11
Evaluation and Technology Assessment

CHARLES P. FRIEDMAN, JEREMY C. WYATT, AND DOUGLAS K. OWENS

After reading this chapter, you should know the answers to these questions:

- Why are empirical studies based on the methods of evaluation and technology assessment important to the successful implementation of information resources to improve health care?
- What challenges make studies in informatics difficult to carry out? How are these challenges addressed in practice?
- Why can all evaluations be classified as empirical studies?
- What are the major assumptions underlying objectivist and subjectivist approaches to evaluation? What are the advantages and disadvantages of each?
- What are the factors that distinguish the three stages of technology assessment?
- How does one distinguish measurement and demonstration aspects of objectivist studies, and why are both aspects necessary?
- What steps are typically undertaken in a measurement study? What designs are typically used in demonstration studies?
- What is the difference between cost-effectiveness and cost-benefit analyses? How can investigators address issues of cost effectiveness and cost benefit of medical information resources?
- What steps are followed in a subjectivist study? What techniques are employed by subjectivist investigators to ensure rigor and credibility of their findings?
- Why is communication between investigators and clients central to the success of any evaluation?

11.1 Introduction and Definitions of Terms

This chapter is about the formal study of medical information resources—computer systems that support health care, education, research, and biomedical research—to address questions of importance to developers, users, and other people. We explore the methods of performing such studies, which are essential to the field of informatics but are often challenging to carry out successfully. Fortunately, every study is not designed from a blank tablet. To guide us, there exist two closely related and highly overlapping bodies of methodological knowledge: evaluation and technology assessment. These methodological fields, which have largely developed over the past four decades, are together the subject of this chapter.[1]

[1]This chapter is heavily drawn from the textbook on evaluation by co-authors Friedman and Wyatt (2006); refer to that text for further details.

11.1.1 Evaluation and Technology Assessment

Most people understand the term evaluation to mean a measurement or description of an organized, purposeful activity. Evaluations are usually conducted to answer questions or to help make decisions. Whether we are choosing a holiday destination or a word processor, we evaluate what the options are and how well they fit key objectives or personal preferences. The forms of the evaluation differ widely, according to what is being evaluated and how important the decision is. Thus, in the case of holiday destinations, we may ask our friend which Hawaiian island she prefers and may browse color brochures from the travel agent; for a word processor, we may gather technical details, such as the time to open and spell check a 1,000-word document or the compatibility with our printer. Thus, the term **evaluation** describes a wide range of data-collection activities, designed to answer questions ranging from the casual, "What does my friend think of Maui?" to the more focused, "Is word processor A faster than word processor B on my personal computer?"

In medical informatics, we study the collection, processing, and communication of health care information and build **information resources**—usually consisting of computer hardware or software—to facilitate these activities. Such information resources include systems to collect, store, and retrieve data about specific patients (e.g., clinical workstations and databases) and systems to assemble, store, and reason about medical knowledge (e.g., medical knowledge–acquisition tools, knowledge bases, decision-support systems, and intelligent tutoring systems). Thus, there is a wide range of medical information resources to evaluate.

Further complicating the picture, each information resource has many different aspects that can be evaluated. The technically minded might focus on inherent characteristics, asking such questions as, "Is the code compliant with current software engineering standards and practices?" or "Is the data structure the optimal choice for this type of application?" Clinicians, however, might ask more pragmatic questions such as, "Is the knowledge in this system completely up-to-date?" or "How long must we wait until the decision-support system produces its recommendations?" People who have a broader perspective might wish to understand the influence of these resources on users or patients, asking questions such as, "How well does this database support a clinical audit?" or "What effects will this decision-support system have on clinical practice, working relationships, and patients?" Thus, evaluation methods in medical informatics must address a wide range of issues, from technical characteristics of specific systems to systems' effects on people and organizations.

Technology assessment is a field of study closely aligned with evaluation (Garber and Owens, 1994). The Institute of Medicine (1985, p. 2) defines technology assessment as "any process of examining and reporting properties of a medical technology used in health care, such as safety, efficacy, feasibility, and indication for use, cost, and cost effectiveness, as well as social, economic, and ethical consequences, whether intended or unintended."

