Evaluation of Biomedical and Health Information Resources

11

Charles P. Friedman and Jeremy C. Wyatt

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?
- 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 features do all evaluations have in common?
- What are the key factors to take into account as part of a process of deciding what are the most important questions to use to frame a study?
- What are the major assumptions underlying objectivist and subjectivist approaches to evaluation? What are the strengths and weaknesses of each approach?
- How does one distinguish measurement and demonstration aspects of objectivist studies, and why are both aspects necessary? In the demonstration aspect of objectivist studies, how are control strategies used to draw inferences?
- 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

Most people understand the term evaluation to mean an assessment of an organized, purposeful activity. Evaluations are usually conducted to answer questions or in anticipation of the need to make decisions (Wyatt and Spiegelhalter 1990). Evaluations may be informal or formal, depending on the characteristics of the decision to be made and, particularly, how much is at stake. But all activities labeled as evaluation involve the empirical process of collecting information that is relevant to the decision at hand. For example, when choosing a holiday destination, members of a family may informally ask friends which

C.P. Friedman, PhD, FACMI (✉)
Schools of Information and Public Health,
University of Michigan, 105 S State St,
Ann Arbor, MI 48109, USA
e-mail: cpfried@umich.edu

J.C. Wyatt, MB BS, FRCP, FACMI
Leeds Institute of Health Sciences,
University of Leeds, Charles Thackrah Building,
101 Clarendon Road, Leeds LS2 9LJ, UK
e-mail: j.c.wyatt@leeds.ac.uk

This chapter is adapted from an earlier version in the third edition authored by Charles P. Friedman, Jeremy C. Wyatt, and Douglas K. Owens.

E.H. Shortliffe, J.J. Cimino (eds.), *Biomedical Informatics*,
DOI 10.1007/978-1-4471-4474-8_11, © Springer-Verlag London 2014

Hawaiian island they prefer and browse various websites including those that provide ratings of specific destinations. After factoring in costs and convenience, the family reaches a decision. More formally, when a health care organization faces the choice of a new electronic health record system, the leadership will develop a plan to collect comparable data about competing systems, analyze the data according to the plan, and ultimately, again through a predetermined process, make a decision.

The field of biomedical and health informatics focuses on the collection, processing, and communication of health related information and the implementation of **information resources**—usually consisting of digital technology designed to interact with people—to facilitate these activities. These information resources can collect, store, and process data related to the health of individual persons (institutional or personal electronic health records), manage and reason about biomedical knowledge (knowledge acquisition tools, knowledge bases, decision-support systems, and intelligent tutoring systems), and support activities related to public health (disease registries and vital statistics, disease outbreak detection and tracking). Thus, there is a vast range of biomedical and health information resources that can be foci of evaluation.

Information resources have many different aspects that can be studied (Friedman and Wyatt 2005, Chap. 3). Where safety is an issue, as it often is, (Fox 1993), we might focus on inherent characteristics of the resource, asking such questions as, "Are the code and architecture 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 "Can I retrieve all the information about a patient or just the information generated in my own clinic?" Executives and public officials might wish to understand the effects of these resources on individuals and populations, asking questions such as, "Has this resource improved the quality of care?" or "What effects will a patient portal have on working relationships between practitioners and patients?" Thus, evaluation methods in biomedical informatics must address a wide range of issues, from technical characteristics of specific systems to systems' effects on people and organizations. The outcomes or effects attributable to the use of health information resources will almost always be a function of how individuals choose to use them, and the social, cultural, organizational, and economic context in which these uses take place (Lundsgaarde 1987).

For these reasons, there is no formula for designing and executing evaluations; every evaluation, to some significant degree, must be custom-designed. In the end, decisions about what evaluation questions to pursue and how to collect and analyze data to pursue them, are exquisitely sensitive to each study's special circumstances and constrained by the resources that are available for it. Evaluation is very much the art of the possible. But neither is evaluation an exercise in alchemy, pure intuition, or black magic. There exist many methods for evaluation that have stood the test of time and proved useful in practice. There is a literature on what works and where, and there are numerous published examples of successful evaluation studies. In this chapter, we will introduce many of these methods, and present frameworks that guide the application of methods to specific decision problems and study settings.

11.2 Why Are Formal Evaluation Studies Needed?

11.2.1 Computing Artifacts Have Special Characteristics

Why are empirical studies of information resources needed at all? Why is it not possible, for example, to model (and thus predict) the performance of information resources, and thus save a lot of time and effort? The answer lies, to a great extent, in the complexity of computational artifacts and their use. For some disciplines, specification of the structure of an artifact allows one to predict how it will function, and engineers

can even design new objects with known performance characteristics directly from functional requirements. Examples of such artifacts are elevators and conventional road bridges: The principles governing the behavior of materials and structures made of these materials are sufficiently well understood that a new elevator or bridge can be designed to a set of performance characteristics with the expectation that it will perform exactly as predicted. Laboratory testing of models of these devices is rarely needed. Field testing of the artifact, once built, is conducted to reveal relatively minor anomalies, which can be rapidly remedied, or to tune or optimize performance. However, when the object concerned is a computer-based resource, not a bridge, the story is different (Littlejohns et al. 2003). Software designers and engineers have theories linking the structure to the function of only the most trivial computer-based resources (Somerville 2002). Because of the complexity of computer-based systems themselves, their position as part of a complex socio-technical system including the users and the organization in which they work, and the lack of a comprehensive theory connecting structure and function, there is no way to know exactly how an information resource will perform until it is built and tested (Murray 2004); and similarly there is no way to know that any revisions will bring about the desired effect until the next version of the resource is tested.

In sum, the only practical way to determine if a reasonably complex body of computer code does what it is intended to do is to test it. This testing can take many shapes and forms. The informal design, test, and revise activity that characterizes the development of all computer software is one such form of testing and results in software that usually functions as expected *by the developers*. More formal and exhaustive approaches to software design, verification and testing using synthetic test cases (e.g., Scott et al. 2011) and other approaches help to guarantee that the software will do what it was designed to do. Even these approaches, however, do not guarantee the success of the software when put into the hands of the intended end-users. This requires more formal studies of the types that will be described in this chapter, which can be undertaken before, during, and after the initial development of an information resource. Such evaluation studies can guide further development; indicate if the resource is likely to be safe for use in real health care, public health, research, or educational settings; or elucidate if it has the potential to improve the professional performance of the users and disease outcomes in their clients.

Many other writings elaborate on the points offered here. Some of the earliest include Spiegelhalter (1983) and Gaschnig et al. (1983) who discussed these phases of evaluation by drawing analogies from the evaluation of new drugs or the conventional software life cycle, respectively. Wasson et al. (1985) discussed the evaluation of clinical prediction rules together with some useful methodological standards that apply equally to information resources. Many other authors since then have described, with differing emphases, the evaluation of health care information resources, often focusing on decision-support tools, which pose some of the most extreme challenges. One relevant book (Friedman et al. 2005) discusses the challenges posed by evaluation in biomedical informatics and offers a wide range of methods described in considerable detail to help investigators explore and resolve these challenges. Other books have explored more technical, health technology assessment or organizational approaches to evaluation methods (Szczepura and Kankaanpaa 1996; van Gennip and Talmon 1995; Anderson et al. 1994; Brender 2005).

11.2.2 The Special Issue of Safety

Before disseminating any biomedical information resource that stores and communicates health data or knowledge and is designed to influence real-world practice or personal health decisions, it is important to verify that the resource is safe when used as intended. In the case of new drugs, European and US regulators have imposed a statutory duty on developers to perform extensive in vitro testing, and in vivo testing in animals, before any human receives a dose of the drug.

Analogous testing is currently not required of information resources. However, since 2000, the safety of biomedical information resources has come increasingly into the spotlight (Rigby et al. 2001; Koppel et al. 2005). For biomedical information resources, safety tests analogous to those required for drugs would include assessment of the accuracy of the data stored and retrieved, determining whether and how easily end-users can employ the resource for its intended purposes, and estimating how often the resource furnishes misleading or incorrect information (Eminovic et al. 2004). It may be necessary to repeat these assessments following any substantial modifications to the information resource, as the correction of safety-related problems may itself generate new problems or uncover previously unrecognized ones.

Determining if an information resource is safe and effective goes fundamentally to the process of evaluation we address in this chapter. All of the methodological issues we raise apply to safety assessments. Casual assessments that fail to address these issues will not resolve the safety question, and will not reveal safety defects that can be remedied. Many of these issues are issues of sampling that we introduce in Sect. 11.4.2. For example, the advice or other "output" generated by most information resources depends critically on the quality and quantity of data available to it and on the manner in which the resource is used by patients or practitioners. People or practitioners who are untrained, in a hurry, or exhausted at 3 A.M., are more likely to fail to enter key data that might lead to the resource generating misleading advice, or to fail to heed an alarm that is not adequately emphasized by the user interface. Thus, to generate valid results, functional tests must put the resources in actual users' hands under the most realistic conditions possible, or in the hands of people with similar knowledge, skills and experience if real users are not available.

Other safety issues are, from a methodological perspective, issues of measurement that we address in Sect. 11.4.2. For example, should "usability" of an information resource be determined by documenting that the resource development process followed best practices to inculcate usability, asking end-users if they believed the resource was usable, or by documenting and studying their "click streams" to determine if end-users actually navigated the resource as the designers intended? There is no single clear answer to this question (see Jackob Nielsen's invaluable resource on user testing[1]), but we will see that all measurement processes have features that make their results more or less dependable and useful. We will also see that the measurement processes built into evaluation studies can themselves be designed to make the results of the studies more helpful to all stakeholders, including those focused on safety.

11.3 Two Universals of Evaluation

11.3.1 The Full Range of What Can Be Formally Studied

Deciding what to study is fundamentally a process of winnowing down from a universe of potential questions to a parsimonious set of questions that can be realistically addressed given the priorities, time, and resources available. This winnowing process can begin with the full range of what can potentially be studied. To both ensure that the most important questions do get "on the table" and to help eliminate the less important ones, it can be useful to start with such a comprehensive list. While experienced evaluators do not typically begin study planning from this broadest perspective, it is always helpful to have the full range of options in mind.

There are five major aspects of an information resource, or an identified class of resources, that can be studied:

1. Need for the resource: Investigators study the status quo *absent* the resource, including the nature of problems the resource is intended to address and how frequently these problems arise. (When an information resource is already deployed, the "status quo" might be the currently deployed resource, and the resource under study is a proposed replacement for it or enhancement to it.)

