates a use of *hierarchical modeling,* a methodological modeling technique that allows representation of methodological concerns, as does likelihood debiasing (see page 95). Specifically, our beliefs in the gamma parameters are parameterized as beta distributions, with different parameters for the experimental and control cohorts. In general, parameters that parameterize belief in other parameters are called *hyperparameters.* I have modeled the hyperparameters for $\gamma_{\text{exp}}^{\text{severity}}$ as uncertain, but the hyperparameters for the control-cohort gamma parameter as certain, if we know the value of the experimental-cohort gamma parameter and the values of two new methodological parameters, $\Delta\alpha$ and $\Delta\beta$, which I shall call the *delta* parameters.

The delta parameters represent the degree to which the distributions of the two gamma parameters are the same. If they are the same (both delta parameters being zero), then we will conclude that the two cohorts were randomized the same way; if they are different, we will conclude otherwise. The definition of the criterion for being the *same* may be similar to the probability criterion used in the Bayesian hypothesis test (see Figure 4.11). Furthermore, the delta parameters give us a location to represent in a prior-probability distribution subjective notions such as the expected quality of randomization, due, for instance, to the reader's knowledge about the investigators' integrity. The delta parameters may also provide the site for updating

[5] In classical statistics, randomization is a prerequisite for the use of classical-statistical tests, as well.

belief about randomization, given qualitative statements in the report, such as the method of treatment assignment.

The gamma parameters, in contrast, provide a location to incorporate observed data, such as the table of baseline characteristics of patients in the study.

Use of the model in interpreting a written report proceeds as follows. First, as in other additive models, we assess our knowledge about the population-level parameters. This knowledge is domain dependent.

Next, we assess our prior belief in $[\alpha, \beta]_{\text{exp}}^{\text{severity}}$. We might consider the number of patients with severe disease expected *from the outset of the study* to be admitted to the study on the basis of our knowledge of the domain and of the projected sample size stated by the authors. We also consider the method of randomization. Thus, if the investigators expected to enroll 200 patients, of whom 50 would be expected to have severe disease on the basis of our knowledge of the disease in question, and the patients were supposed to be completely randomized (one-half to each treatment group), then the prior belief in $\gamma_{\text{exp}}^{\text{severity}}$ would be $\mathcal{BE}(25, 75)$; The α-parameter value of 25 represents the mean of our prior belief in $\alpha_{\text{exp}}^{\text{severity}}$. Because this parameter is bounded on one side, we might represent our belief in this parameter with a log-normal distribution. The mean of that distribution would be 25, and the variance would depend on how certain we are *before getting details* that the randomization protocol was executed as described by the authors; this variance measures our trust in these particular authors.

Next, we consider whether we think there might have been uneven assignment: Would patients be assigned preferentially to the control group? If so, we would set the mean of our beliefs in $\Delta\alpha$ and $\Delta\beta$ to numbers reflecting the expected deviation from assignment protocol. Because the differences are potentially unbounded, we could represent the belief in each difference with a normal distribution. Again, the variance of the distribution reflects our trust in the investigators.

Finally, we incorporate evidence from the study. The distribution of severity status in patients actually enrolled in the study updates belief in the gamma parameters, whereas the patient outcomes update belief in the log-odds parameters. With the prior beliefs and evidence collected, we can generate the posterior beliefs in the seven basic parameters.

This model, then, represents the different aspects of randomization that I raised as concerns at the beginning of this section. The power of the model is that it separates a number of functions that have traditionally been lumped together in classical analyses. Specifically, the hypothesis-testing analysis of the table of baseline characteristics has borne the burden of verifying the following: (1) that the patients were indeed randomized as the authors claimed, (2) that the distribution of baseline characteristics between the two treatment cohorts was the same, and (3) that the distribution of characteristics was a fair portrayal of the population of patients. The model for randomization presented here separates these issues. First, the degree of randomization is assessed by examining the posterior-probability distribution for the delta parameters. Second, the distribution of baseline characteristics does *not* need to be identical for the analysis to proceed. The model does require, however, data or prior beliefs in the log-odds parameters, if the gamma parameters are not close enough. Third, the representativeness of the study cohorts can be embodied in an additive model between the population and study levels; I shall not discuss this model here.

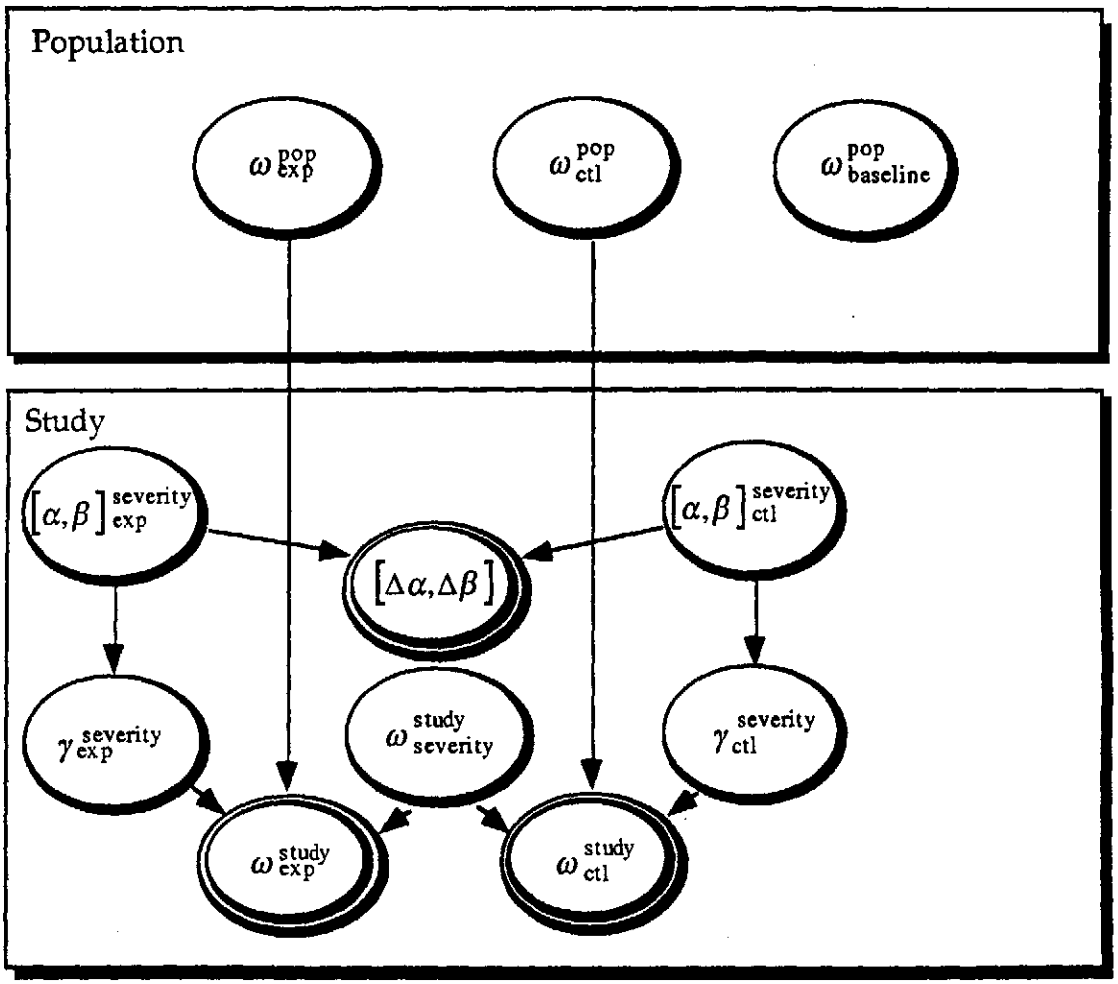


Figure 8.5: Randomization. This model represents the belief in the gamma parameters as a hierarchical model: The belief in the parameter is further parameterized by a beta distribution, about whose parameters we may be uncertain. The uncertain beta *hyperparameters* are denoted by the vectors, $[\alpha,\beta]^{severity}_{exp}$ and $[\alpha,\beta]^{severity}_{ctl}$. The differences between the experimental and control hyperparameters are the delta parameters, $[\Delta\alpha,\Delta\beta]$. If we were to find that the latter two parameters were zero—analagously to the Bayesian hypothesis test in Figure 4.11—then we could conclude that the control cohort and the experimental cohort were randomly composed in the same way. Evidence is typically available for the two gamma parameters and the study mortality rates.