But what is a medical technology? **Medical technology** usually is defined broadly and consists of the "techniques, drugs, equipments, and procedures used by health care professionals in delivering medical care to individuals, and the systems within which such

care is delivered" (Institute of Medicine, 1985, pp. 1–2). Medical information resources clearly fit within this definition. Technology assessment is relevant to informatics because many of the techniques from this field are applicable to the study of information resources.

We shall not dwell here on the differences between evaluation and technology assessment. Such differences are ones of emphasis and focus. Individuals who do evaluation and technology assessment are interested in much the same issues and use similar methods.

11.1.2 *Reasons for Performing Studies*

Like all complex and time-consuming activities, evaluation and technology assessment can serve multiple purposes. There are five major reasons why we study clinical information resources (Wyatt and Spiegelhalter, 1990):

- *Promotional*: If we are to encourage the use of information resources in medicine, we must be able to reassure physicians that these systems are safe and that they benefit both patients and institutions through improved cost effectiveness.
- *Scholarly*: One of the main activities in medical informatics is developing clinical information resources using computer-based tools. To obtain a deeper understanding of the links between the structure, function, and effects of these information resources on clinical decisions and actions requires careful evaluation. The knowledge we gain from such studies will help to build the foundations of medical informatics as a discipline (Heathfield and Wyatt, 1995).
- *Pragmatic*: Without evaluating their systems, developers will never know which techniques or methods are more effective or why certain approaches failed. Equally, other developers will not be able to learn from previous mistakes and may reinvent a square wheel.
- *Ethical*: Clinical professionals are under an obligation to practice within an ethical framework. For example, before using an information resource, health care providers must ensure that it is safe. Equally, those responsible for commissioning the purchase of a hospital-wide clinical information system costing several million dollars must be able to justify this in preference to other information resources or the many other health care innovations that compete for the same budget.
- *Medicolegal*: To reduce the risk of liability, developers of an information resource should obtain accurate information to allow them to assure users that the resource is safe and effective. Users need evaluation results to enable them to exercise their professional judgment before using systems so that the law will regard these users as "learned intermediaries." An information resource that treats users merely as automata, without allowing them to exercise their skills and judgment, risks being judged by the strict laws of product liability instead of by the more lenient principles applied to provision of professional services (Brahams and Wyatt, 1989) (also see Chapter 10).

The motivation for every study is one or more of these factors. Awareness of the major reason for conducting an evaluation will often help the investigators to frame the questions to be addressed and to avoid disappointment.

11.1.3 *The Stakeholders in Evaluation Studies and Their Roles*

Figure 11.1 shows the actors who pay for (solid arrows) and regulate (shaded arrows) the health care process. Each of them may be affected by a medical information resource, and each may have a unique view of what constitutes benefit. More specifically, in a typical clinical information resource project, the key stakeholders are the developers, the users, the patients whose management may be affected, and the people responsible for purchasing and maintaining the system. Each may have different questions to be answered (Figure 11.2).

Whenever we design evaluation or technology assessment studies, it is important to consider the perspectives of all stakeholders in the information resource. Because studies are often designed to answer specific questions, any one study is unlikely to satisfy all of the questions that concern stakeholders. Sometimes, due to the intricacy of health care systems and processes, it can be a challenge for an evaluator to identify all the relevant stakeholders and to distinguish those whose questions must be satisfied from those whose satisfaction is optional.

11.2 The Challenges of Study Design and Conduct

The work of evaluation and technology assessment in informatics lies at the intersection of three areas, each notorious for its complexity: (1) medicine and health care delivery,