[1] www.useit.com (Accessed 4/19/13).

2. Design and development process: Investigators study the skills of the development team, and the development methodologies employed by the team, to understand if the resulting resource is likely to function as intended.
3. Resource static structure: Here the focus of the evaluation includes specifications, flow charts, program code, and other representations of the resource that can be inspected without actually running it.
4. Resource usability and dynamic functions: The focus is on the usability of the resource and how it performs when it is used in pilots prior to full deployment.
5. Resource use, effect and impact: Finally, after deployment, the focus switches from the resource itself to the extent of its use and its effects on professional, patient or public users, and health care organizations.

In a theoretically "complete" evaluation, sequential studies of a particular resource might address all of these aspects, over the life cycle of the resource. In the real world, however, it is difficult, and rarely necessary, to be so 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. Some aspects of an information resource will be studied informally using anecdotal data collected via casual methods. Other aspects will be studied more formally in ways that are purposefully designed to inform specific decisions and that involve systematic collection and analysis of data. Distinguishing those aspects that will be studied formally from those left for informal exploration is a task facing all evaluators.

11.3.2 The Structure of All Evaluation Studies, Beginning with a Negotiation Phase

If the list offered in the previous section can be seen as the universe of what can be studied, Fig. 11.1 can be used as a framework for planning all evaluation studies. The first stage in any study is negotiation between the individuals who will be carrying out the study (the "evaluators") and the "stakeholders" who have interests in or otherwise will be concerned about the study results. Before a study can proceed, the key stakeholders who are supporting the study financially and providing other essential resources for it—such as the institution where the information resource is deployed—must be satisfied with the general plan. The negotiation phase identifies the broad aim and objectives of the study, what kinds of reports and other deliverables will result and by when, where the study personnel will be based, the resources available to conduct the study, and any constraints on what can be studied.

The results of the negotiation phase are expressed in a document, generally known as a contract between the evaluators and the key stakeholders. The contact guides the planning and execution of the study and, in a very significant way, protects all parties from misunderstandings about intent and execution. Like any contract, an **evaluation contract** can be changed later with consent of all parties.

Following the negotiation process and its reflection in a contract, the planning of the evaluation proceeds in a sequence of logical steps, starting with the formulation of specific questions to be addressed, then the selection of the

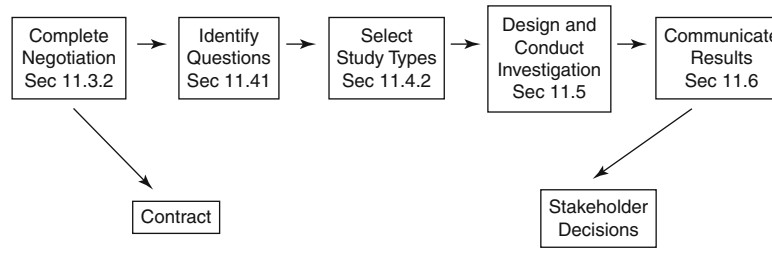

Fig. 11.1 Generic structure of all evaluation studies

type(s) of study that will be used, the investigation that entails the collection and analysis of data, and ultimately the communication of the findings back to the stakeholders, which typically inform a range of decisions. Although Fig. 11.1 portrays a one-way progression through this sequence of stages, in the real world of evaluation there are often detours and backtracks.

11.4 Deciding What to Study and What Type of Study to Do: Questions and Study Types

11.4.1 The Importance of Identifying Questions

Once the study's scope and other applicable "ground rules" have been established, the real work of study planning can begin. The next step, as suggested by Fig. 11.1, is to convert the perspectives of the concerned parties, and what these individuals or groups want to know, into a finite, specific set of questions. It is important to recognize that, for any evaluation setting that is interesting enough to merit formal evaluation, the number of potential questions is infinite. This essential step of identifying a tractable number of questions has a number of benefits:

- It helps to crystallize thinking of both investigators and key members of the audience who are the stakeholders in the evaluation.
- It guides the investigators and clients through the critical process of assigning priority to certain issues and thus productively narrowing the focus of a study.
- It converts broad statements of aim (e.g., "*to evaluate a new order communications system*") into specific questions that can potentially be answered (e.g., "*What is the impact of the order communications system on how clinical staff spend their time*, *the rate and severity of adverse drug events and the length of patient stay*?").
- It allows different stakeholders in the evaluation process—patients, professional groups, managers – to see the extent to which their own concerns are being addressed, and to ensure that these feed into the evaluation process.
- Most important, perhaps, it is hard if not impossible to develop investigative methods without first identifying questions, or at least focused issues, for exploration. The choice of methods follows from the evaluation questions: not from the novel technology powering the information resource or the type of resource being studied. Unfortunately, some investigators choose to apply the same set of the methods to any study, irrespective of the questions to be addressed, or even to limit the evaluation questions addressed to those compatible with the methods they prefer. We do not endorse this approach.

Consider the distinction made earlier between informal evaluations that people undertake continuously as they make choices as part of their everyday personal or professional lives, and more formal evaluations that are planned and then executed according to that plan. In short, formal evaluations are those that conform to the architecture of Fig. 11.1. In these formal evaluations, the questions that actually get addressed survive a narrowing process that begins with a broad set of candidate questions. When starting a formal evaluation, therefore, a major decision is whom to consult to establish the questions that will get "on the table", how to log and analyze their views, and what weight to place on each of these views. There is always a wide range of potential players in any evaluation (see Box 11.1) and there is no formula that defines whom to consult or in what order. Through this process, the investigators apply their common sense and, with experience, learn to follow their instincts; it is often useful to establish a steering group to advise the evaluators to ensure that their efforts remain true to the interests and preferences of the stakeholders. The only universal mistake is to fail to consult one or more of the key stakeholders, especially those paying for the study or those ultimately making the key decisions to be informed by the evaluation.

Box 11.1: Some of the Potential Players in an Evaluation Study

- Those commissioning the evaluation study, who will typically have questions or decisions that rely on the data collected
- Those paying for the evaluation study
- Those paying for the development and/or deployment of the resource
- End-users of the resource, who are often providers of data for the study
- Developers of the resource and their managers
- Care providers and their managers
- Staff responsible for resource implementation and user training
- Information technology staff and leaders in the organization where the resource is deployed
- Senior managers in the organization where the resource is deployed
- The patients whose care the resource may directly or indirectly influence
- Quality improvement and safety professionals in the organization in which the resource is implemented

Through discussions with various stakeholder groups, the hard decisions regarding the questions to be addressed in the study are made. A significant challenge for investigators is the risk of getting swamped by detail resulting from the multiplicity of questions that can be asked in any study. To manage through the process, reflect on the major issues identified after each round of discussions with stakeholders, and then identify the questions that map to these issues. Where possible keep questions at the same level of granularity. It is important to keep a sense of perspective, distinguishing the issues as they arise and organizing them into some kind of hierarchy, for example low or operational level issues, tactical or medium level and high level strategic issues. Interdependencies should be noted and care should be taken to avoid intermingling global issues with more focused issues. For example, when evaluating an electronic lab notebook system for researchers, it is important to distinguish operational, low level issues, such as the time taken to enter data for equipment orders, from strategic issues such as the impact of the resource on research productivity. While this distinction may seem trivial in the abstract, in practice these issues often get muddled as different stakeholders argue for their particular needs and interests to be represented.

It is critical that the specific questions serving as the beacon guiding the study be determined and endorsed by all key stakeholders, before any significant decisions about the detailed design of the study are made. We will see later that evaluation questions can, in many circumstances, change over the course of a study; but that fact does not obviate the need to specify a set of questions at the outset.

Consider as an example a new information resource that sends SMS text messages to patients with chronic illnesses, reminding them of upcoming appointments, to take their medication (e.g. Lester et al. 2010) or about other key events (for example, recurring blood draws) that are important to their care. An example set of initial negotiated questions for this study is shown below, along with the stakeholder groups that will have direct or primary interest in each question.

11.4.2 Selecting a Study Type

After developing evaluation questions, the next step is to understand which study type(s) the evaluation questions naturally invoke. These study types are specific to the study of information resources, and are particularly informative to the design of evaluation studies in biomedical informatics. The study types are described below and also summarized in Table 11.1. The second column of Table 11.1 links the study types to the aspect of the resource that is studied, as introduced in Sect. 11.3.1. Each study type is likely to appeal to certain stakeholders in the evaluation process, as suggested in the rightmost column of the table. A wide range of data collection and analysis methods, as discussed in Sect. 11.5, can

Table 11.1 Classification of generic study types by broad study question and the stakeholders most concerned

Study type	Aspect studied	Broad study question	Audience/stakeholders primarily interested in results
1. Needs assessment	Need for the resource	What is the problem?	Resource developers, funders of the resource
2. Design validation	Design and development process	Is the development method in accord with accepted practices?	Funders of the resource; professional and governmental certification agencies e.g., Food and Drug Administration, Office of the National Coordinator for HIT
3. Structure validation	Resource static structure	Is the resource appropriately designed to function as intended?	Professional indemnity insurers, resource developers; professional and governmental certification agencies
4. Usability test	Resource dynamic usability and function	Can intended users navigate the resource so it carries out intended functions?	Resource developers, users, funders
5. Laboratory function study	Resource dynamic usability and function	Does the resource have the potential to be beneficial?	Resource developers, funders, users, academic community
6. Field function study	Resource dynamic usability and function	Does the resource have the potential to be beneficial in the real world?	Resource developers, funders, users
7. Lab user effect study	Resource effect and impact	Is the resource likely to change user behavior?	Resource developers and funders, users
8. Field user effect study	Resource effect and impact	Does the resource change actual user behavior in ways that are positive?	Resource users and their clients, resource purchasers and funders
9. Problem impact study	Resource effect and impact	Does the resource have a positive impact on the original problem?	The universe of stakeholders

be used to answer the questions embraced by all nine study types. *Choice of a study type typically does not constrain the methods that can be used to collect and analyze data.*

1. **Needs assessment** studies seek to clarify the information problem the resource is intended to solve. These studies take place before the resource is designed–usually in the setting where the resource is to be deployed, although simulated settings may sometimes be used. Ideally, these potential users will be studied while they work with real problems or cases, to understand better how information is used and managed, and to identify the causes and consequences of inadequate information flows. The investigator seeks to understand users' skills, knowledge and attitudes, as well as how they make decisions or take actions. An example is a study on 300 primary care physicians to understand their trade-offs among the reliability of an electronic patient record, where they could access the resource, and who could have access to it (Wyatt et al. 2010). To ensure that developers have a clear model of how a proposed information resource will fit with working practices and structures, they may also need to study health care or research processes, team functioning, or relevant aspects of the larger organization in which work is done.
2. **Design validation** studies focus on the quality of the processes of information resource design and development, for example by asking experts to review these processes. The experts may review documents, interview the development team, compare the suitability of the software engineering methodology and programming tools used with others that are available, and generally apply their expertise to identify potential flaws in the approach

used to develop the software, as well as constructively to suggest how these might be corrected.