Chapter 9

Conclusion

In this final chapter, I shall answer potential criticisms of this work and shall suggest future directions. Criticisms regarding internal validity are discussed in Section 9.1, and those concerning external validity are covered in Section 9.2. I shall discuss the contributions of this dissertation in Section 9.3, and I shall outline planned future work in Section 9.4. I shall close the dissertation, in Section 9.5, with final comments.

9.1 Internal Validity

The question of internal validity asks whether the thesis is defensible on its own terms. One minor methodological issue is the propriety of acquiring knowledge from a biostatistician in this project, where physicians are the goal users (see Section 2.1). In response, we note that the core knowledge garnered from Professor Brown is consistent with such knowledge found in many other sources of information for this domain (see Section 2.2.3.2).

A more forceful criticism is that THOMAS may be too difficult to use (Section 8.1). There are three components to the difficulty: the physical appearance, the need for primary data, and the conceptual complexity of the system. The thesis depends

least on the physical appearance of the interface, which can be improved. What I have demonstrated with three physician subjects is that the *conceptual* interface has meaning for clinicians: Physicians can create proper statistical models without having to know mathematical statistics.

THOMAS's need for primary data derives from the fact that the proper analysis of survival data requires the history of each patient and not a summary statistic (such as the rate of deaths in the cohort of patients assigned to a treatment). This is a problem as much with the published literature—which does not present this information to the reader—as with the computer artifact. A future version of THOMAS should be able to read automatically the data directly from an electronic article that would contain the needed data. Yet, the necessity will always remain for the physician manually to direct the system, deciding *which* data should be read in and assigning the proper *label* to a datum.

As for the conceptual complexity, the experience with clinician users has suggested that physicians can relate to the novel concepts inherent in the approach. However, the approach clearly needs the ability to explain its results to the user. It is not enough to say that the life expectancy is greater for patients who receive one drug, or to say that the posterior belief in one mortality rate has a mean that is lower than the posterior belief in another mortality rate. The system must answer questions such as these: Which data contributed most heavily to the differences? What model components led to the calculated uncertainty in the posterior beliefs? If the system cannot arrive at a single answer, what model components are responsible? (See page 167.) Currently, a statistician reviewing an analysis can give the answers to these questions; it is not clear how to use the influence-diagram representation to provide the same answers.

9.2 External Validity

There are two classes of concerns regarding the external validity of the thesis. First, would the approach scale up? Second, would the biomedical community use the approach?

9.2.1 Scaling Up

For the approach of this dissertation to scale up for larger problems, an implemented system must be able to deal with larger statistical models, and the system must be able to construct more complex models.

In terms of calculation, posterior-mode analysis is a matrix-based algorithm that calculates approximate solutions. The performance bounds, therefore, should be polynomial. In the discussion related to Table 8.2, we saw, however, that performance also depends on the shape of the posterior distribution. Better algorithms are needed.

In terms of complexity, there are two questions. Can the framework be enlarged to include more complicated concepts? Section 8.3 has answered this question in the affirmative. Can physicians use more complex systems? The resolution of this question depends on system builders' abilities to build different conceptual interfaces to different methodological issues and to enable systems to access primary data.

9.2.2 Use by the Biomedical Community

Two major obstacles will deter the biomedical community from putting the conclusions of this dissertation into use. One is the concern with cheating (Hamaker, 1977), another is the possible superfluity of the approach, and the third is the novelty of the approach (Weaver, 1991).

The bottom-line argument against the use of Bayesian methods in interpreting clinical research data is that it is too easy for users to claim a prior belief that results

in a desired posterior belief. Thus, if a given reader wants the data to support a conclusion she has, she can simply fabricate the prior belief that, combined with the primary likelihood function for the data from the study, results in the supporting posterior belief.

One response to the concern regarding cheating is that a system such as THOMAS forces a user to maintain an *audit trail* of her beliefs (see Section 4.6.2). A second response is that cheating can occur, regardless of our probabilistic philosophy. In classical statistics, the process is vulnerable during the choice of statistical model,[1] and during the combination of statistical-test results with prior beliefs, where, because there is no quantification of the prior belief and because there are no commonly accepted rules for combining qualitative prior belief with p values, the posterior conclusion is relatively arbitrary (see Section 3.5.1).

A second obstacle suggests itself from our examination of Table 8.3, where we found that real prior beliefs potentially have little impact on conclusions of the analysis. We may gather from this observation that the entire enterprise of worrying about prior beliefs is superfluous and unnecessary. There are two answers to this criticism. First, this observation explains one reason why frequentist analyses, which essentially assume noninformative priors, are successful: In the presence of copious data, the prior belief does not matter. However, where there are few data available, the prior belief does become important, and must be represented in the analysis. Second, a goal of this dissertation is to provide a tool for adjudicating arguments between readers (see Section 2.2.1). If two readers initially were to disagree as to the decision implications of a study, but then were to use a system like THOMAS—which elicits their prior beliefs—to analyze the study, and were to find that they would reach the same conclusion, then the system would be successful.

[1]There is subjectivity in matching the study that was executed to a statistical model for which a test is available, and there is no assurance that the test was not chosen after data "dredging," when the authors might have made sure that the resulting p value would be "significant."

Regarding the novelty of the approach, we can expect fundamental change to occur through persuasion, demonstration of improved benefits (Avorn and Soumerai, 1983; Kanouse and Jacoby, 1988; Greer, 1988; Hill and Weisman, 1991), and, in this case, with improvements in the software that simplify the work.

9.3 Contributions

Because it is interdisciplinary, this work makes contributions in a number of fields: medical informatics, biostatistics, Bayesian biostatistics, artificial intelligence, and medical education.

9.3.1 Medical Informatics

I maintain that the analysis and evaluation of biomedical research is an important component of medical informatics. Few tools have been built for physicians to help them to perform these tasks. The Bayesian methodological formulation developed in Chapter 7 provides an example of the type of semantic layer needed between the physician and the statistical analysis. The result of incorporating this layer into statistical systems is that the physician is able to take control of the interaction with the system, and, therefore, is able to take charge of an important source of information at her disposal. Potentiating such control is one goal of medical-informatics research.

There are implications from this work for the storage of, and access to, the biomedical literature—another concern of medical informatics. In the age of the electronic article, we need to think about the appropriate format for creating and storing the new "literature." The format I propose for the literature is that articles be interactive documents, allowing the reader to investigate the study's methodology in ways I have suggested in creating THOMAS. For instance, the reader should be able to combine her prior knowledge with the data from the study, taking methodological

concerns into account. This capability has implications for how the electronic journal article should be stored, and even for what information electronic-journal editors should request from potential contributors. Specifically, electronic articles should contain the probabilistic and statistical models the investigators think are appropriate and relevant to the study, and should contain the authors' prior beliefs in basic parameters. Statistical models should be represented as influence diagrams, so that a reader's system like THOMAS could tailor the conclusions of the article, using the reader's methodological knowledge. This format further implies that articles might be indexed by the parameters for which the articles supply evidence. Both of these implications provide suggestions for future medical-informatics research.

9.3.2 Biostatistics

This dissertation participates in the discourse within the biostatistical community regarding the propriety of Bayesian methods for interpreting clinical studies. Specifically, I have provided a novel argument for the use of Bayesian statistics—an argument based on a knowledge-level analysis of the use of research results for making clinical decisions (see Sections 3.5 and 4.6). I have shown that the classical approach to data analysis leads to a decision-making process that is heuristic in a way that violates cherished aims of the biomedical scientific community; the p value, I showed in Section 3.5.1, acts as such a heuristic. I have also shown that the Bayesian approach does not violate these ideals.

In addition, this dissertation speaks to a problem of great importance to the biostatistical community: How should research results be reported? Statisticians are often frustrated by the small amount of space journal editors give them and by the degree to which they must simplify their analyses. The format I suggested in Section 9.3.1 allows for statisticians to report analyses that are more complex, at the same time that it allows for the reader to modify the conclusions of the study to take

into account domain knowledge and local experience. My approach thus allows the reader to participate actively in the reporting and interpretation processes.