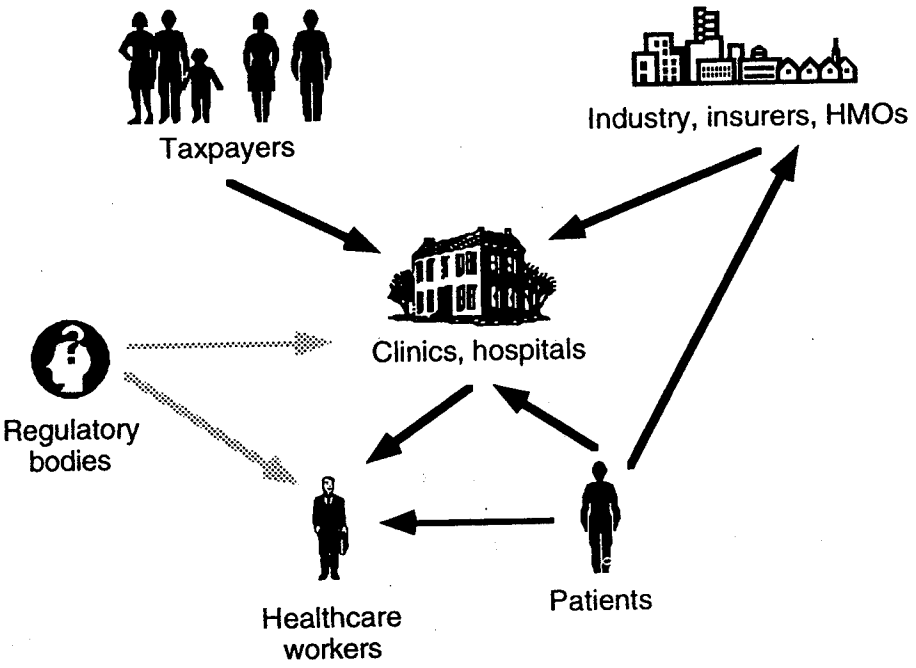

Figure 11.1. Some of the actors involved in health care delivery, administration, policy making, and regulation, each of whom may have a stake in an evaluation study. (*Source*: Friedman and Wyatt, 1997a.)

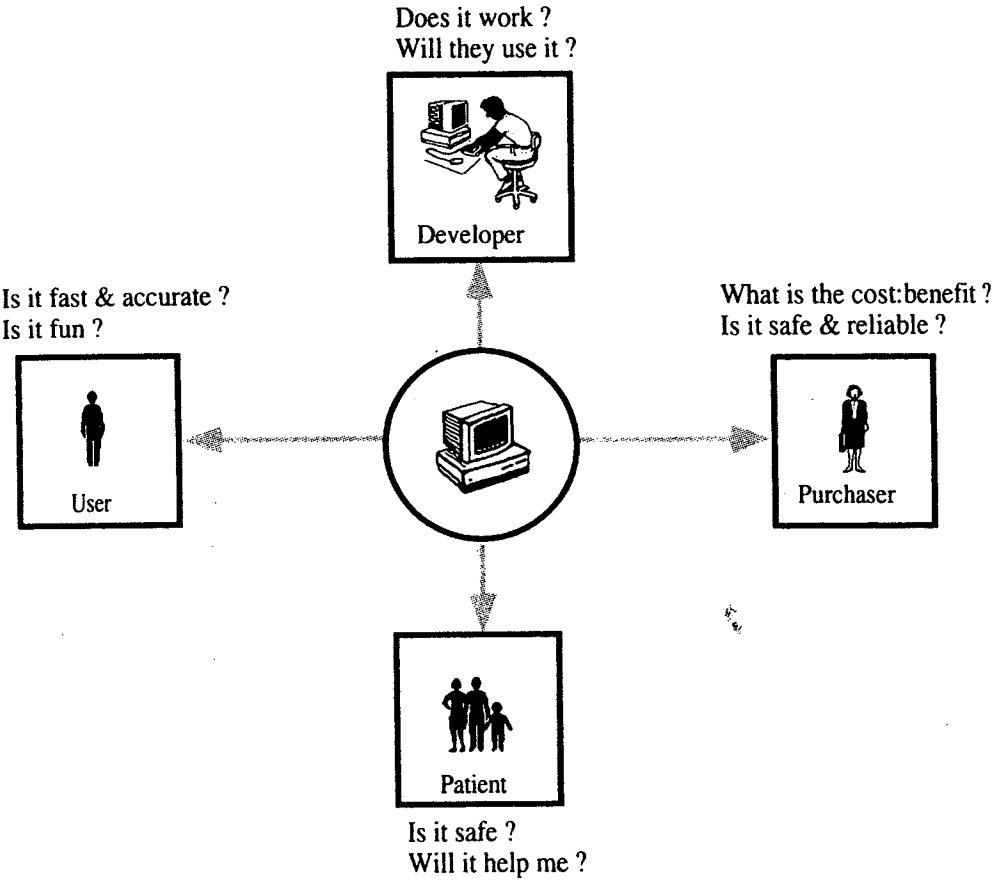

Figure 11.2. Different stakeholders may have quite different perspectives on a clinical information resource and questions that they wish to be answered by an evaluation study. (*Source*: Friedman and Wyatt, 1997a.)