3. **Structure validation** studies address the static form of the software, usually after a first prototype has been developed. This type of study is most usefully performed by an expert or a team of experts with experience in developing software for the problem domain and concerned users. For these purposes, the investigators need access to both summary and detailed documentation about the system architecture, the structure and function of each module, and the interfaces among them. The expert might focus on the appropriateness of the algorithms that have been employed and check that they have been correctly implemented. Experts might also examine the data structures (e.g., whether they are appropriately normalized) and knowledge bases (e.g., whether they are evidence-based, up to date, and modelled in a format that will support the intended analyses or reasoning). Most of this will be done by inspection and discussion with the development team. Sometimes specialized software may be used to test the structure of the resource (Somerville 2002).

   Note that the study types listed up to this point do not require a functioning information resource. However, beginning with usability testing below, the study types require the existence of at least a functioning prototype.

4. **Usability testing** studies focus on system function and addresses whether intended users can actually operate or navigate the software, to determine whether the resource has the potential to be helpful to them (see also Chap. 4). In this type of study, use of a prototype by typical users informs further development and should improve its usability. Although usability testing is often performed by obtaining opinions of usability experts, usability can also be tested by deploying the resource in a laboratory or classroom setting, introducing users to it, and then allowing them either to navigate at will and provide unstructured comments or to attempt to complete some scripted tasks (see Nielsen, www.useit.com). Data can be collected by the computer itself, from the user, by a live observer, via audio or video capture of users' actions and statements, or by specialized instrumentation such as eye-tracking tools. Many software developers have usability testing labs equipped with sophisticated measurement systems, staffed by experts in human computer interaction to carry out these studies—an indication of the importance increasingly attached to this type of study.
5. **Laboratory function studies** go beyond usability to explore more specific aspects of the information resource, such as the quality of data captured, the speed of communication, the validity of the calculations carried out, or the appropriateness of advice given. These functions relate less to the basic usability of the resource and more to how the resource performs in relation to what it is trying to achieve for the user or the organization. When carrying out any kind of function testing, the results will depend crucially on what problems the users are asked to solve, so the "tasks" employed in these studies should correspond as closely as possible to those to which the resource will be applied in real working life.
6. **Field function studies** are a variant of laboratory function testing in which the resource is "pseudo-deployed" in a real work place and employed by real users, up to a point. However, in field function tests, although the resource is used by real users with real tasks, there is no immediate access by the users to the output or results of interaction with the resource that might influence their decisions or actions, so no effects on these can occur. The output is recorded for later review by the investigators, and perhaps by the users themselves.

   Studies of the effect or impact of information resources on users and problems are in many ways the most demanding. As the focus of its study moves from its functions to its possible effects on health decisions or care processes, the conduct of research, or educational practice, there is often the need to establish cause and effect.

7. In **laboratory user effect studies**, simulated user actions are studied. Practitioners employ the resource and are asked what they "would do" with the results or advice the resource generates, but no action is taken. Laboratory user effect studies are conducted with prototype or released versions of the resource, outside the practice environment. Although such studies involve individuals who are representative of the "end-user" population, the primary results of the study derive from simulated actions, so the care of patients or conduct of research is not affected by a study of this type. An example is a study in which junior doctors viewed realistic prescribing scenarios and interacted with a simulated prescribing tool while they were exposed to simulated prescribing alerts of various kinds (Scott 2011).
8. In a **field user effect study**, the actual actions or decisions of the users of the resource are studied after the resource is formally deployed. This type of study provides an opportunity to test whether the resource is actually used by the intended users, whether they obtain accurate and useful information from it, and whether this use affects their decisions and actions in significant ways. In field user effect studies, the emphasis is on the behaviors and actions of users, and not the consequences of these behaviors. For example, one study examined the impact of SMS reminders on anti-retroviral medication adherence in Africans with HIV and showed a dramatic improvement (Lester et al. 2010).
9. **Problem impact studies** are similar to field user effect studies in many respects, but differ profoundly in the questions that are the focus of exploration. Problem impact studies examine the extent to which the original problem that motivated creation or deployment of the information resource has been addressed. Often this requires investigation that looks beyond the actions of care providers, researchers, or patients to examine the consequences of these actions. For example, an information resource designed to reduce medical errors may affect the behavior of some clinicians who employ the resource, but for a variety of reasons, the error rate remains unchanged. The causes of errors may be system-level factors and the changes inculcated by the information resource may address only some of these factors. Patients may be motivated to exercise through interaction with an information resource, but fail to meet weight loss objectives because they cannot afford concomitant changes in their diets. In other domains, an information resource may be widely used by researchers to access biomedical information, as determined by a user effect study, but a subsequent problem impact study may or may not reveal effects on scientific productivity. New educational technology may change the ways students learn, but may or may not increase their performance on standardized examinations. In the Lester study of SMS alerts (Lester et al. 2010), increased adherence to antiretroviral therapy (a user action) was also accompanied by improved viral load suppression. Fully comprehensive problem impact studies will also be sensitive to unintended consequences. Sometimes, the solution to the target problem creates other, unintended and unanticipated problems that can affect perceptions of success. As electronic mail became an almost universal mode of communication, almost no one anticipated the problems of "spam" or "phishing".

11.4.3 Factors Distinguishing the Nine Study Types

Table 11.2 further distinguishes the nine study types, as described above, using a set of key differentiating factors discussed in detail in the paragraphs that follow.

The setting in which the study takes place: Studies of the design process, the resource structure, and many resource functions are typically conducted outside the active practice or decision environment, in a "laboratory" setting. Studies to elucidate the need for a resource and studies of its impact on users would usually take place in ongoing practice settings—known generically as the "field"—where health care practitioners, researchers, students, or administrators are doing

Table 11.2 Factors distinguishing the nine generic study types

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 and quality of data collected or displayed; accuracy of advice given…
6. Field function study	Field	Prototype or released version	Proxy, real users	Real	Speed and 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

real work in the real world. The same is true for studies of the impact of a resource on persons and organizations. These studies can take place only in a setting where the resource is available for use at the time and where professional activities occur and/or important decisions are made. To an investigator planning studies, an important consideration that determines the kind of study possible is the degree of access to users in the field setting. If, as a practical matter, access to the field setting is very limited, then several study types listed in Tables 11.1 and 11.2 are not possible, and the range of evaluation questions that can be addressed is limited accordingly.

The version of the resource used: For some kinds of studies, a simulated or prototype version of the resource may be sufficient (Scott et al. 2011), whereas for studies in which the resource is employed by intended users to support real decisions and actions, a fully robust and reliable version is needed (e.g., Lester et al. 2010).

The sampled resource users: Most biomedical information resources are not autonomous agents that operate independently of users. More typically, information resources function through interaction with one or more such "users" who often bring to the interaction their own domain knowledge and knowledge of how to operate the resource. In some types of evaluation studies, the users of the resource are not the end users for whom the resource is ultimately designed, but are members of the development or evaluation teams, or other individuals we can call "proxy users" who are chosen because they are conveniently available or because they are affordable. In other types of studies, the users are sampled from the end-users for whom the resource is ultimately designed. The type of users employed gives shape to a study and can affect its results profoundly. The usability of a resource is easily overestimated if the "users" in a usability study are those who designed the system. Volunteer users of a consumer-oriented website

may be more literate than the general population the resource is designed to benefit.

The sampled tasks: For function and effect studies, the resource is actually "run." The users included in the study actually interact with the resource. This requires tasks, typically clinical or scientific case or problems, for the users to undertake. These tasks can be invented or simulated; they can be abstracted versions of real cases or problems, shortened to suit the specific purposes of the study; or they can be live cases or research problems as they present to resource users in the real world. Clearly, the kinds of tasks employed in a study have serious implications for the study results and the conclusions that can be drawn from them.

The observations that are made: All evaluation studies entail observations that generate data that are subsequently analyzed to make decisions. As seen in Table 11.2, many different kinds of observations[2] can be made.

In the paragraphs above we have introduced the term "sampling" for both tasks and users. It is important to establish that in real evaluation studies, tasks and users are always sampled from some real or hypothetical populations. Sampling of users and tasks are major challenges in evaluation study design since it is never possible, practical, or desirable to try to study everyone doing everything possible with an information resource. Sampling issues are addressed later in this chapter.

11.5 Conducting Investigations: Collecting and Drawing Conclusions from Data

11.5.1 Two Grand Approaches to Study Design, Data Collection, and Analysis

Several authors have developed classifications, or **typologies**, of evaluation methods or approaches. Among the best is that developed in 1980 by Ernest House (1980). A major advantage of House's typology is that each approach is linked elegantly to an underlying philosophical model, as detailed in his book. This classification divides current practice into eight discrete approaches, four of which may be viewed as **objectivist** and four of which may be viewed as **subjectivist**. While the distinctions between the eight approaches House describes are beyond the scope of this chapter, the grand distinction between objectivist and subjectivist approaches is very important. Note that these approaches are not entitled "objective" and "subjective", because those labels carry strong and fundamentally misleading connotations of scientific precision in the former case and of idiosyncratic imprecision in the latter. We will see in this section how both objectivist (often called quantitative) and subjectivist (often called qualitative) approaches find rigorous application across the range of study types described earlier.

To appreciate the fundamental difference between the approaches, it is necessary to address their very different philosophical roots. The objectivist approaches derive from a **logical–positivist** philosophical orientation—the same orientation that underlies the classic experimental sciences. The major premises underlying the objectivist approaches are as follows:

- In general, attributes of interest are properties of the resource under study. More specifically, this position suggests that the merit and worth of an information resource—the attributes of most interest in evaluation—can in principle be measured with all observations yielding the same result. It also assumes that an investigator can measure these attributes without affecting how the resource under study functions or is used.
- Rational persons can and should agree on what attributes of a resource are important to measure and what results of these measurements would be identified as a most desirable, correct, or positive outcome. In medical informatics, making this assertion is tantamount to stating that a gold standard of resource performance can always be identified and that all rational individuals can be brought to consensus on what this gold standard is.