Furthermore, my use of influence diagrams in describing statistical models suggests that these structures may be useful in the teaching and communication of statistical concepts.

Finally, this dissertation creates no new statistical knowledge, but, rather, attempts to synthesize knowledge from many sources, without reinventing concepts that classical statisticians have already shown to be useful. In fact, a strength of the approach is its use of typical statistical models that obviates the need to resort to ad hoc heuristics. My approach to *credibility* is a case in point, where this concept is represented through the distribution of prior beliefs in methodological parameters throughout the statistical model (see Section 5.8 and page 199).

9.3.3 Bayesian Biostatistics

Biostatisticians have been interested in Bayesian methods for many years, but the approach is only now being put into practice in the course of biostatistical analyses (Spiegelhalter and Freedman, 1988; Breslow, 1989). This dissertation contributes to a number of areas within Bayesian biostatistics.

One Bayesian concern has been how to report the primary likelihood function so that users can combine their prior belief with that likelihood function, and can calculate the corresponding posterior belief. The solution in the past has been to provide a variety of graphs and nomograms (Hildreth, 1963; Dickey, 1973), where the user can find the posterior belief that corresponds to her prior knowledge. THOMAS extends this approach by providing the reader the ability to modify the primary likelihood function itself, and to view the results of that modification.

Another Bayesian concern is how to construct the statistical model appropriate to

a reported study. Eddy and colleagues (Eddy et al., 1991) have provided such an approach, in their Confidence Profile Method. I discussed this approach in Sections 4.7 and 7.8.

The notion of model construction, as we saw in Section 4.4, is a heuristic replacement for the proper Bayesian task of assigning prior beliefs to all models in the universe of models. Thus, in asking the user to supply the statistical structure, we are enlarging the scope of the notion of assessing clients' prior beliefs.

A more global Bayesian concern is how to modify the planning and execution of clinical studies to take decision-theoretical concerns into account. Hilden (1987) and Hilden and Habemma (1990) have considered the implications of a decision-theoretic approach to clinical trials. Spiegelhalter (1986), Spiegelhalter and Freedman (1986) and Spiegelhalter, Freedman and Blackburn (1986) have suggested ways of resolving classical concerns, using Bayesian approaches, in sample-size calculation, in power calculation, and in clinical-trial monitoring. The approach I take provides a technological solution to this problem: If software were available for easy, but meaningful, calculation of Bayesian statistics at the conclusion of a study, then perhaps trials would be designed at the outset to produce the appropriate inputs to those calculations; this design would alter current design practice.

9.3.4 Artificial Intelligence

An important problem in developing AI systems for statistics is the coordination of the various types of knowledge needed by the system: probabilistic, methodological, statistical, frequency, and domain. The framework for THOMAS provides a knowledge representation that isolates each of these categories of knowledge, making system development amenable to a novel type of modular design, as we saw in Chapter 7. The design also enables the system to create names for new entities, on the basis of this knowledge (see Section 7.4.1).

This dissertation extends the use of influence diagrams as a knowledge representation in solving knowledge-based problems, and examines statistical modeling as a domain appropriate for AI (but see Gale and Pregibon (1985)). In the process, I introduce the notion of hierarchical and typed influence diagrams. Such diagrams might be used in systems where many entities fall into strictly defined classes. Furthermore, the use of the metadata-state diagram represents a new strategy for dynamic construction of influence diagrams (see Section 7.2).

Finally, this dissertation takes a more complex view of statistics than may be generally acknowledged in AI research. For instance, some researchers view statistics as referring simply to observed frequencies (Bacchus, 1989). Furthermore, investigators of machine learning often do not incorporate knowledge about the methodological relationship between the ideals they wish to infer and the data they are given.

9.3.5 Medical Education

By examining physicians' use of research data from a systemic point of view, I have developed an approach that suggests the skills and knowledge physicians should have when using the clinical research literature for taking clinical action. Specifically, we should teach physicians probabilistic thinking, utility assessment, and the elements of good research design and methodology. This approach is obviously of importance in teaching critical appraisal of literature, in general.

9.4 Future Work

This dissertation only scratches the surface of making Bayesian techniques accessible to the practicing physician (see the introduction to Chapter 5). THOMAS's capabilities need to be expanded along a number of dimensions. In each extension, there is a conceptual problem and a programming problem involved.

We need a greater variety of physical interfaces for physicians to use (see Section 6.1.1). Conceptually, we need to determine the interfaces to which physicians respond the best. Once we know that, we can improve a system by giving the user more choices at each step, or perhaps by developing a more detailed—and customizable—user model.

We need to extend THOMAS's current utility model (see Sections 4.3.2, 5.3 and 7.7, and Equation 5.1), to provide a richer set of models, especially those taking morbidity into account. Conceptually, we must either develop a library of utility models or decide on the primitive components of such models. With a library of models, a system can enumerate more models than just the pragmatic-difference model when the program seeks the definition of clinical significance. With primitive components, the system can provide tools to assist the user in constructing a wide variety of decision models. Because utility models are subsets of influence diagrams, the capability to help users to construct utility models is similar to that of THOMAS itself—dynamic construction of influence diagrams; we would hope that the insights derived from building THOMAS could be used in this solution. The domain problem with this extension is that, for any study where morbidity is the outcome under consideration, belief about mortality must be included in the decision model. Assessing the morbidity–mortality balance when morbidity is the central issue is not trivial.

Another type of extention of THOMAS's utility model is that of making the model represent the objectives of decision makers other than the patient. For physicians, this extention may involve medicolegal notions. For editors, this extention may include notions of newsworthiness. For students, this extention may concern educational objectives.

We need a wider selection of probabilistic models, beyond the exponential model for mortality (see Section 5.4). Conceptually, we need to develop a knowledge base for *diagnosing* when to use such models, or to insist that the appropriate model be

reported by study investigators. This model-selection subsystem could be incorporated into THOMAS at the step of establishing the study design (step II.A.2.b.i.1 in Table 6.1). Such a reasoner is allowed to be heuristic in this Bayesian system, because it is performing model selection (see Section 4.4); therefore, a traditional rule-based system may be used (see the concluding comments of Chapter 7). Regardless of the diagnosing system used, we are left with the conceptual problem of translating the semantics of a model's parameter to the user. For instance, the shape parameter of a Weibull distribution does not have straightforward verbal semantics; we might best communicate them by displaying the distributions determined by different values, and might have the user choose the curve (or set of curves) closest to her belief.

We need to represent a wider spectrum of methodological concerns. Conceptually, we need models appropriate to issues such as randomization, and the influence of pre-existing conditions (see Section 8.3). The construction of such models is an open area of research. In implementation, we need to construct a visual metaphor consistent with each particular model and with the Bayesian paradigm. We have no assurance that the patient-flow diagram is the appropriate metaphor for baseline characteristics, for instance. In fact, the appreciation for the link between the form of the methodological model and the user interface is a key theme of this dissertation.

The probability-updating algorithm must be broader based. Extending the algorithm beyond beta-distributed variables is conceptually straightforward. The one caveat is that the assumption may be invalid that the form of the prior distribution depends solely on the limits of the variable (e.g., relative risks would be distributed log-normally simply because one limit is finite (zero) and the other is infinite). This assumption properly should be superseded by specific domain knowledge. However, such domain knowledge may lead to nonconjugate forms, which lead to pragmatically intractable calculations, because of the system's need to perform numerical integrations over high-dimensional spaces; Gibbs sampling (Hrycej, 1990) shows some

promise in this respect.

Extending the probabilistic-updating algorithm to allow for dependence among basic parameters is conceptually simple. Again, however, system performance would degrade, because the matrix calculations used in the approximation algorithm would take longer. Specifically, the Cramer matrix (see Equation 7.3) is diagonal when the basic parameters are marginally independent; with dependencies added among basic parameters, the structure becomes arbitrary, and more difficult to use.

Extending the algorithm to allow for dependence among pieces of evidence is precisely what THOMAS was created to *avoid;* the entire likelihood debiasing-approach taken here (see Section 4.2.3) is one of modeling any interaction among pieces of evidence in terms of parameters, leaving the evidence to be conditionally independent, given some effective parameter, of all other pieces of evidence.