(2) computer-based information systems, and (3) the general methodology of study conduct itself. Because of the complexity of each area, any work that combines them necessarily poses serious challenges.

11.2.1 The Complexity of Medicine and Health Care Delivery

Donabedian (1966) informs us that any health care innovation may influence three aspects of the health care system. The first is the health care system's **structure**, including the space it takes up; the equipment available; the financial resources required; and the number, skills, and interrelationships of the staff. The second is the **processes** that take place during health care activity, such as the number and appropriateness of diagnoses, and the investigations and therapies administered. The third is the health

care **outcomes** for both individual patients and the community, such as quality of life, complications of procedures, and length of survival. Thus, when we study the influence of an information resource on a health care system, we may see effects on any of these three aspects. An information resource may lead to an improvement in one area (e.g., patient outcomes) but to deterioration in another (e.g., the costs of running the service).

Also, it is well known that the roles of nursing and clinical personnel are well defined and hierarchical in comparison to those in many other professions. Thus information resources designed for one specific group of professionals, such as a residents' information system designed for one hospital (Young, 1980), may be of little benefit to other groups.

Because health care is a safety-critical area, with more limited budgets and a less tangible currency than, for example, retail or manufacturing, rigorous proof of safety and effectiveness is required in evaluation studies of clinical information resources. Complex regulations apply to people who develop or market clinical therapies or investigational technology. It is not yet clear whether these regulations apply to all computer-based information resources or to only those that manage patients directly, without a human intermediary (Brannigan, 1991).

Medicine is well known to be a complex domain. Students spend a minimum of 7 years gaining qualifications. A single internal-medicine textbook contains approximately 600,000 facts (Wyatt, 1991b); practicing experts have as many as 2 to 5 million facts at their fingertips (Pauker et al., 1976). Also, medical knowledge itself (Wyatt, 1991b), and methods of health care delivery, change rapidly so that the goalposts for a medical information resource may move during the course of an evaluation study.

Patients often suffer from multiple diseases, which may evolve over time at differing rates and may be subject to a number of interventions and other influences over the course of the study period, confounding the effects of changes in information management. There is even variation in how doctors interpret patient data (e.g., prostate-specific antigen results) across medical centers. Thus, simply because an information resource is safe and effective when used in one center on patients who have a given diagnosis, we are not entitled to prejudge the results of using it in another center or with patients who have a different disease profile.

The causal links between introducing an information resource and achieving improvements in patient outcome are long and complex compared with those for direct patient care interventions such as medications. In addition, the functioning and influence of an information resource may depend critically on input from health care workers or patients. It is thus unrealistic to look for quantifiable changes in patient outcomes after the introduction of many information resources until we have documented changes in the structure or processes of health care delivery.

The processes of medical decision making are complex and have been studied extensively (Elstein et al., 1978; Patel. et. al, 2001). Clinicians make many kinds of decisions—including diagnosis, monitoring, therapy, and prognosis—using incomplete and fuzzy data, some of which are appreciated intuitively and are not recorded in the clinical notes. If an information resource generates more effective management of both patient data and medical knowledge, it may intervene in the process of medical decision

making in a number of ways, making difficult the determination of which component of the resource is responsible for observed changes.

There is a general lack of gold standards in medicine. Thus, for example, diagnoses are rarely known with 100 percent certainty, because it is unethical to do all possible tests on every patient, (to follow up patients without good cause), because tests and ability to interpret them are imperfect, and because the human body is simply too complex. When a clinician attempts to establish a diagnosis or the cause of death, even if it is possible to perform a postmortem examination, correlating the patients' symptoms or clinical findings before death with the observed changes may prove impossible. Determining the correct management for a patient is even more complicated, because there is wide variation in **consensus opinions** (Leitch, 1989), as reflected in wide variations in clinical practice even in neighboring areas.