[2] We use "observations" here very generically to span a range of activities that includes watching someone work with an information resource as well as highly-instrumented tracking or measurement.

- Because numerical measurement allows precise statistical analysis of performance over time or performance in comparison with some alternative, numerical measurement is prima facie superior to a verbal description. Verbal, descriptive data (generally known as qualitative data) are useful in only preliminary studies to identify hypotheses for subsequent, precise analysis using quantitative methods.
- Through these kinds of comparisons, it is possible to demonstrate to a reasonable degree that a resource is or is not superior to what it replaced or to a competing resource.

Contrast these assumptions with a set of assumptions that derives from an **intuitionist–pluralist** or de-constructivist philosophical position that spawns a set of subjectivist approaches to evaluation:

- What is observed about a resource depends in fundamental ways on the observer. Different observers of the same phenomenon might legitimately come to different conclusions. Both can be objective in their appraisals even if they do not agree; it is not necessary that one is right and the other wrong.
- Merit and worth must be explored in context. The value of a resource emerges through study of the resource as it functions in a particular patient care or educational environment.
- Individuals and groups can legitimately hold different perspectives on what constitutes the most desirable outcome of introducing a resource into an environment. There is no reason to expect them to agree, and it may be counterproductive to try to lead them to consensus. An important aspect of an evaluation would be to document the ways in which they disagree.
- Verbal description can be highly illuminating. Qualitative data are valuable, in and of themselves, and can lead to conclusions as convincing as those drawn from quantitative data. The value of qualitative data, therefore, goes far beyond that of identifying issues for later "precise" exploration using quantitative methods.
- Evaluation should be viewed as an exercise in argument or rhetoric, rather than as a demonstration, because any study can appear equivocal when subjected to serious scrutiny.

The approaches to evaluation that derive from this subjectivist philosophical perspective may seem strange, imprecise, and unscientific when considered for the first time. This perception stems in large part from the widespread acceptance of the objectivist worldview in biomedicine. Over the last two decades, however, thanks to some early high quality studies (e.g., Forsythe et al. 1992; Ash et al. 2003) the importance and utility of these subjectivist approaches in evaluation has been established within biomedical informatics. As stated earlier, the evaluation mindset includes methodological eclecticism. It is important for people trained in classic experimental methods at least to understand, and possibly even to embrace, the subjectivist worldview if they are to conduct fully informative evaluation studies.

Some of these issues are further considered in Appendix B, which describes some more recent perspectives on evaluation.

11.5.2 Conduct of Objectivist Studies

Figure 11.2 expands the generic process for conducting evaluation studies to illustrate the steps involved in conducting an objectivist study. Figure 11.2 illustrates the linear sequence in which the investigation portion of an evaluation study is typically carried out. We will focus in this chapter on issues of study design that are the biggest challenge to the validity of objectivist studies and we will further focus on that subset of objectivist study designs which are comparative in nature. More details on the other aspects of objectivist studies are available in a textbook on the subject (Friedman et al. 2005) and in standard references on experimental design (Campbell & Stanley 1963).

11.5.2.1 Structure and Terminology of Comparative Studies

Most objectivist evaluations performed in the world make a comparison of some type. For informatics, aspects of performance of individuals, groups, or organizations *with* the

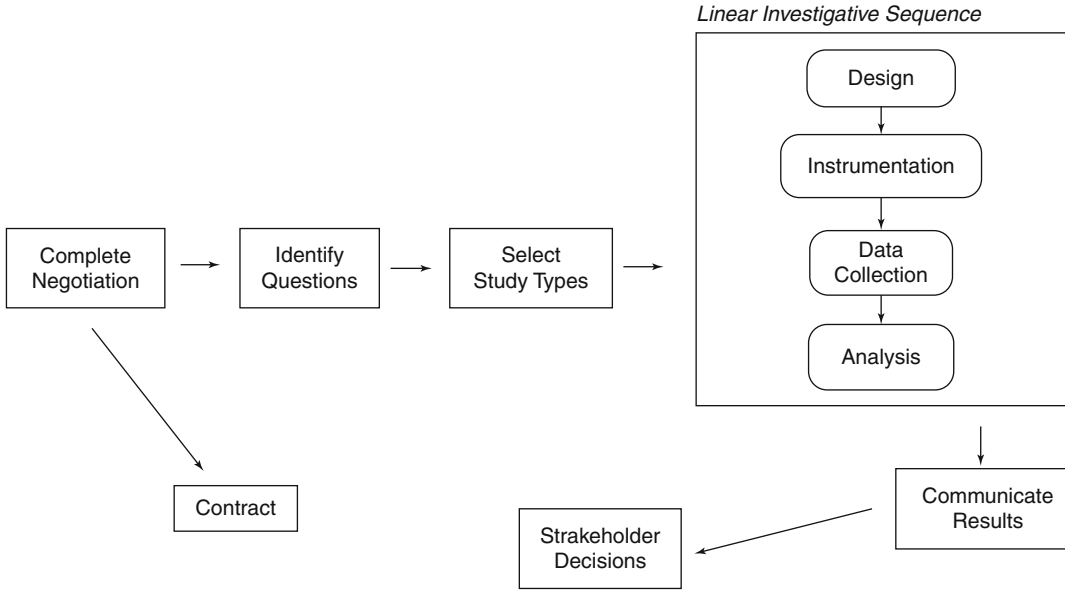

Fig. 11.2 Generic structure depicting an objectivist investigation

information resource are compared to those same aspects *without* the resource or with some alternative resource. After identifying a sample of participants for the study, the researcher assigns each participant, often randomly, to one or a set of conditions. Some outcomes of interest are measured for each participant. The averaged values of these outcomes are then compared across the conditions. If all other factors are controlled, then any measured difference in the averaged outcomes can be attributed to the resource.

This relatively simple description of a comparative study belies the many issues that affect their design, execution, and ultimate usefulness. To understand these issues, we must first develop a precise terminology.

The **participants** in a study are the entities about which data are collected. It is key to emphasize that participants are often people—for example, care providers or recipients—but also may be information resources, groups of people, or organizations. Because many of the activities in informatics are conducted in hierarchical settings with naturally occurring groups (a "doctor's patients"; the "researchers in a laboratory"), investigators must, for a particular study, define the participants carefully and consistently.

Variables are specific characteristics of the participants that either are measured purposefully by the investigator or are self-evident properties of the participants that do not require measurement. Some variables take on a continuous range of values. Others have a discrete set of levels corresponding to each of the possible measured values. For example, in a hospital setting, physician members of a ward team can be classified as residents, fellows, or attending physicians. In this case, the variable "physician's level of qualification" is said to have three "levels".

The **dependent variables** are those variables in the study that captures the outcomes of interest to the investigator. (For this reason, dependent variables are also called **outcome variables**.) A study may have one or more dependent variables. In a typical study, the dependent variables will be computed, for each participant, as an average over a number of tasks. For example, clinicians' diagnostic performance may be measured over a set of cases, or "tasks", that provide a range of diagnostic challenges.

The **independent variables** are included in a study to try and explain the measured values of the dependent variables. For example, whether an information resource is available, or not, to support certain clinical tasks could be the major independent variable in a study designed to evaluate the effects of that resource.

Measurement challenges almost always arise in the assessment of the outcome or dependent variable for a study (Friedman 2003). Often, for example, the dependent variable is some type of performance measure that invokes concerns about reliability (precision) and validity (accuracy) of measurement. The independent variables may also raise measurement challenges. When the independent variable is patient gender, for example, the measurement problems are relatively straightforward—though in some studies classifying trans-gender individuals may need some thought. If the independent variable is an attitude or other "state of mind", profound measurement challenges can arise.

11.5.2.2 Issues of Measurement

Measurement is the process of assigning a value corresponding to the presence, absence, or degree of a specific attribute in a specific object, as illustrated in Fig. 11.3. When we speak specifically of measurement, it is customary to use the term "object" to refer to the entity on which measurements are made. Measurement usually results in either (1) the assignment of a numerical score representing the extent to which the attribute of interest is present in the object, or (2) the assignment of an object to a specific category. Taking the temperature (attribute) of a patient (object) is an example of the process of measurement.

From the premises underlying objectivist studies (see Sect. 11.5.1), it follows that proper execution of such studies requires careful and specific attention to methods of measurement. It can never be assumed, particularly in informatics, that attributes of interest are measured without error. Accurate and precise measurement must not be an afterthought. Measurement is of particular importance in biomedical informatics because, as a relatively young field, informatics does not have a well-established tradition of "variables worth measuring" or proven instruments for measuring them (Friedman 2003). By and large, people planning studies in informatics are faced first with the task of deciding what to measure and then with that of developing their own measurement methods. For most researchers, these tasks prove to be harder and more time-consuming than initially anticipated.

We can underscore the importance of measurement by establishing a formal distinction between studies undertaken to develop methods

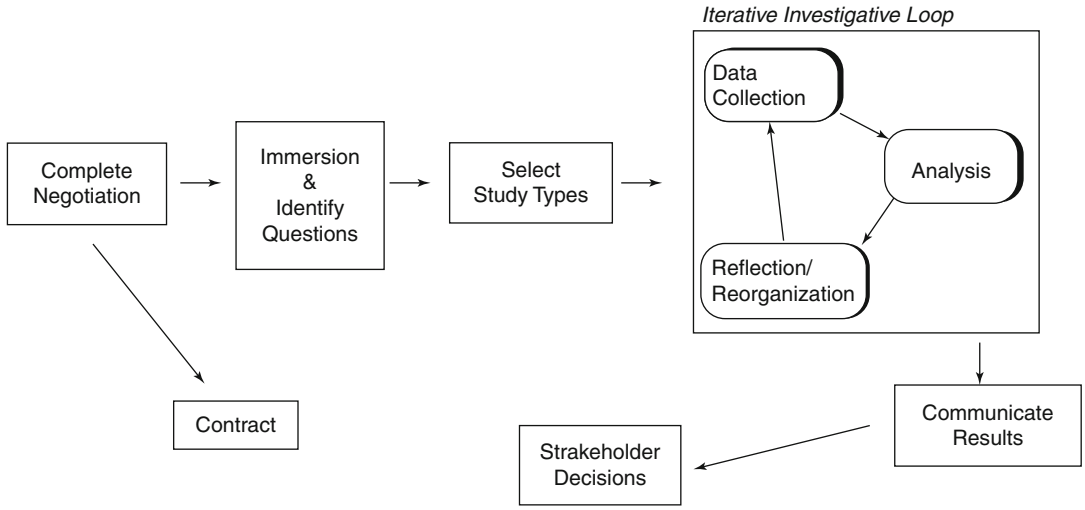

Fig. 11.3 Generic structure depicting a subjectivist investigation

for making measurements, which we call measurement studies, and the subsequent use of these methods to address questions of direct importance in informatics, which we call demonstration studies. **Measurement studies** seek to determine how accurately and precisely an attribute of interest can be measured in a population of objects. In an ideal objectivist measurement, all observers will agree on the result of the measurement. Any disagreement is therefore due to error, which should be minimized. The more agreement among observers or across observations, the better the measurement. Measurement procedures developed and validated through measurement studies provide researchers with what they need to conduct **demonstration studies** that directly address questions of substantive and practical concern to the stakeholders for an evaluation study. Once we know how accurately we can measure an attribute using a particular procedure, we can employ the measured values of this attribute as a variable in a demonstration study to draw inferences about the performance, perceptions, or effects of an information resource. For example, once measurement studies have determined how accurately and precisely the usability of a class of information resources can be measured, a subsequent demonstration study could explore which of two resources that are members of this class has greater usability.