Moving beyond the domain of the single research report brings us to the field of Bayesian meta-analysis (Eddy et al., 1991). Creating a system to house meta-analysis requires incorporating notions of relationships *between* studies. A straightforward model is that two studies provide evidence for the same population parameter (Figure 8.3). More difficult notions to represent are the dependence of the design of one study on the results of a previous study, or the dependence of outcomes of two studies that results from their having examined the same set of patients.

Thus, we can make certain improvements to THOMAS simply through more programming. Other improvements require research. Still others require changes in the biomedical research literature itself.

9.5 Concluding Remarks

Changes in quantity often lead to changes in quality. Recent innovations in computer software and greater computer availability have given physicians increased access to

the biomedical research literature. This quantitatively increased access will lead, inevitably, to qualitative changes in the way the biomedical community reasons with, and about, research data. The work presented in this dissertation may play a part in preparing for those fundamental changes.

Appendix A

Glossary

A.1 Abbreviations

AI	artificial intelligence (page 11)
ANOVA	analysis of variance (page 220)
CCT	controlled clinical trial (page 21)
CPM	Confidence Profile Method (page 120)
iid	independent and identically distributed (page 56)
MI	myocardial infarction (page 3)
pdf	probability distribution (page 53)
RCT	randomized clinical trials (page 21)

A.2 Notation

$\mathcal{A}$	a set of decision alternatives (page 103)

α	first parameter of a beta distribution (page 54); a methodological parameter for the proportion of patients who violated study protocol (page 134)
β	second parameter of a beta distribution (page 54)
$\mathcal{B}(p)$	Bernoulli distribution with success probability parameter, p (page 54)
$\mathcal{BE}(\alpha, \beta)$	beta distribution with parameters α and β (page 54)
$\mathcal{BI}(\pi, n)$	binomial distribution with success probability, π, and sample size, n (page 54)
δ	decision alternative (page 103)
$\mathcal{E}(\lambda)$	exponential model with instantaneous rate parameter (page 54)
L	lifespan (page 126)
$\ell(\cdot\|\cdot)$	likelihood of the first argument, given the second argument (page 53)
λ	instantaneous failure (mortality) rate for the exponential model (page 54)
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean, μ, and variance, σ^2 (page 54)
τ	methodological parameter for the proportion of time patients were compliant with therapy (page 138)
θ	a generic parameter, or a timed-mortality-rate parameter, e.g., the mortality rate (within 3 months) of patients assigned to metoprolol
$\theta^{\text{pop}}_{\text{assigned}}$	the population outcome parameter in patients assigned to therapy—superscript denotes level, subscript denotes history
$u(\cdot)$	utility of the argument (page 103)
x or X_i	single observation (page 53)
X	random variable (page 53)
$\langle X \rangle$	the mean of the pdf for X (page 87)
$\mathbf{X}$	column vector of observations (page 53)
$\sim$	is distributed as

A.3 Graphical Conventions

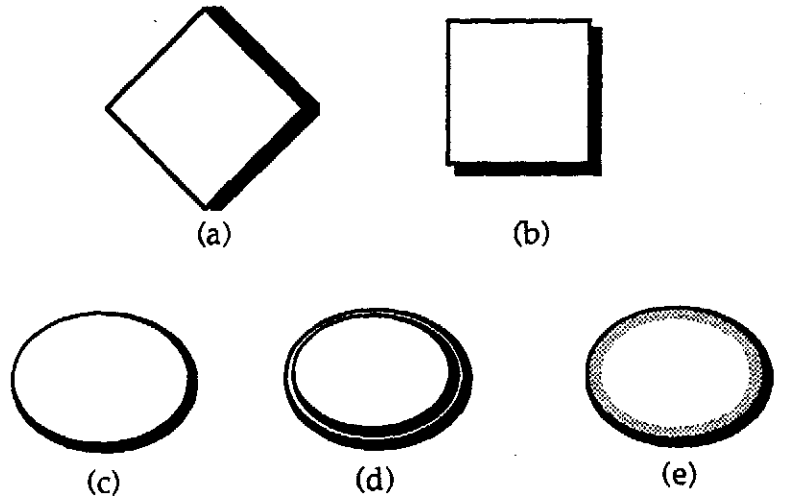


Figure A.1: Influence-diagram nodes. (a) Utility node, representing the overall utility to the decision maker of a state of the world; the node's value is a function of its predecessors. (b) Decision node, which contains the decision alternatives (not shown). (c) Chance node, which represents a random variable, whose belief is a function of its predecessors. (d) Deterministic node, whose value is a function of its predecessors. (e) Evidence node, a chance node whose value can be, or has been, observed.

If two nodes are connected by a directed (arrowed) arc, then the source node is the *predecessor* and the destination node is the *successor*.

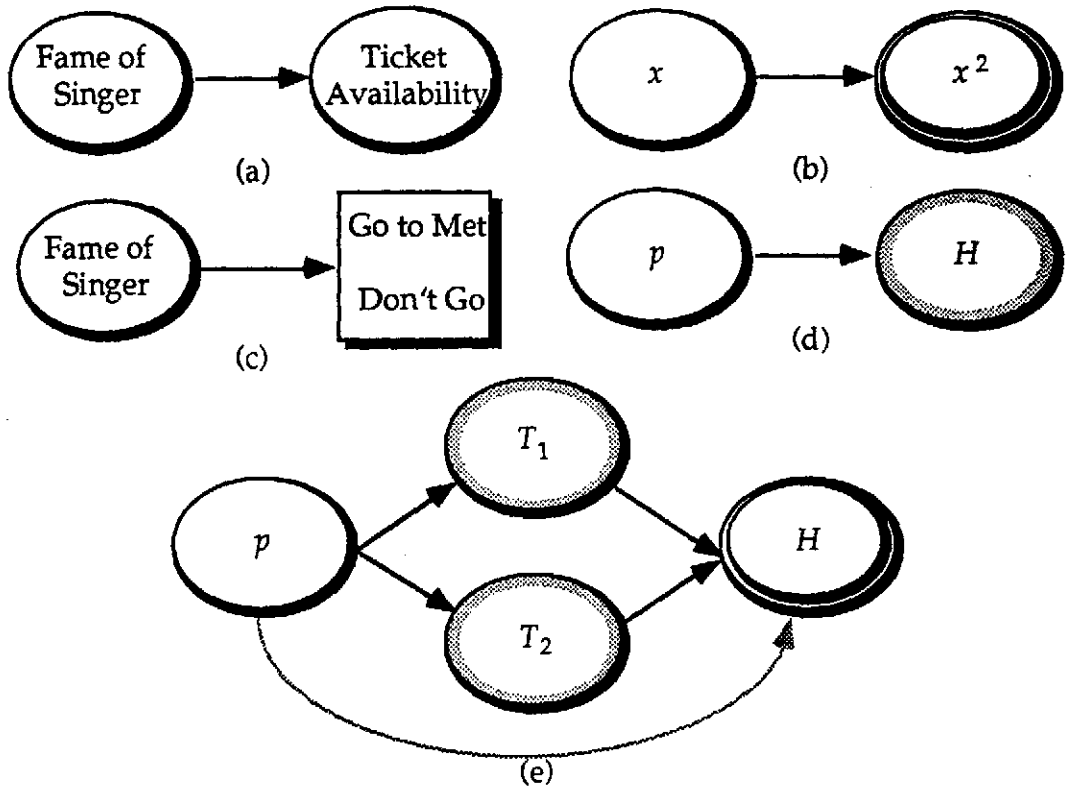


Figure A.2: Influence-diagram arcs. All arcs in an influence diagram are directed and denote the dependence of the value of the destination node on the value of the source node. The exact nature of the arc depends on the types of nodes involved. (a) Probabilistic arc. This figure denotes the model that my belief in how many tickets remain available for a particular opera depends on the fame of the singer, about which I am also uncertain. (b) Deterministic arc. This figure shows that the value of x^2 can be found with certainty, if I know the value of x. (c) Informational arc. This figure indicates that, at the time that I am deciding whether to go to the Metropolitan Opera House on the evening of a performance, I know how famous the singer is. (d) Likelihood arc. This figure shows that the number of heads I observe in tossing a coin depends on the probability of that coin falling heads. (e) Derived-likelihood arc. This figure expands Figure d: The coin is tossed twice, and the observed outcomes are added deterministically ($H = T_1 + T_2$). By virtue of certain properties of the Bernoulli distribution (the basis of the likelihood arcs impinging on each toss node), the probabilistic dependence on p of the calculated value—the number of heads—is known to be binomial distributed. The arc is superfluous in defining the relationships among the nodes in this diagram.