Doctors practice under strict legal and ethical obligations to give their patients the best care that is available, to do patients no harm, to keep patients informed about the risks of all procedures and therapies, and to maintain confidentiality. These obligations may well impinge on the design of evaluation studies. For example, because health care workers have imperfect memories and patients take holidays and participate in the unpredictable activities of real life, it is impossible to impose strict discipline in data recording, and study data are often incomplete. Similarly, before a randomized controlled trial can be undertaken, health care workers and patients are entitled to a full explanation of the possible benefits and disadvantages of being allocated to the control and intervention groups before giving their consent.

11.2.2 The Complexity of Computer-Based Information Resources

From the perspective of a computer scientist, the goal of evaluating a computer-based information resource might be to predict that resource's function and effects from a knowledge of its structure. Although software engineering and formal methods for specifying, coding, and evaluating computer programs have become more sophisticated, even systems of modest complexity challenge these techniques. To formally verify a program rigorously (to obtain proof that it performs all and only those functions specified), we must invest effort that increases exponentially with the program's size—the problem is "**NP hard.**" Put simply, to test a program rigorously requires the application of every combination of possible input data in all possible orders. Thus, it entails at least n factorial experiments, where n is the number of input data items. The size of n factorial increases exponentially with small increases in n, so the task rapidly becomes unfeasible. In some technology-led projects, the goals of the new information resources are not defined precisely. Developers may be attracted by technology and may produce applications without first demonstrating the existence of a clinical problem that the application is designed to meet (Heathfield and Wyatt, 1993). An example was a conference entitled "Medicine Meets Virtual Reality: Discovering Applications for 3D Multimedia." The lack of a clear need for an information resource makes it hard to evaluate the ability of the information resource to alleviate a clinical problem. Although one can still evaluate

the structure and function of the system in isolation, it will be hard to interpret the results of such an evaluation in clinical terms.

Some computer-based systems are able to adapt themselves to their users or to data already acquired, or they may be deliberately tailored to a given institution; it may then be difficult to compare the results of one evaluation with a study of the same information resource conducted at a different time or in another location. Also, the notoriously rapid evolution of computer hardware and software means that the time course of an evaluation study may be greater than the lifetime of the information resource itself.

Medical information resources often contain several distinct components, including the interface, database, reasoning and maintenance programs, patient data, static medical knowledge, and dynamic inferences about the patient, the user, and the current activity of the user. Such information resources may perform a wide range of functions for users. Thus, if evaluators are to answer questions such as, "What part of the information resource is responsible for the observed effect?" or "Why did the information resource fail?" they must be familiar with each component of the information resource, know its functions, and understand potential interactions (Wyatt, 1989, 1991a).

11.2.3 The Complexity of Study Methods

Studies do not focus solely on the structure and function of an information resource, they also address the resource's effects on the care providers who are customarily its users and on patient outcomes. To understand users' actions, investigators must confront the gulf between peoples' private opinions, public statements, and actual behavior. Humans vary widely in their responses to stimuli, both from minute to minute and from one to another, making the results of measurements subject to random and systematic errors. Thus, studies of medical information resources require analytical tools from the behavioral and social sciences, statistics, and other fields.

Studies require test material, such as clinical cases, and information resource users, such as physicians or nurses. Both are often in shorter supply than the study design requires; the availability of patients also is usually overestimated, sometimes many times over. In addition, it may be unclear what kind of cases or users should be recruited for a study. Often, study designers are faced with a trade-off between selecting cases, users, and study settings with high fidelity to real life and selecting those who will help to achieve adequate experimental control. Finally, one of the more important determinants of the results of an evaluation study is the manner in which case data are abstracted and presented to users. For example, we would expect differing results in a study of an information resource's accuracy depending on whether the test data were abstracted by the developers or by the intended users.

There are many reasons for performing studies, ranging from assessing a student's work to formulating health policy to understanding a specific technical advance. Such reasons will in turn determine the kinds of questions that will be asked about the information resource. To help those who are trying to determine the broad goals of an evaluation study, in Table 11.1 we list some of the many questions that can arise about information resources and about their influence on users, patients, and the health care system.

Table 11.1. Possible questions that may arise during the study of a medical information resource.

About the resource itself	About the resource's impact
Is there a clinical need for it?	Do people use it?
Does it work?	Do people like it?
Is it reliable?	Does it improve users' efficiency?
Is it accurate?	Does it influence the collection of data?
Is it fast enough?	Does it influence users' decisions?
Is data entry reliable?	For how long do the observed effects last?
Are people likely to use it?	Does it influence users' knowledge or skills?
Which parts cause the effects?	Does it help patients?
How can it be maintained?	Does it change consumption of resources?
How can it be improved?	What might ensue from widespread use?