A detailed discussion of measurement issues is beyond the scope of this chapter. The bottom line is that investigators should know that their measurement methods will be adequate before they collect data for their studies. It is necessary to perform a measurement study, involving data collection on a small scale, to establish the adequacy of all measurement procedures if the measures to be used do not have an established track record (e.g., Ramnarayan et al. 2003; Demiris et al. 2000). Even if the measurement procedures of interest do have a track record in a particular health care environment and with a specific mix of cases and care providers, they may not perform equally well in a different environment, so measurement studies may still be necessary. Researchers should always ask themselves, "How good are my measures in this particular setting?" whenever they are planning a study, before they proceed to the demonstration phase. The importance of measurement studies for informatics was explained in 1990 by Michaelis and co-workers (Michaelis et al. 1990). Another study (Friedman et al. 2003) has demonstrated that studies of clinical information systems have not systematically addressed the adequacy of the methods used to measure the specific outcomes reported in these studies.

Whenever possible, investigators planning studies should employ established measurement methods with a "track record", rather than developing their own. While there exist relatively few compendia and measurement instruments specifically for health informatics, a web-based resource listing over 50 instruments associated with the development, usability, and impact of management information systems is available on the Internet.[3]

11.5.2.3 Sampling Strategies

Selection of Participants

The participants selected for objectivist studies must resemble those to which the evaluator and others responsible for the study wish to apply the results. For example, when attempting to quantify the likely impact of a clinical information resource on clinicians at large, there is no point in studying its effects on the clinicians who helped develop it, especially if they built it, as they are likely to be more familiar with the resource than average practitioners. Characteristics of clinical participants that typically need to be taken into account include age, experience, role, attitude toward digital information resources, and extent of their involvement in the development of the resource. Analogous factors would apply to patients or health care consumers as participants.

Volunteer Effect

A common bias in the selection of participants is the use of volunteers. It has been established in many areas that people who volunteer as participants, whether to complete questionnaires, partici-

[3] http://www.misq.org/skin/frontend/default/misq/surveys98/surveys.html#toc (Accessed January 27, 2013).

pate in psychology experiments, or test-drive new cars or other technologies, are atypical of the population at large (e.g., Pinsky et al. 2007). Although volunteers may make willing participants for pilot studies, they should be avoided in definitive demonstration studies, as they considerably reduce the generality of findings. One strategy is to include all participants meeting the selection criteria in the study. However, if this would result in too many participants, rather than asking for volunteers, it is better randomly or otherwise systematically to select a representative sample of all eligible clinicians, following up invitation letters with telephone calls to achieve as near 100 % recruitment of the selected sample as possible.

Number of Participants Needed

The financial investment required for an evaluation study depends critically on the number of participants needed. The required number in turn depends on the precision of the answer required from the study and the risk investigators are willing to take of failing to detect a significant effect. (All other things being equal, the larger the sample size, the greater the likelihood of detecting an effect against a predetermined criterion for statistical significance.) Statisticians can advise on this point and carry out sample-size calculations to estimate the number of participants required. Sometimes, in order to recruit the required number of participants, an element of volunteer effect must be tolerated; often there is a trade-off between obtaining a sufficiently large sample and ensuring that the sample is representative. Also, the impact of sample size on effect detection is non-linear. The value of adding, say, 10 more representative participants to a sample of 100 is far less than that of adding 10 more participants to a sample of 30.

Selection of Tasks

In the same way that participants must be carefully selected to resemble the people likely to use the information resource, any tasks the participants complete must also resemble those that will generally be encountered in the field setting where the information resource is deployed. Thus when evaluating a clinical order-entry system intended for general use, it would be unwise to use only complex cases from, for example, a pediatric intensive care setting. Although the order-entry system might well be of considerable benefit in intensive care cases, it is inappropriate to generalize results from such a limited sample to the full range of cases seen in ambulatory pediatrics. An instructive example is provided by the study of Van Way et al. (1982) who developed a scoring system for diagnosing appendicitis and studied the resource's accuracy using exclusively patients who had undergone surgery for suspected appendicitis. Studying this group of patients had the benefit of allowing the true cause of the abdominal pain to be obtained with near certainty as a by-product of the surgery itself. However, in these patients who had all undergone surgery for suspected appendicitis the symptoms were more severe and the incidence of appendicitis was five to ten times higher than for the typical patient for whom such a scoring system would be used. Thus the accuracy obtained with postsurgical patients would be a poor estimate of the system's accuracy in routine clinical use.

If the performance of an information resource is measured on a small number of hand-picked cases, the functions it performs may appear spuriously complete and its usability overestimated. This is especially likely if these cases are similar to, or even identical with, the training set of cases used to develop or tune the information resource before the evaluation is carried out. When a statistical model that powers an information resource is carefully adjusted to achieve maximal performance on training data, this adjustment may worsen its accuracy on a fresh set of data due to a phenomenon called over fitting (Wasson 1985). Thus it is important to obtain a new set of cases and evaluate performance on this new test set. Sometimes developers omit cases from a sample if they do not fall within the scope of the information resource, for example if the final diagnosis for a case is not represented in a diagnostic system's knowledge base. This practice violates the principle that a test set should be representative of all cases in which the information resource will be used, and will overestimates its effectiveness with unseen data.

11.5.2.4 Control Strategies in Comparative Studies

One of the most challenging questions in comparative study design is how to obtain control (Liu et al. 2011). We need a way to account for all the other changes taking place that are not attributable to the information resource. In the following sections we review a series of control strategies. We employ, as a running example of an information resource under study, a reminder system that prompts doctors to order prophylactic antibiotics for orthopedic patients to prevent postoperative infections. In this example, the intervention is the installation and commissioning of the reminder system; the participants are the physicians; and the tasks are the patients cared for by the physicians. The dependent variables derive from the outcome measurements made and would include physicians' ordering of antibiotics and the rate of postoperative infections averaged across the patients cared for by each physician.

Descriptive (Uncontrolled) Studies

In the simplest possible design, an uncontrolled or **descriptive study**, we install the reminder system, allow a suitable period for training, and then make our measurements. There is no independent variable. Suppose that we discover that the overall postoperative infection rate is 5% and that physicians order prophylactic antibiotics in 60 % of orthopedic cases. Although we have two measured dependent variables, it is hard to draw meaningful conclusions from these figures. It is possible that there has been no change due to the system.

Historically Controlled Experiments

As a first improvement to a descriptive study, let us consider a **historically controlled experiment**, sometimes called a **before–after study**. The investigator makes baseline measurements of antibiotic ordering and postoperative infection rates before the information resource is installed, and then makes the same measurements after the information resource is in routine use. The independent variable is time and has two levels: before and after resource installation. Let us say that, at baseline, the postoperative infection rates were 10 % and doctors ordered prophylactic antibiotics in only 40 % of cases; the post-intervention figures are the same as before (see Table 11.3).

Table 11.3 Results from a hypothetical before-after study of the impact of reminders on post operative infection rates

	Reminder group
Baseline infection rate	10 %
Post-intervention infection rate	5 %

The investigators may claim that the halving of the infection rate can be safely ascribed to the information resource, especially because it was accompanied by a 20 % improvement in doctors' antibiotic prescribing. Many other factors might, however, have changed in the interim to cause these results, especially if there was a long interval between the baseline and postintervention measurements. New staff could have taken over, the case mix of patients could have changed, new prophylactic antibiotics may have been introduced, or clinical audit meetings may have highlighted the infection problem and thus caused greater clinical awareness. Simply assuming that the reminder system alone caused the reduction in infection rates is naive. Other factors, known or unknown, could have changed meanwhile, making untenable the simple assumption that our intervention is responsible for all of the observed effects (Liu et al. 2011). An improvement on this design is to add either internal or external controls—preferably both. The internal control should be a measure likely to be affected by any nonspecific changes happening in the local environment, but unaffected by the intervention. The external control can be exactly the same measure as in the target environment, but in a similar external setting, e.g., another hospital. If the measure of interest changes while there is no change in either internal or external controls, a skeptic needs to be quite resourceful to claim that the system is not responsible (Wyatt and Wyatt 2003).

Simultaneous Nonrandomized Controls

To address some of the problems with historical controls, we might use **simultaneous controls**, which requires us to make our outcome measure-

Table 11.4 Results of a hypothetical non-randomized parallel group study of reminders and post op infection rates

	Reminder group (%)	Control group (%)
Baseline rate	10	10
Post-intervention rate	5	11

ments in doctors and patients who are not influenced by the prophylactic antibiotic reminder system but who are subject to the other changes taking place in the environment. Taking measurements both before and during the intervention strengthens the design, because it gives an estimate of the changes due to the nonspecific factors taking place during the study period.

This study design would be a parallel group comparative study with simultaneous controls. Table 11.4 gives hypothetical results of such a study, focusing on postoperative infection rates as a single outcome measure or dependent variable. The independent variables are time and group, both of which have two levels of intervention and control.

There is the same improvement in the group where reminders were available, but no improvement—indeed a slight deterioration—where no reminders were available. This design provides suggestive evidence of an improvement that is most likely to be due to the reminder system. This inference is stronger if the same doctors worked in the same wards during the period the system was introduced, and if similar kinds of patients, subject to the same nonspecific influences, were being operated on during the whole time period.