Appendix B

Influence Diagrams

Influence diagrams represent decision problems under uncertainty. There are two components to an influence diagram: the uncertainty component, and the decision component. The uncertainty component is represented by a *belief network*—an acyclic directed graph[1] whose nodes represent variables and whose arcs represent probabilistic dependencies. Given any set of consistent probability statements about a set of domain variables representing a full joint-probability space, a belief network can be constructed corresponding exactly to the joint-probability distribution, and the joint-probability can be calculated directly from the diagram. Because the graph is acyclic, the nodes can be placed in an order where children precede parents. Given this order, the joint-probability distribution for n variables is factored as follows:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P\left(A_i \mid \bigcap_{j=i+1}^{n} A_j\right). \tag{B.1}$$

For instance, with only three propositions,

$$P(A, B, C) = P(A \mid B, C) \cdot P(B|C) \cdot P(C). \tag{B.2}$$

[1]That is, all arcs are directed and it is impossible to return to any node by following a path of arcs beginning from that node.

If we know that $P(A \mid B, C) = P(A \mid B)$, then we are saying that our knowledge about A depends *only* on B, and is *independent* of C; we draw the diagram shown in Figure B.1b. The knowledge engineer constructs the belief network by representing (conditional) independence as the absence of an arc between nodes deemed irrelevant to each other. Given such a network, any dependency or independency statement derived from the graph is present in the original probability statements, or is implied by them (Geiger and Pearl, 1988). Belief networks are special cases of graph-theoretic objects called semigraphoids, which satisfy the properties of symmetry, decomposition, weak union, and contraction (Pearl, 1988).

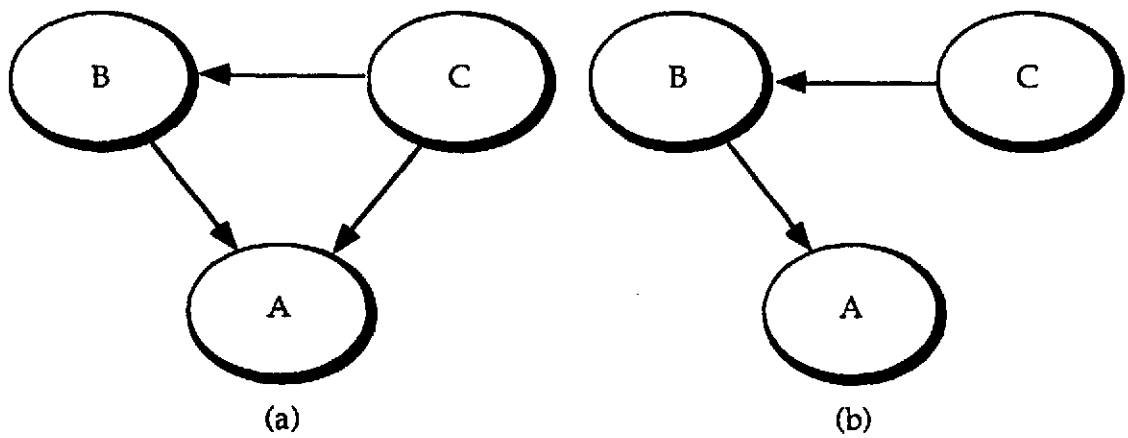


Figure B.1: Influence-diagram example. This diagram depicts the joint probability of $A, B,$ and C. (a) The joint distribution is represented as a fully connected graph. (b) The knowledge about conditional independence that $P(A \mid B, C) = P(A \mid B)$ allows the deletion of the arc between B and C : $P(A, B, C) = P(A \mid B) \cdot P(B \mid C) \cdot P(C)$.

A belief network is fully specified when (1) the possible values of each variable are defined, (2) the dependency of each variable on its parent variables is specified (probability distributions and functions), and (3) the prior beliefs are encoded. Thus, the belief network in Figure B.1b may be specified by stating that (1) each variable is dichotomous, (2) $P(B \mid C) = 0.9, P(B \mid \text{not } C) = 0.7, P(A \mid B) = 0.5,$ and $P(A \mid$

not B) = 0.6, and (3) $P(C) = 0.5$. A fully specified belief network induces beliefs in all its constituents. If variables have been observed, then the network implicitly contains the updated beliefs in all variables related to the observed variables. The task of making explicit those implicit beliefs is the job of probabilistic-updating algorithms. The creation of these algorithms has been a focus of research in the past ten years (Pearl, 1986; Lauritzen and Spiegelhalter, 1988)

The decision component of an influence diagram is represented by decision and utility nodes. An influence diagram is fully specified when (1) its uncertainty component is fully specified, and (2) its utility model is specified. Algorithms for arriving at the optimal decision using an influence diagram are given by Shachter (1988b) and Cooper (1988).

Lehmann (1990) provides a broad introduction to representations of uncertainty in AI, and Heckerman (1990) gives a more specific introduction to influence diagrams. A recent issue of the journal *Networks*[2] and the conference proceedings edited by Oliver and Smith (1990) provide an introduction to specific research areas, along with useful references.

[2]Vol 20, number 5 (1990).

Bibliography

Agosta, J. M. (1988). The structure of Bayes nets for vision recognition. In *Proceedings of the Fourth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 1–7, University of Minnesota, August 19–20.

Andersen, S. K., Olesen, K. G. et al. (1989). HUGIN: A shell for building Bayesian belief universes for expert systems. In *Proceedings of the Eleventh International Joint Conference in Artificial Intelligence,* pages 1080–1085, Detroit, August 20–25.

Anjewierden, A. (1987). The KADS system. In *Proceedings of the First European Workshop on Knowledge Acquisition*, pages E2:1–12. Redding, UK.

Antman, E. M. and Braunwald, E. (1990). Acute MI management in the 1990s. *Hospital Practice*, 25:65–82.

Antman, K., Amato, D. et al. (1985). Selection bias in clinical trials. *Journal of Clinical Oncology*, 3(8):1142–1147.

Armitage, P. (1983). *Statistical methods in medical research.* Blackwell Scientific, Oxford.

Avorn, J. and Soumerai, S. B. (1983). Improving drug-therapy decisions through educational outreach: A randomized controlled trial of academically based 'detailing'. *New England Journal of Medicine*, 308:1457–1463.

Bacchus, F. (1989). Lp: A logic for statistical information. In *Proceedings of the Fifth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 1–6, Windsor, Ontario, August 18–20.

Beck, J. R., Kassirer, J. P., and Pauker, S. G. (1982). A convenient approximation of life expectancy (The 'Deale'). *The American Journal of Medicine*, 73:883–897.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The new S language: A programming environment for data analysis and graphics.* Computer Science Series. Wadsworth & Brooks/Cole, Belmont, CA.

Beinlich, I. and Herskovits, E. (1990). Ergo: A graphical environment for constructing Bayesian belief networks. In *Proceedings of the Sixth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 114–121, Cambridge, MA, July 27–29.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer Series in Statistics. Springer-Verlag, New York, second edition.

Berger, J. O. (1988). In defense of the Likelihood Principle: Axiomatics and coherency. In Bernardo, J. M., DeGroot, M. H. et al., editors, *Bayesian Statistics 3*, pages 33–65. Oxford University Press, Oxford.

Berger, J. O. and Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76:159–165.

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*, volume 6 of *Lecture Notes–Monograph Series.* Institute of Mathematical Statistics, Hayward, CA.

Berlin, J. A., Laird, N. M. et al. (1989). A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine*, 8:141–151.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 41:113–147.

Berndt, E. K., Hall, B. H. et al. (1974) Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 3/4:653–665.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the Amercian Statistical Association*, 57:269–326.

Bonduelle, Y. (1987). *Aggregating expert opinions by resolving sources of disagreement.* PhD thesis, Stanford University, Stanford, California.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, MA.