(*Source*: Friedman and Wyatt, 1997a.)

11.3 The Full Range of What Can Be Studied

When evaluating a medical information resource, there are five major aspects of interest: (1) the clinical need the resource is intended to address, (2) the process used to develop the resource, (3) the resource's intrinsic structure, (4) the functions that the resource carries out, and (5) the resource's effects on users, patients, and other aspects of the clinical environment. In a theoretically complete evaluation, separate studies of a particular resource might address each aspect. In the real world, however, it is difficult to be comprehensive. Over the course of its development and deployment, a resource may be studied many times with the studies in their totality touching on many or most of these aspects, but few resources will be studied completely and many will, inevitably, be studied only minimally.

The evaluation focus changes as we study the different aspects:

1. *The need for the resource*: Evaluators study the clinical status quo absent the resource. They determine the nature of the problems that the resource is intended to address and the frequency with which these problems arise.
2. *The development process*: Evaluators study the skills of the development team and the methodologies employed to understand whether the design is likely to be sound.
3. *The resource's intrinsic structure*: Evaluators study specifications, flowcharts, program codes, and other representations of the resource that they can inspect without running the program.
4. *The resource's functions*: Evaluators study how the resource performs when it is used.
5. *The resource's effects*: Evaluators study not the resource itself but rather its influence on users, patients, and health care organizations.

Several factors characterize an evaluation study:

- *The focus of study*: The focus can be the status quo before introduction of the information resource, the design process adopted, the resource's structure or function, the resource users' simulated decisions or real decisions, or the clinical actions and patient outcomes once the resource is made available in the workplace.

- *Study setting*: Studies of the design process, the resource's structure, and the resource's functions can be conducted outside the active clinical environment, in a laboratory setting, which is easier logistically and may allow greater control over the evaluation process. Studies to elucidate the need for a resource and studies of the resource's effects on users both usually take place in clinical settings. The effects of a resource on patients and health care organizations can take place in only a true clinical setting where the resource is available for use at the time and place where patient-management decisions are made.
- *Clinical data employed*: For many studies, the resource will actually be run. That will require clinical data, which can be simulated data, data abstracted from real patients' records, or actual patient data. Clearly, the kind of data employed in a study has serious implications for the study results and the conclusions that can be drawn.
- *User of the resource*: Most information resources function in interaction with one or more users. In any particular study, the users of the resource can be members of the development team or the evaluation team, or other individuals not representative of those people who will interact with the resource after it is deployed; or the users in a study could be representative of the end users for whose use the resource is ultimately designed. Again, the selection of resource users can affect study results profoundly.
- *The decisions affected by use of the resource*: Many information resources, by providing information or advice to clinicians, seek to influence the decisions made by these clinicians. As a study moves from the laboratory to the clinical setting, the information provided by the resource potentially has greater implications for the decisions being made. Depending on a study's design and purposes, only simulated decisions may be affected (clinicians are asked what they would do, but no action is taken), or real decisions involved in the care of actual patients may be affected.

Table 11.2 lists nine broad types of studies of clinical information resources that can be conducted: the focus of each type, the setting in which it occurs, the kind of clinical data employed as input to the resource, the person who uses the resource during the study, and the kind of clinical decisions affected by the resource during the study. For example, a laboratory-user impact study would be conducted outside the active clinical environment based on simulated or abstracted clinical data. Although it would involve individuals representative of the end-user population, the study would yield primary results derived from simulated clinical decisions, so the clinical care of patients would not be affected. Read across each row of the table to obtain a feel for the contrasts among these study types.

11.4 Approaches to Study Design

Having established a large number of reasons why it can be difficult to study medical information resources, we now introduce the methods that have been developed to address these challenges. We begin by describing a generic structure that all studies share. Then we introduce, in turn, more specific methods of evaluation and the closely related methods of technology assessment.

Tables 11.2 Generic types of evaluation studies of clinical information resources.

Study type	Study setting	Version of the Resource	Sampled users	Sampled tasks	What is observed
1. Needs Assessment	Field	None, or pre-existing resource to be replaced	Anticipated resource users	Actual tasks	User skills, knowledge, decisions

Tables 11.2 Generic types of evaluation studies of clinical information resources.