Even though the controls in this example are simultaneous, skeptics may still refute our argument by claiming that there is some systematic, unknown difference between the clinicians or patients in the two groups. For example, if the two groups comprised the patients and clinicians in two adjacent wards, the difference in the infection rates could be attributable to systematic or chance differences between the wards. Perhaps hospital-staffing levels improved in some wards but not in others, or there was cross infection by a multiple-resistant organism only among the patients in the control ward. To overcome such criticisms, we could expand the study to include all wards in the hospital—or even other hospitals—but that would clearly take considerable resources. We could try to measure everything that happens to every patient in both wards and to build complete psychological profiles of all staff to rule out systematic differences. We would still, however, be vulnerable to the accusation that some variable that we did not measure—did not even know about—explains the difference between the two wards. A much simpler strategy is to ensure that the controls really are comparable by randomizing them.

Simultaneous Randomized Controls

The crucial problem in the previous example is that, although the controls were simultaneous, there may have been systematic, unmeasured differences between them and the participants receiving the intervention (Liu and Wyatt 2011). A simple and effective way of removing systematic differences, whether due to known or unknown factors, is to randomize the assignment of participants to control or intervention groups. Thus, we could randomly allocate one-half of the doctors on both wards to receive the antibiotic reminders and the remaining doctors to work as they did before. We would then measure and compare postoperative infection rates in patients managed by doctors in the reminder and control groups. Provided that the doctors never look after one another's patients, any difference that is statistically "significant" (conventionally, a result that is statistically determined to have a probability of 0.05 or less of occurring by chance) can be attributed reliably to the reminders.

Table 11.5 shows the hypothetical results of such a study. The baseline infection rates in the patients managed by the two groups of doctors are similar, as we would expect, because the patients were allocated to the groups by chance. There is a greater reduction in infection rates in patients of reminder physicians compared with those of control physicians. Because random assignment means that there was no systematic difference in patient characteristics between groups, the only systematic difference between

Table 11.5 Results of a hypothetical randomized controlled trial of the impact of reminders on post op infection rates

	Reminder physicians (%)	Control physicians (%)
Baseline infection rate	11	10
Post intervention infection rate	6	8
Difference in infection rate	−5	−2

the two groups of patients is receipt of reminders by their doctors.

Provided that the sample size is large enough for these results to be statistically significant, we might begin to conclude with some confidence that providing doctors with reminders caused the reduction in infection rates. One lingering question is why there was also a small reduction, from baseline to installation, in infection rates in control cases, even though the control group should have received no reminders.

11.5.3 Conduct of Subjectivist Studies

The objectivist approaches to evaluation, described in the previous section, are useful for addressing some, but not all, of the interesting and important questions that challenge investigators in medical informatics. The subjectivist approaches described here address the problem of evaluation from a different set of premises. They use different but equally rigorous methods. Figure 11.3 expands the generic process for conducting evaluation studies to illustrate the stages involved in conducting a subjectivist study, and emphasizes the "iterative loop" of data collection, analysis and reflection as the major distinguishing characteristic of a subjectivist investigation. Another distinctive feature of subjectivist studies is an immersion in the environment where the resources is being or will be deployed. Because subjectivist approaches may be less familiar to readers of this chapter, we describe subjectivist studies in more detail than we did their objectivist counterparts.

11.5.3.1 The Rationale for Subjectivist Studies

Subjectivist methods enable us to address the deeper questions that arise in informatics: the detailed "whys" and "according to whoms" in addition to the aggregate "whethers" and "whats." Subjectivist approaches seek to represent the viewpoints of people who are users of the resource or are otherwise significant participants in the environment where the resource operates. The goal is illumination rather than judgment. The investigators seek to build an argument that promotes deeper understanding of the information resource or environment of which it is a part. The methods used derive largely from **ethnography** (Forsythe 1992). The investigators immerse themselves physically in the environment where the information resource is or will be operational, and they collect data primarily through observations, interviews, and reviews of documents. The designs—the data-collection plans—of these studies are not rigidly predetermined and do not unfold in a fixed sequence. They develop dynamically and nonlinearly as the investigators' experience accumulates.

11.5.3.2 A Rigorous, but Different, Methodology

The subjectivist approaches to evaluation, like their objectivist counterparts, are empirical methods. Although it is easy to focus only on their differences, these two broad classes of evaluation approaches share many features. In all empirical studies, for example, evidence is collected with great care; the investigators are always aware of what they are doing and why. The evidence is then compiled, interpreted, and ultimately reported. Investigators keep records of their procedures, and these records are open to audit by the investigators themselves or by individuals outside the study team. The principal investigator or evaluation-team leader is under an almost sacred scientific obligation to report their methods. Failure to do so will invalidate a study. Both classes of approaches also share a dependence on theories that guide investigators to explanations of the observed phenomena, as well as to a dependence on the pertinent empirical literature

such as published studies that address similar phenomena or similar settings. In both approaches, there are rules of good practice that are generally accepted; it is therefore possible to distinguish a good study from a bad one.

There are, however, fundamental differences between objectivist and subjectivist approaches. First, subjectivist studies are **emergent** in design. Objectivist studies typically begin with a set of hypotheses or specific questions, and with a plan for addressing each member of this set. The investigator assumes that, barring major unforeseen developments, the plan will be followed exactly. Deviation, in fact, might introduce bias. The investigator who sees negative results emerging from the exploration of a particular question or use of a particular measurement instrument might change strategies in hope of obtaining more positive findings. In contrast, subjectivist studies typically begin with general **orienting issues** that stimulate the early stages of investigation. Through these initial investigations, the important questions for further study emerge. The subjectivist investigator is willing, at virtually any point, to adjust future aspects of the study in light of the most recent information obtained. Subjectivist investigators tend to be **incrementalists**; they change their plans from day-to-day and have a high tolerance for ambiguity and uncertainty. In this respect, they are much like good software developers. Also like software developers, subjectivist investigators must develop the ability to recognize when a project is finished, when further benefit can be obtained only at too great a cost in time, money, or work.

A second feature of subjectivist studies is a **naturalistic** orientation, a reluctance to manipulate the setting of the study, which in most cases is the environment in to which the information resource is introduced. They do not alter the environment to study it. Control groups, placebos, purposeful altering of information resources to create contrasting interventions, and other techniques that are central to the construction of objectivist studies typically are not used. Subjectivist studies will, however, employ quantitative data for descriptive purposes and may offer quantitative comparisons when the research setting offers a "natural experiment" where such comparisons can be made without deliberate intervention. For example, when physicians and nurses both use a clinical system to enter orders, their differing experiences with the system offer a natural basis for comparison (Ash 2003). Subjectivist researchers are opportunists where pertinent information is concerned; they will use what they see as the best information available to illuminate a question under investigation.

A third important distinguishing feature of subjectivist studies is that their end product is a report written in narrative prose. These reports may be lengthy and may require significant time investment from the reader; no technical understanding of quantitative research methodology or statistics is required to comprehend them. Results of subjectivist studies are therefore accessible—and may even be entertaining—to a broad community in a way that results of objectivist studies are not. Objectivist study reports often can be results of inferential statistical analyses that most readers will not find easy to read and will typically not understand. Reports of subjectivist studies seek to engage their audience.

11.5.3.3 Natural History of a Subjectivist Study

Figure 11.3 illustrates the stages that characterize a subjectivist study (see also Chap. 9 in Friedman et al. 2005). These stages constitute a general sequence, but, as we mentioned, subjectivist investigators must always be prepared to revise their thinking and possibly to return to earlier stages in light of new evidence. Backtracking is a legitimate step in this model.

1. *Negotiation of the ground rules of the study*: In any empirical research, and particularly in evaluation studies, it is important to negotiate an understanding between the study team and the people commissioning the study. This understanding should embrace the general aims of the study; the kinds of methods to be used; the access to various sources of information, including health care providers, patients, and various documents; and the format for interim and final reports. The aims of the study may be formulated in a set of initial

orienting questions. Ideally, this understanding will be expressed in a **memorandum of understanding**, analogous to a contract.

2. *Immersion into the environment*: At this stage, the investigators begin spending time in the work environment. Their activities range from formal introductions to informal conversations, or to silent presence at meetings and other events. Investigators use the generic term **field** to refer to the setting, which may be multiple physical locations, where the work under study is carried out. Trust and openness between the investigators and the people in the field are essential elements of subjectivist studies to ensure full and candid exchange of information.

   Even as immersion is taking place, the investigator is already collecting data to sharpen the initial questions or issues guiding the study. Early discussions with people in the field, and other activities primarily targeted toward immersion, inevitably begin to shape the investigators' views. Almost from the outset, the investigator is typically addressing several aspects of the study simultaneously.
3. *Iterative loop*: At this point, the procedural structure of the study becomes akin to an iterative loop, as the investigator engages in cycles of data collection, analysis and reflection, member checking, and reorganization. Data collection involves interview, observation, document analysis, and other methods. Data are collected on planned occasions, as well as serendipitously or spontaneously. The data are recorded carefully and are interpreted in the context of what is already known. Analysis and reflection entail the contemplation of the new findings during each cycle of the loop. **Member checking** is the sharing of the investigator's emerging thoughts and beliefs with the participants themselves. Reorganization results in a revised agenda for data collection in the next cycle of the loop.

   Although each cycle within the iterative loop is depicted as linear, this representation is misleading. Net progress through the loop is clockwise, as shown in Fig. 11.3, but backward steps are natural and inevitable. They are not reflective of mistakes or errors. An investigator may, after conducting a series of interviews and studying what participants have said, decide to speak again with one or two participants to clarify their positions on a particular issue.
4. *Communicate results*: Subjectivist students tend to have a multi-staged reporting and communication process. The first draft of the study report should itself be viewed as a research instrument. By sharing this report with a variety of individuals, the investigator obtains a major check on the validity of the findings. Typically, reactions to the preliminary report will generate useful clarifications and a general sharpening of the study findings. Because the report usually includes a prose narrative, it is vitally important that it be well written in language understandable by all intended audiences. Circulation of the report in draft can ensure that the final document communicates as intended. Use of anonymous quotations from interviews and documents makes a report highly vivid and meaningful to readers.

   The final report, once completed, should be distributed as negotiated in the original memorandum of understanding. Distribution is often accompanied by "meet the investigator" sessions that allow interested persons to ask the author of the report to expand or explain what has been written.