Breese, J. (1987). *Knowledge representation and inference in intelligent decision systems.* PhD thesis, Stanford University, Stanford, CA.

Breese, J. S. and Fehling, M. R. (1988). Decision-theoretic control of problem solving: Principles and architecture. In *Proceedings of the Fourth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 30–37, University of Minnesota, August 19–20.

Breslow, N. (1989). Biostatistics and Bayes. In *Proceedings of the American Statistical Association*, pages 51–71, Washington, D. C.

Breuker, J., Wielinga, B. J. et al. (1987). Model driven knowledge acquisition: Interpretation models. Deliverable A1, Esprit Project 1098 Memo 87, VF Project Knowledge Acquisition in Formal Domains, Social Science Informatics, University of Amsterdam, Amsterdam.

Brown, Jr. , B. W. and Hollander, M. (1977). *Statistics: A biomedical introduction.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Brown, Jr., B. W. (1984). The randomized clinical trial. *Statistics in Medicine*, 3:307–311.

Buchanan, B. G. and Shortliffe, E. H. (1984). *Rule-based expert systems: The* MYCIN *experiments of the Stanford Heuristic Programming Project.* The Addison-Wesley Series in Artificial Intelligence. Addison-Wesley, Reading, MA.

Buntine, W. (1989). Decision tree induction systems: A Bayesian analysis. In Kanal, L. N. and Lemmer, J. F., editors, *Uncertainty in artificial intelligence 3*, pages 109–127. North-Holland, Amsterdam.

The Canadian Cooperative Study Group. (1978). A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *New England Journal of Medicine*, 299:53–59.

Chalmers, T. C., Celano, P. et al. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine*, 309:1358–1361.

Chalmers, T. C., Smith, Jr. et al. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, 2:31–49.

Chavez, R. M. (1991). *Architectures and approximation algorithms for probabilistic expert systems.* PhD thesis, Stanford University, Stanford, CA.

Chen, P. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions in Database Systems*, 1(1):9-36.

Chytil, M. K. (1986). Metadata and its role in medical informatics. In Salomon, R., Blum, B., and Jörgensen, M. J., editors, *MEDINFO 86*, pages 149–154. IFIP-IMIA, North-Holland.

Clancey, W. J. (1989). The knowledge level reinterpreted: Modeling how systems interact. *Machine Learning*, 4:285–291.

Clayton, M. K., Geisser, S., and Jennings, D. E. (1986). A comparison of several model selection procedures. In Goel, P. K. and Zellner, A., editors, *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti*, chapter 27, pages 425–439. North-Holland, Amsterdam.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Houghton Mifflin, Boston.

Cooper, G. F. (1988). A method for using belief networks as influence diagrams. In *Proceedings of the Fourth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 55–63, University of Minnesota, August 19–20.

Cornfield, J. (1966a). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *Journal of the American Statistical Association*, 61(315):577–594.

Cornfield, J. (1966b). Recent methodological contributions to clinical trials. *American Journal of Epidemiology*, 104(4):408–421.

Cox, D. R. (1977). The teaching of the strategy of statistics. *Bulletin of the International Statistical Institute*, 47(Book I):553–558.

de Finetti, B. (1974). *Theory of probability: A critical introductory treatment*, volume 1 of *The Wiley Series in Probability and Mathematical Statistics.* Wiley, London.

DeGroot, M. H. (1970). *Optimal statistical decisions.* McGraw-Hill Book Company, New York.

DerSimonian, R. and Laird, N. (1986). Meta-Analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *Annals of Probability*, 8:745–764.

Diamond, G. A. and Forrester, J. S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Annals of Internal Medicine*, 98:385–394.

Dickey, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *Journal of the Royal Statistical Society, Series B*, 35:285–305.

Dixon, W. J. (1985). BMDP *statistical software manual*. University of California Press, Berkeley, CA.

Draper, N. R. and Guttman, I. (1986). A common model selection criterion. In Viertl, R., editor, *Probability and Bayesian statistics*, pages 139–150. Plenum, New York.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahnemann, D., Slovic, P., and Tversky, A., editors, *Judgment under uncertainty: Heuristics and biases*, chapter 18, pages 249–267. Cambridge University Press, Cambridge, UK.

Eddy, D. M. (1989). The Confidence Profile Method: A Bayesian method for assessing health technologies. *Operations Research*, 37(2):210–228.

Eddy, D. M. (1990). The challenge. *Journal of the American Medical Association*, 263(2):287–290.

Eddy, D. M., Hasselblad, V., and Shachter, R. (1990). An introduction to a Bayesian method for meta-analysis: The Confidence Profile Method. *Medical Decision Making*, 10(1):15–23.

Eddy, D. M., Hasselblad, V., and Shachter, R. (1991). *The statistical synthesis of evidence: Meta-analysis by the Confidence Profile Method.* Academic Press, Boston.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge University Press, Cambridge, UK.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrics*, 58(3):403–417.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40:1–5.

Ellenberg, S. S. (1984). Randomization designs in comparative clinical trials. *New England Journal of Medicine*, 310(21):1404–1408.

Ellman, T. (1986). Representing statistical computations: Toward a deeper understanding. In Gale, W. A., editor, *Artificial intelligence and statistics*, pages 228–238. Addison-Wesley, Reading, MA.

Emerson, J. D. and Colditz, G. A. (1983). Use of statistical analysis in *The New England Journal of Medicine. New England Journal of Medicine*, 309:709–713.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31:195–233.

Feinstein, A. R. (1985). *Clinical epidemiology: The architecture of clinical research.* Saunders, Philadelphia.

Felson, D. T., Anderson, J. J., and Meenan, R. F. (1990). Time for changes in the design, analysis, and reporting of rheumatoid arthritis clinical trials. *Arthritis and Rheumatism*, 33:140–149.

Field, M. J. and Lohr, K. N., editors (1990). *Clinical practice guidelines: Directions for a new program*, Committee to Advise the Public Health Service on Clinical

Practice Guidelines, Institute of Medicine, National Academy Press, Washington, DC.

Fisher, R. A. (1959). *Statistical methods and scientific inference.* Oliver and Boyd, Edinburgh, 2d edition.

Freedman, L. S. (1990). The effect of partial noncompliance on the power of a clinical trial. *Controlled Clinical Trials*, 11:157–168.

Friedman, S. B. and Phillips, S. (1981). What's the difference? Pediatric residents and their inaccurate concepts regarding statistics. *Pediatrics*, 68:644–646.

Gale, W. A. (1986a). REX review In Gale, W. A., editor, *Artificial Intelligence and Statistics*, pages 173–227. Addison-Wesley, Reading, MA.

Gale, W. A. (1986b). *Artificial intelligence and statistics.* Addison-Wesley, Reading, MA.

Gale, W. A. and Pregibon, D. (1985). Artificial intelligence research in statistics. *AI Magazine*, 5:72–75.

Gardner, M. J. and Bond, J. (1990). An exploratory study of statistical assessment of papers published in the *British Medical Journal. Journal of the American Medical Association*, 263(10):1355–137.

Gehlbach, S. H. (1982). *Interpreting the medical literature: A clinician's guide.* Collamore, Lexington, MA.

Geiger, D. and Pearl, J. (1988). On the logic of causal models. In *Proceedings of the Fourth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 136–147, University of Minnesota, August 19-20.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelband, H. (1983). *The impact of randomized clinical trials on health policy and medical practice*, volume OTA-BP-H-22. Congress of the United States, Office of Technology Assessment, Washington, D.C.

Goel, P. K. (1988) Software for Bayesian analysis: Current status and additional needs. In Bernardo, J. M., DeGroot M. H. et al., editors, *Bayesian Statistics 3*, pages 173–188. Oxford University Press, Oxford.

Goffman, W. (1981). The ecology of the biomedical literature and information retrieval. In Warren, K. S., editor, *Coping with the biomedical literature*, chapter 3, pages 31–46. Praeger, New York.

Goldman, R. P. (1990). *A probabilistic approach to language understanding.* PhD thesis, Brown University, Providence, RI.

Good, I. J. (1980). The philosophy of exploratory datum analysis. In *Proceedings of the Business and Economics Statistics Section*, pages 1–7. American Statistical Association.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* University of Minnesota Press, Minneapolis.