Study type	Study setting	Version of the Resource	Sampled users	Sampled tasks	What is observed
1. Needs Assessment	Field	None, or pre-existing resource to be replaced	Anticipated resource users	Actual tasks	User skills, knowledge, decisions or actions; care processes, costs, team function or organization; patient outcomes
2. Design Validation	Development lab	None	None	None	Quality of design method or team
3. Structure Validation	Lab	Prototype or released version	None	None	Quality of resource structure, components, architecture
4. Usability Test	Lab	Prototype or released version	Proxy, real users	Simulated, abstracted	Speed of use, user comments, completion of sample tasks
5. Laboratory Function Study	Lab	Prototype or released version	Proxy, real users	Simulated, abstracted	Speed & quality of data collected or displayed; accuracy of advice given
6. Field Function Study	Field	Prototype or released version	Proxy, real users	Real	Speed & quality of data collected or displayed; accuracy of advice given
7. Lab User Effect Study	Lab	Prototype or released version	Real users	Abstracted, real	Impact on user knowledge, simulated / pretend decisions or actions
8. Field User Effect Study	Field	Released version	Real users	Real	Extent and nature of resource use. Impact on user knowledge, real decisions, real actions
9. Problem Impact Study	Field	Released version	Real users	Real	Care processes, costs, team function, cost effectiveness

11.4.1 The Anatomy of All Studies

The structural elements that all studies share are illustrated in Figure 11.3. Evaluations are guided by someone's or some group's need to know. No matter who that someone is—the development team, the funding agency, or other individuals and groups—the evaluation must begin with a process of negotiation to identify the questions that will be a starting point for the study. The outcomes of these negotiations are an understanding of how the evaluation is to be conducted, usually stated in a written contract or agreement, and an initial expression of the questions the evaluation seeks to answer. The next element of the study is investigation: the collection of data to address these questions and, depending on the approach selected, possibly other questions that arise during the study. The mechanisms are numerous, ranging from the performance of the resource on a series of benchmark tasks to observations of users working with the resource.

The next element is a mechanism for reporting the information back to the individuals who need to know it. The format of the report must be in line with the stipulations of the contract; the content of the report follows from the questions asked and the data collected. The report is most often a written document, but it does not have to be—the purposes of some evaluations are well served by oral reports or by live demonstrations. We emphasize that it is the evaluator's obligation to establish a process through which the results of her study are communicated, thus creating the potential for the study's findings to be put to constructive use. No investigator can guarantee a constructive outcome for a study, but there is much they can do to increase the likelihood of a salutary result. Also note that a salutary result of a study is not necessarily one that casts the resource under study in a positive light. A salutary result is one where the stakeholders learn important information from the study findings.

11.4.2 Philosophical Bases of Approaches to Evaluation

Several authors have developed classifications, or **typologies**, of evaluation methods or approaches. Among the best is that developed in 1980 by Ernest House. A major

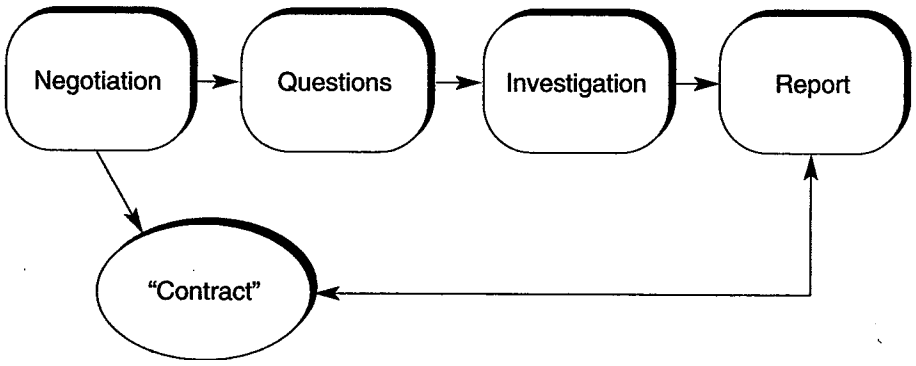

Figure 11.3. Anatomy of all evaluation studies. (*Source*: Friedman and Wyatt, 1997a.)