11.5.3.4 Subjectivist Data-Collection and Data-Analysis Methods

What data-collection strategies are in the subjectivist researcher's black bag? There are several, and they are typically used in combination. We shall discuss each one, assuming a typical setting for a subjectivist study in medical informatics, the introduction of an information resource into patient care activities in a hospital.

Observation

The investigators typically immerse themselves into the setting under study in one of two ways. The investigator may act purely as a detached observer, becoming a trusted and unobtrusive

feature of the environment but not a participant in the day-to-day work and thus reliant on multiple "informants" as sources of information. True to the naturalistic feature of this kind of study, great care is taken to diminish the possibility that the presence of the observer will skew the work activities that occur or that the observer will be rejected outright by the team. An alternative approach is participant observation, where the investigator becomes a member of the work team. Participant observation is more difficult to engineer; it may require the investigator to have specialized training in the study domain. It is time consuming but can give the investigator a more vivid impression of life in the work environment. During both kinds of observation, data accrue continuously. These data are qualitative and may be of several varieties: statements by health care providers and patients, gestures and other nonverbal expressions of these same individuals, and characteristics of the physical setting that seem to affect the delivery of health care.

Interviews

Subjectivist studies rely heavily on interviews. Formal interviews are occasions where both the investigator and interviewee are aware that the answers to questions are being recorded (on paper or tape) for direct contribution to the evaluation study. Formal interviews vary in their degree of structure. At one extreme is the **unstructured interview**, where there are no predetermined questions. Between the extremes is the **semi structured interview**, where the investigator specifies in advance a set of topics that he would like to address but is flexible as to the order in which these topics are addressed, and is open to discussion of topics not on the prespecified list. At the other extreme is the **structured interview**, with a schedule of questions that are always presented in the same words and in the same order. In general, the unstructured and semi structured interviews are preferred in subjectivist research. Informal interviews—spontaneous discussions between the investigators and members of a team that occur during routine observation—are also part of the data collection process. Informal interviews are invariably considered a source of important data. Group interviews, akin to focus groups, may also be employed (e.g., Haddow et al. 2011). Group interviews are very efficient ways to reach large numbers of participants, but investigators should not assume that individual participants will express in a group setting the same sentiments they will express if interviewed one-on-one.

Sampling also enters into the interview process. There are usually more participants to interview than resources. Unlike in objectivist studies, where random sampling is a form of gold standard to inform statistical inference, subjectivist studies employ more purposeful strategies. Investigators might actively seek interviewees they suspect to have unique or particularly insightful opinions. They might remain in more frequent contact with key informants who, for various reasons, have the most insight into what is happening.

Document and Artifact Analysis

Every project produces a trail of papers and other artifacts. These include patient charts, the various versions of a computer program and its documentation, memoranda prepared by the project team, perhaps a cartoon hung on the office door by a ward clerk. Unlike the day-to-day events of patient care, these artifacts do not change once created or introduced. They can be examined retrospectively and referred to repeatedly, as necessary, over the course of a study. Also included under this heading are **unobtrusive measures**, which are the records accrued as part of the routine use of the information resource. They include, for example, user trace files of an information resource. Data from these measures are often quantifiable.

Anything Else That Seems Useful

Subjectivist investigators are supreme opportunists. As questions of importance to a study emerge, the investigators will collect any information that they perceive as bearing on these questions. This data collection could include clinical chart reviews, questionnaires, tests, simulated patients, and other methods more commonly associated with the objectivist approaches.

When to end data collection is another challenge in otherwise open-ended subjectivist studies. "Saturation" is important principle to help investigators know when to stop. Stated simply, a data collection process is saturated when it becomes evident that, as more data are collected, no new findings or insights are emerging.

Analysis of Subjectivist Data

There are many alternative procedures for analysis of qualitative data. The important point is that the analysis is conducted systematically. In general terms, the investigator looks for insights, themes or trends emerging from several different sources. She collates individual statements and observations by theme, as well as by source. Some investigators transfer these observations to file cards so they can be sorted and resorted in a variety of ways. Others use software especially designed to facilitate analysis of qualitative data (Fielding and Lee 1991). Because they allow electronic recording of the data while the investigator is "in the field", tablets, smartphone Apps and other hand-held devices are changing the way subjectivist research is carried out.

The subjectivist analysis process is fluid, with analytic goals shifting as the study matures. At an early stage, the goal is primarily to focus the questions that themselves will be the targets of further data elicitation. At the later stages of study, the primary goal is to collate data that address these questions. Conclusions derive credibility from a process of "triangulation", which is the degree to which information from different independent sources generate the same theme or point to the same conclusion. Subjectivist analysis also employs a strategy known as "member checking" whereby investigators take preliminary conclusions back to the persons in the setting under study, asking if these conclusions make sense, and if not, why not. In subjectivist investigation, unlike objectivist studies, the agenda is never completely closed. The investigator is constantly on the alert for new information that can require a significant reorganization of the findings and conclusions that have been drawn to date.

11.6 Communicating Evaluation Results

Once a study is complete, the results need to be communicated to the stakeholders and others who might be interested. In many ways, communication of evaluation results, a term we prefer over "reporting", is the most challenging aspect of evaluation. Elementary theory tells us that, in general, successful communication requires a sender, one or more recipients, and a channel linking them, along with a message that travels along this channel (Ong and Coiera 2011).

Seen from this perspective, successful communication of evaluation results is challenging in several respects. It requires that the recipient of the message actually receive it. That is, for evaluations, the recipient must read the written report or attend the meeting intended to convey evaluation results, and the investigator is challenged to create a report the stakeholders will want to read or to choreograph a meeting they will be motivated to attend. Successful communication also requires that the recipient understand the message, which challenges investigators to draft written documents at the right reading level, with audience-appropriate technical detail. Sometimes there must be several different forms of the written report to match several different audiences. Overall, we encourage investigators to recognize that their obligation to communicate does not end with the submission of a written document comprising their technical evaluation report. The report is one means or channel for communication, not an end in itself.

Depending on the nature, number, and location of the recipients, there are a large number of options for communicating the results of a study, including:

- Written reports
  - Document(s) prepared for specific audience(s)
  - Internal newsletter article
  - Published journal article, with appropriate permissions
  - Monograph, picture album, or book

- One-to-one or small group meetings
  - With stakeholders or specific stakeholder groups
  - With general public, if appropriate
- Formal oral presentations
  - To groups of project stakeholders
  - Conference presentation with poster or published paper in proceedings
  - To external meetings or seminars
- Internet
  - Project Web site or blog
  - Web "chat", forum or Twitter feed to socialize results
  - Online preprint
  - Internet based journal
- Other
  - Video or podcast describing study and information resource
  - Interview with journalist on newspaper, TV, radio

A written, textual report is not the sole medium for communicating evaluation results. Verbal, graphical, or multimedia approaches can be helpful as ways to enhance communication with specific audiences. Another useful strategy is to hold a "town meeting" to discuss a traditional written report after it has been released. Photographs or videos can portray the work setting for a study, the people in the setting, and the people using the resource. If appropriate permissions are obtained, these images—whether included as part of a written report, shown at a town meeting, or placed on a Web site—can be worth many thousands of words. The same may be true for recorded statements of resource users. If made available, with permission, as part of a multimedia report, the voices of the participants can convey a feeling behind the words that can enhance the credibility of the investigator's conclusions (Fig. 11.4).

Fig. 11.4 A picture is worth 1000 words: in the report of a study to establish the need for an electronic patient record, a casual photograph like this may prove much more persuasive than a table of data or paragraphs of prose

In addition to the varying formats for communication described above, investigators have other decisions to make after the data collection and analysis phases of a study are complete. One key decision is what personal role they will adopt after the formal investigative aspects of the work are complete. They may elect only to communicate the results, but they may also choose to persuade stakeholders to take specific actions in response to the study results, and perhaps even assist in the implementation of these actions. This raises a key question: Is the role of an evaluator simply to record and communicate study findings and then to move on to the next study, or is it to engage with the study stakeholders and help them change how they work as a result of the study?

To answer this question about the role of an evaluator, we need to understand that an evaluation study, particularly a successful one, has the potential to trigger a series of events, starting with the analysis of study results through communication to interpretation, recommendation, and even implementation. Some evaluators—perhaps enthused by the clarity of their results and an opportunity to use them to improve health

care, biomedical research, or education—prefer to go beyond reporting the results and conclusions to making recommendations, and then helping the stakeholders to implement them. The dilemma often faced by evaluators is whether to retain their scientific detachment and merely report the study results, or to stay engaged somewhat longer. Evaluators who choose to remain may become engaged in helping the stakeholders interpret what the results mean, guiding them in reaching decisions and perhaps even in implementing the actions decided upon. The longer they stay, the greater the extent to which evaluators must leave behind their scientific detachment and take on a role more commonly associated with change agents. Some confounding of these roles is inevitable when the evaluation is performed by individuals within the organization that developed the information resource under study. There is no hard-and-fast rule for deciding on the most appropriate role for the evaluator; the most important realization for investigators is that the different options exist and that a decision among them must inevitably be made.

11.7 Conclusion: Evaluation as an Ethical and Scientific Imperative

Evaluation takes place, either formally or informally, throughout the resource development cycle: from defining the need to monitoring the continuing impact of a resource once it is deployed (Stead et al. 1994). We have seen in this chapter that different issues are explored, at different degrees of intensity, at each stage of resource development. For meaningful evaluation to occur, adequate resources must be allocated for studies when time and money are budgeted for a development effort. Evaluation cannot be left to the end of a project. While formal evaluations, as we have described them here, are still seen as optional for resources of the types that are the foci of biomedical and health informatics, the increasing complexity and prevalence of these resources have raised concerns about their safety and effectiveness when used in the real world (e.g., Koppel et al. 2005). For the moment, we would argue that formal evaluations, using the range of methods described in this chapter, are mandated by the professional ethics of biomedical informatics as an applied scientific discipline (see Chap. 10).