Greenland, S. (1984). Bias in methods for deriving standardized morbidity ratio and attributable fraction estimates. *Statistics in Medicine*, 3:131–141.

Greenland, S. (1987). Bias in indirectly adjusted comparisons due to taking the total study population as the reference group. *Statistics in Medicine*, 6:193–195.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419.

Greer, A. L. (1988). The state of the art versus the state of the science: The diffusion of new medical technologies into practice. *International Journal of Technology Assessment in Health Care*, 4:5–26.

Hamaker, H. C. (1977). Subjective probabilities and exchangeability from an objective point of view. *International Statistical Review*, 45:223–231.

Hand, D. J. (1986). Patterns in statistical strategy. In Gale, W. A., editor, *Artificial intelligence and statistics*, pages 355–387. Addison-Wesley, Reading, MA.

Hanson, N. R. (1961). *Patterns of discovery: An inquiry into the conceptual foundations of science.* Cambridge University Press, Cambridge, UK.

Hayes-Roth, B., Waterman, D. A., and Lenat, D. B. (1983). *Building expert systems.* Addison-Wesley, Reading, MA.

Haynes, R. B., McKibbon, K. A. et al. (1986). How to keep up with the medical literature: I. Why try to keep up and how to get started. *Annals of Internal Medicine*, 105:149–153.

Heckerman, D. and Jimison, H. (1987). A perspective on confidence and its use in focusing attention during knowledge acquisition. In *Proceedings of the Third Annual Workshop on Uncertainty in Artificial Intelligence*, pages 123–131, University of Washington, July 10–12.

Heckerman, D. E. (1990). *Probabilistic similarity networks.* PhD thesis, Stanford University, Stanford, California.

Heckerman, D. E. and Horvitz, E. J. (1988). The myth of modularity in rule-based systems for reasoning with uncertainty. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty in artificial intelligence 2*, pages 23–34. North-Holland, Amsterdam.

Heckerman, D. E., Horvitz, E. J., and Nathwani, B. N. (1989). Update on the Pathfinder project. In Kingsland, III, L. C., editor, *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 203–207, Washington, DC. IEEE Computer Society Press.

Heckerman, D. E., Horvitz, E. J., and Nathwani, B. N. (1990). Toward normative expert systems: The Pathfinder project. Technical Report KSL-90-08, Stanford University, Stanford, California.

Herskovits, E. (1991). *Computer-based probabilistic network construction.* PhD thesis, Stanford University, Stanford, CA.

Hilden, J. (1987). Reporting clinical trials from the viewpoint of a patient's choice of treatment. *Statistics in Medicine*, 6:745–752.

Hilden, J. and Habbema, J. D. F. (1990). The marriage of clinical trials and clinical decision science. *Statistics in Medicine*, 9:1243–1257.

Hildreth, C. (1963). Bayesian statisticians and remote clients. *Econometrica*, 31(3):422–438.

Hill, M. N. and Weisman, C. S. (1991). Physicians' perceptions of consensus reports. *International Journal of Assessment in Health Care*, 7:30–41.

Hjalmarson, Å., Herlitz, J. et al. (1981). Effect on mortality of metoprolol in acute myocardial infarction. *Lancet*, 2(9251):823–827.

Holtzman, S. (1989). *Intelligent decision systems.* Addison-Wesley, Menlo Park, CA.

Horvitz, E. J., Cooper, G. F., and Heckerman, D. E. (1989). Reflection and action under scarce resources: Theoretical principles and empirical study. In *Proceedings of the Eleventh International Joint Conference in Artificial Intelligence*, pages 1121–1127, Detroit, MI, USA, August 20–25.

Howard, R. A. (1983). Decision analysis in systems engineering. In Howard, R. A. and Matheson, J. E., editors, *Readings on the Principles and Applications of Decision Analysis*, pages 57–94. Strategic Decision Group, Menlo Park, CA.

Howard, R. A. (1988). From influence to relevance to knowledge. An influence-diagram approach to medical-technology assessment. In Oliver, R. M. and Smith, J. Q., editors, *Influence diagrams, belief nets, and decision analysis*, pages 3–23. Wiley, Chichester.

Howard, R. A. and Matheson, J. (1981). *Readings on the Principles and Applications of Decision Analysis.* Strategic Decision Group, Menlo Park, CA.

Howson, C. and Urbach, P. (1989). *Scientific reasoning: The Bayesian approach.* Open Court, La Salle, IL.

Hrycej, T. (1990). Gibbs sampling in Bayesian networks. *Artificial Intelligence*, 46:351–363.

Jasinski, P. J. (1986). Databases, statistics and expert systems: An integrated approach to clinical data management and analysis. In Salomon, R., Blum, B., and Jörgensen, M. J., editors, *MEDINFO 86*, pages 1127–1129. IFIP-IMIA, North-Holland.

Kanouse, D. E. and Jacoby, I. (1988). When does information change practitioners' behavior? *International Journal of Technology Assessment in Health Care*, 4:27–33.

Klein, D. A., Lehmann, H. P., and Shortliffe, E. H. (1990). A value-theoretic expert system for evaluating randomized clinical trials. In Miller, R., editor, *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, pages 330–334, Washington, DC. IEEE Computer Society Press.

Kleinbaum, D. G., Morgenstern, H., and Kupper, L. L. (1981). Selection bias in epidemiologic studies. *American Journal of Epidemiology*, 113(4):452–463.

Kolmogorov, A. N. (1933/1965). *Foundations of the theory of probability.* Chelsea, New York.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* University of Chicago Press, Chicago.

Kyburg, Jr., H. E. (1983). The reference class. *Philosophy of Science*, 50:374–397.

Kyburg, Jr., H. E. (1987). Higher order probabilities. In *Proceedings of the Third Annual Workshop on Uncertainty in Artificial Intelligence*, pages 30–38, University of Washington, July 10–12.

L'Abbé, K. A., Detsky, A. S., and O'Rourke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine*, 107:224–233.

Lakatos, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials*, 7:189–199.

Laudan, L. (1984). *Science and values: The aims of science and their role in scientific debate.* University of California Press, Berkeley.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 5(2):157–224.

Lehmann, E. L. (1986). *Testing statistical hypotheses.* A Wiley publication in mathematical statistics. Wiley, New York, second edition.

Lehmann, H. P. (1988). Knowledge acquisition for probabilistic expert systems. In Greenes, R. A., editor, *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care*, pages 73–77, Washington, DC. IEEE Computer Society Press.

Lehmann, H. P. (1990). Uncertainty in artificial intelligence: A review in light of recent workshops. *Journal of Mathematical Psychology*, 34:336–363.

Levitt, T. S. (1988). Model-based probabilistic situation inference in hierarchical hypothesis spaces. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty in artificial intelligence 2*, pages 347–356. North-Holland, Amsterdam.

Lindley, D. V. (1972). *Bayesian statistics.* Society for Industrial and Applied Mathematics, Philadelphia.

Linhart, H. and Zucchini, W. (1986). *Model selection.* Wiley, New York.

Mallows, C. L. and Walley, P. (1980). A theory of data analysis? In *Proceedings of the Business and Economics Section*, pages 8–14, Washington, D. C. American Statistical Association.

Meinert, C. L. and Tonascia, S. (1986). *Clinical trials: Design, conduct, and analysis*, volume 8 of *Monographs in Epidemiology and Biostatistics.* Oxford University Press, New York.

Meinert, C. L., Tonascia, S., and Higgins, K. (1984). Content of reports on clinical trials: A critical review. *Controlled Clinical Trials*, 4:328–347.

Miettinen, O. S. and Cook, E. F. (1981). Confounding: Essence and detection. *American Journal of Epidemiology*, 114(4):593–603.

Miller, R. G. (1981). *Survival Analysis.* Wiley, New York.

Mosteller, F. (1981). Evaluation: Requirements for scientific proof. In Warren, K. S., editor, *Coping with the biomedical literature*, chapter 8, pages 103–122. Praeger, New York.

Musen, M. A. (1989). *Automated generation of model-based knowledge-acquisition tools.* Pittman, London.

Nelde, J. A and Wolstenhome, D. (1986). *A front-end for* GLIM. In *COMPSTAT-86*, pages 113–117, Fort Collins, CO. American Statistical Association.

Newell, A. (1981). The knowledge level. *AI Magazine*, 2:1–20.

Neyman, J. and Pearson, E. S. (1930). *On the problem of two samples.* Imprimerie de l'université, Cracovie.

Nugent, W. H. (1986). Artificial intelligence techniques for retrospective help in data analysis. In *COMPSTAT-86*, pages 118–121, Fort Collins, CO. American Statistical Association.

Oldford, R. W. and Peters, S. C. (1986). Implementation and study of statistical strategy. In Gale, W. A., editor, *Artificial intelligence and statistics*, pages 335–353. Addison-Wesley, Reading, MA.

Oldford, R. W. and Peters, S. C. (1988). DINDE: Towards more sophisticated software environments for statistics. *SIAM Journal of Scientific and Statistical Computing*, 9(1):191–211.

Oliver, R. M. and Smith, J. Q. (1990). *Influence diagrams, belief nets, and decision analysis.* Wiley, Chichester.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288.

Pearl, J. (1987). Do we need higher-order probabilities, and if so, what do they mean? In *Proceedings of the Third Annual Workshop on Uncertainty in Artificial Intelligence*, pages 46–60, University of Washington, July 10–12.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* The Morgan Kauffmann Series in Representation and Reasoning, R. J. Brachman, editor, Morgan Kauffmann, San Mateo, CA.

Peto, R., Pike, M. C. et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer*, 34:585–612.

Pocock, S. J., Hughes, M. D., and Lee, R. J. (1987). Statistical problems in the reporting of clinical trials: A survey of three medical journals. *New England Journal of Medicine*, 317:426–432.

Pregibon, D. (1986). A DIY guide to statistical strategy. In Gale, W. A., editor, *Artificial intelligence and statistics*, pages 389–399. Addison-Wesley, Reading, MA.

Radnitzky, G. (1973). *Contemporary schools of metascience.* Henry Regnery, Chicago.

Reisch, J. S., Tyson, J. E., and Mize, S. G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics*, 84(5):815–827.

Rennels, G. D. (1987). *A computational model of reasoning from the clinical literature*, Lecture Notes in Medical Informatics, volume 32, P. L. Reichertz and D. A. B. Lindberg, editors, Springer-Verlag, Berlin.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, 32:51–63.

Sackett, D. L. (1981). Evaluation: Requirements for clinical application. In Warren, K. S., editor, *Coping with the biomedical literature*, chapter 9, pages 123–157. Praeger, New York.

Sackett, D. L. and Gent, M. (1979). Controversy in counting and attributing events in clinical trials. *New England Journal of Medicine*, 301:1410–1412.

Sackett, D. L., Haynes, R. B et al. (1991). *Clinical epidemiology: A basic science for clinical medicine.* Little, Brown, Boston, second edition.

Sacks, H. S., Berrier, J. et al. (1987). Meta-analyses of randomized controlled trials. *New England Journal of Medicine*, 316:450–455.

Savage, L. J. (1972). *The foundations of statistics.* Dover, New York, second edition.

Scura, G. and Davidoff, F. (1981). Case-related use of the medical literature: Clinical librarian services for improving patient care. *Journal of the American Medical Association*, 245(1):50–52.

Self, M. and Cheeseman, P. (1987). Bayesian prediction for artificial intelligence. In *Proceedings of the Third Annual Workshop on Uncertainty in Artificial Intelligence*, pages 61–69, University of Washington, Seattle, July 10–12.

Shachter, R. D. (1988a). A linear approximation method for probabilistic inference. In *Proceedings of the Fourth Annual Workshop on Uncertainty in Artificial Intelligence*, pages 299–0306, University of Minnesota, August 19–20.

Shachter, R. D. (1988b). Probabilistic inference and influence diagrams. *Operations Research*, 36(4):589–604.

Shachter, R. D. (1988c). DAVID: Influence diagram processing system for the Macintosh. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty in artificial intelligence 2*,pages 191–196. North-Holland, Amsterdam

Shachter, R. D., Eddy, D. M. and Hasselblad, V. (1990). An influence-diagram approach to medical-technology assessment. In Oliver, R. M. and Smith, J. Q., editors, *Influence diagrams, belief nets, and decision analysis*, pages 321–350. Wiley, Chichester.

Shwe, M., Middleton, B. et al. (to appear). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: I. The probabilistic model and inference algorithm. *Method of Information in Medicine.*

Smith, J. et al. (1985). RED: A red-cell antibody identification expert module. *Journal of Medical Systems*, 9:121.

Snedecor, G. M. and Cochran, W. G. (1980). *Statistical Methods.* Science Press, seventh edition.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5:421–433.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7:8–17.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5:1–13.

Spiegelhalter, D. J. and Freedman, L. S. (1988). Bayesian approaches to clinical trials. In Bernardo, J. M., DeGroot, M. H. et al., editors, *Bayesian Statistics 3*, pages 453–477, Oxford. Third Valencia International Meeting, Clarendon Press.

SPSS, Inc. (1983). SPSS^X *user's guide.* SPSS, Inc., Chicago.

Star, S. (1987). Theory-based inductive learning: An integration of symbolic and quantitative methods. In *Proceedings of the Third Annual Workshop on Uncertainty in Artificial Intelligence*, pages 229–236, University of Washington, July 10–12.

Student (Gosset, W. S.). (1937). Comparison between balanced and random arrangements of field plots. *Biometrika*, XXIX:363–379.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *Journal of Philosophy*,75:76–92.

Tierney, L. (1990). *Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics.* Wiley, New York.

Trudeau, P. A. (1991). *Bayesian inference and its application to meta-analysis of epidemiologic data.* PhD thesis, University of Texas H. S. C. at Houston School of Public Health, Houston, TX.

The University Group Diabetes Program (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. *Diabetes*, 19(Suppl 2):747–830.

The Veterans Affairs Cooperative Variceal Sclerotherapy Group, (1991) Prophylactic sclerotherapy for esophageal varices in men with alcoholic liver disease. A randomized, single-blind, multicenter trial. *New England Journal of Medicine*, 324:1779–1784.

von Mises, R. and Geiringer, H. (1928/1964). *The mathematical theory of probability and statistics.* Academic Press, New York.

von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior.* Princeton University Press, Princeton, NJ, second edition.

Wald, A. (1950). *Statistical decision functions.* Wiley Publications in Statistics. Wiley, New York.

Walker, C. J., McKibbon et al. (1989). Methods for assessing the competence of physicians' use of MEDLINE with GRATEFUL MED. In Kingsland, III, L. C., editor, *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 441–444, Washington, DC. IEEE Computer Society Press.

Warren, K. S. (1981). *Coping with the biomedical literature: A primer for the scientist and the clinician.* Praeger Publishers, New York.

Weaver, R. R. (1991). Assessment and diffusion of computerized decision support systems. *International Journal of Technology Assessment in Health Care*, 7:42–50.

Weiner, J. M., Horowitz, R. S. and Bauer, M. Clinical trial expert system", In *COMPSTAT-87*, pages 117–122. American Statistical Association.

Weiss, S. T. and Samet, J. M. (1980). An assessment of physician knowledge of epidemiology and biostatistics. *Journal of Medical Education*, 55:692–697.

Wellman, M. P. (1988). *Formulation of tradeoffs in planning under uncertainty.* PhD thesis, Massachussets Institute of Technology, Cambridge, MA.

Williamson, J. W., German, P. S. et al. (1989). Health science information management and continuing education of physicians: A survey of U.S. primary care practitioners and their opinion leaders. *Annals of Internal Medicine*, 110:151–160.

Wittkowski, K. M. (1986). An expert system for testing statistical hypotheses. In *COMPSTAT-86*, pages 438–442, Fort Collins, CO. American Statistical Association.

Wulff, H. R., Andersen, B. et al. (1987). What do doctors know about statistics? *Statistics in Medicine*, 6:3–10.

Wyatt, J., Cuff, R., et al. (1991) Design-a-Trial: a knowledge-based system to help design clinical trials (abstract). Presented at the Third Conference of the European Society for Artificial Intelligence in Medicine, Maastricht, Holland, June.

Yusuf, S., Peto, R. et al. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, 27(3):335–371.