Formal evaluations of biomedical information resources may someday be a statutory or regulatory requirement in many or all parts of the world, as they are already for new drugs or medical devices. If and when that day comes, the wide variety of questions to be addressed and the diversity of legitimate methods available to address those questions, as described in this chapter, will make it difficult to describe with exactitude how these studies should be done. There have been some published academic checklists or guidelines describing things to study and report in such studies (Talmon et al. 2009), but this is a bridge to be crossed in the future. We express the hope that writers of such guidelines and regulations will not overprescribe the methods to be used, while insisting on rigor in drawing conclusions from data collected using study designs thoughtfully matched to carefully identified questions. We hope the reader has learned from this chapter that rigor in evaluation is achievable in many ways, that information resources differ from drugs in many ways, and that overly rigid prescription of evaluation methods for informatics, however well intentioned, could defeat the well-intentioned purpose. However, it is also clear that the intensity of the evaluation effort should be closely matched to the resource's maturity (Stead et al. 1994). For example, one would not wish to conduct an expensive field trial of an information resource that is barely complete, is still in prototype form, may evolve considerably before taking its final shape, or is so early in its development that it may fail because simple programming bugs have not been eliminated. Seen from this perspective, biomedical information resources are merely a subset of complex intervention, and their development and evaluation needs to follow a logical pathway, such as

the MRC Framework for Complex Interventions (Campbell et al. 2000).

Suggested Readings

Anderson, J. G., & Aydin, C. E. (Eds.). (2005). *Evaluating the organizational impact of health care Information systems*. New York: Springer. This is an excellent edited volume that covers a wide range of methodological and substantive approaches to evaluation in informatics.

Brender, J. (2006). *Handbook for evaluation for health informatics*. Burlington: Elsevier Academic Press. Along with the Friedman and Wyatt text cited below, one of few textbooks available that focuses on evaluation in health informatics.

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. Cambridge, MA: MIT Press. This is a nicely written, detailed book that is focused on evaluation of artificial intelligence applications, not necessarily those operating in medical domains. It emphasizes objectivist methods and could serve as a basic statistics course for computer science students.

Fink, A. (2004). *Evaluation fundamentals: Insights into the outcomes, effectiveness, and quality of health programs* (2nd ed.). Thousand Oaks: Sage Publications. A popular text that discusses evaluation in the general domain of health.

Friedman, C. P., & Wyatt, J. C. (2006). *Evaluation methods in biomedical informatics*. New York: Springer. This is the book on which the current chapter is based. It offers expanded discussion of almost all issues and concepts raised in the current chapter.

Jain, R. (1991). *The art of computer systems performance analysis: Techniques for experimental design, measurement, simulation, and modelling*. New York: Wiley. This work offers a technical discussion of a range of objectivist methods used to study computer systems. The scope is broader than Cohen's book (1995) described earlier. It contains many case studies and examples and assumes knowledge of basic statistics.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks: Sage Publications. This is a classic book on subjectivist methods. The work is very rigorous but also very easy to read. Because it does not focus on medical domains or information systems, readers must make their own extrapolations.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks: Sage Publications. This is a valuable textbook on evaluation, emphasizing objectivist methods, and is very well written. It is generic in scope, and the reader must relate the content to biomedical informatics. There are several excellent chapters addressing pragmatic issues of evaluation. These nicely complement the chapters on statistics and formal study designs.

Questions for Discussion

1. Associate each of the following hypothetical evaluation scenarios with one or more of the nine types of studies listed in Table 11.1. Note that some scenarios may include more than one type of study.
   (a) An order communication system is implemented in a small hospital. Changes in laboratory workload are assessed.
   (b) The developers of the order communication system recruit five potential users to help them assess how readily each of the main functions can be accessed from the opening screen and how long it takes users to complete them.
   (c) A study team performs a thorough analysis of the information required by psychiatrists to whom patients are referred by a community social worker.
   (d) A biomedical informatics expert is asked for her opinion about a PhD project on a new bioinformatics algorithm. She requests copies of the student's code and documentation for review.
   (e) A new intensive care unit system is implemented alongside manual paper charting for a month. At the end of this time, the quality of the computer-derived data and data recorded on the paper charts is compared. A panel of intensive care experts is asked to identify, independently, episodes of hypotension from each data set.
   (f) A biomedical informatics professor is invited to join the steering group for a series of apps to support people living with diabetes. The only documentation available to critique at the first meeting is a statement of the project goal, description of the

planned development method, and the advertisements and job descriptions for team members.

(g) Developers invite educationalists to test a prototype of a computer-aided learning system as part of a user-centered design workshop

(h) A program is devised that generates a predicted 24-h blood glucose profile using seven clinical parameters. Another program uses this profile and other patient data to advise on insulin dosages. Diabetologists are asked to prescribe insulin for a series of "paper patients" given the 24-h profile alone, and then again after seeing the computer-generated advice. They are also asked their opinion of the advice.

(i) A program to generate alerts to prevent drug interactions is installed in a geriatric clinic that already has a computer-based medical record system. Rates of clinically significant drug interactions are compared before and after installation of the alerting program.

2. Choose any alternative area of biomedicine (e.g., drug trials) as a point of comparison, and list at least four factors that make studies in medical informatics more difficult to conduct successfully than in that area. Given these difficulties, discuss whether it is worthwhile to conduct empirical studies in medical informatics or whether we should use intuition or the marketplace as the primary indicators of the value of an information resource.
3. Assume that you run a philanthropic organization that supports biomedical informatics. In investing the scarce resources of your organization, you have to choose between funding a new system or resource development, or funding empirical studies of resources already developed. What would you choose? How would you justify your decision?
5. To what extent is it possible to be certain how effective a medical informatics resource really is? What are the most important criteria of effectiveness?
4. Do you believe that independent, unbiased observers of the same behavior or outcome should agree on the quality of that outcome?
5. Many of the evaluation approaches assert that a single unbiased observer is a legitimate source of information in an evaluation, even if that observer's data or judgments are unsubstantiated by other people. Give examples drawn from our society where we vest important decisions in a single experienced and presumed impartial individual.
6. Do you agree with the statement that all evaluations appear equivocal when subjected to serious scrutiny? Explain your answer.

Appendices

Appendix A: Two Evaluation Scenarios

Here we introduce two scenarios that collectively capture many of the dilemmas facing those planning and conducting evaluations in biomedical informatics:

1. A prototype information resource has been developed, but its usability and potential for benefit need to be assessed prior to deployment;
2. A commercial resource has been deployed across a large enterprise, and there is need to understand its impact on users as well as on the organization.

These scenarios do not address the full scope of evaluations in biomedical informatics, but they cover a lot of what people do. For each, we introduce sets of evaluation questions that frequently

arise and examine the dilemmas that investigators face in the design and execution of evaluation studies.

Scenario 1: A Prototype Information Resource has Been Developed, but its Usability and Potential for Benefit Need to Be Assessed Prior to Deployment

The primary evaluation issue here is the upcoming decision to continue with the development of the prototype information resource. Validation of the design and structure of the resource will have been conducted, either formally or informally, but not yet a usability study. If this looks promising, a laboratory evaluation of key functions is also advised before making the substantial investment required to turn a promising prototype into a system that is stable and likely to bring more benefits than problems to users in the field. Here, typical questions will include:

- Who are the target users, and what are their background skills and knowledge?
- Does the resource make sense to target users?
- Following a brief introduction, can target users navigate themselves around important parts of the resource?
- Can target users carry out a selection of relevant tasks using the resource, in reasonable time and with reasonable accuracy?
- What user characteristics correlate with the ability to use the resource and achieve fast, accurate performance with it?
- What other kinds of people can use it safely?
- How to improve the layout, design, wording, menus etc.
- Is there a long learning curve? What user training needs are there?
- How much on-going help will users require once they are initially trained?
- What concerns do users have about the system – e.g., accuracy, privacy, effect on their jobs, other side effects
- Based on the performance of prototypes in users' hands, does the resource have the potential to meet user needs?

These questions fall within the scope of the usability and laboratory function testing approaches listed in Table 11.1. A wide range of techniques–borrowed from the human-computer interaction field and employing both objectivist and subjectivist approaches–can be used, including:

- Seeking the views of potential users after both a demonstration of the resource and a hands-on exploration. Methods such as focus groups may be very useful to identify not only immediate problems with the software and how it might be improved, but also potential broader concerns and unexpected issues that may include user privacy and long term issues around user training and working relationships.
- Studying users while they carry out a list of pre-designed tasks using the information resource. Methods for studying users includes watching over their shoulder, video observation (sometimes with several video cameras per user); think aloud protocols (asking the user to verbalize their impressions as they navigate and use the system); and automatic logging of keystrokes, navigation paths, and time to complete tasks.
- Use of validated questionnaires to capture user impressions, often before and after an experience with the system, one example being the Telemedicine Preparedness questionnaire (Demiris et al. 2000).
- Specific techniques to explore how users might improve the layout or design of the software. For example, to help understand what users think of as a "logical" menu structure for an information resource, investigators can use a card sorting technique. This entails listing each function available on all the menus on a separate card and then asking users to sort these cards into several piles according to which function seems to go with which [www.useit.com].

Depending on the aim of a usability study, it may suffice to employ a small number of potential users. Nielsen has shown that, if the aim is to identify only major software faults, the proportion identified rises quickly up to about 5 or 6 users then much more slowly to plateau at about 15–20 users (Nielsen 1994). Five users will often

identify 80 % of software problems. However, investigators conducting such small studies, useful though they may be for software development, cannot then expect to publish them in a scientific journal. The achievement in this case is having found answers to a very specific question about a specific software prototype. This kind of local reality test is unlikely to appeal to the editors or readers of a journal. By contrast, the results of formal laboratory function studies, that typically employ more users, are more amenable to journal publication.

Scenario 2: A Commercial Resource Has Been Deployed Across a Large Enterprise, and There Is Need to Understand its Impact on Users as Well as on the Organization

The type of evaluation questions that arise here include:

- In what fraction of occasions when the resource could have been used, was it actually used?
- Who uses it, why, are these the intended users, and are they satisfied with it?
- Does using the resource improve influence information/communication flows?
- Does using the resource influence their knowledge or skills?
- Does using the resource improve their work?
- For clinical information resources, does using the resource change outcomes for patients?
- How does the resource influence the whole organization and relevant sub units?
- Do the overall benefits and costs or risks differ for specific groups of users, departments, the whole organization?
- How much does the resource really cost the organization?
- Should the organization keep the resource as it is, improve it or replace it?
- How can the resource be improved, at what cost, and what benefits would result?

To each of the above questions, one can add: "Why, or why not ?", to get a broader understanding of what is happening as a result of use of the resource.

This evaluation scenario, suggesting a problem impact study, is often what people think of first when the concept of evaluation is introduced. However, we have seen in this chapter that it is one of many evaluation scenarios, arising relatively late in the life cycle of an information resource. When these impact-oriented evaluations are undertaken, they usually result from a realization by stakeholders, who have invested significantly in an information resource, that the benefits of the resource are uncertain and there is need to justify recurring costs. These stakeholders usually vary in the kind of evaluation methods that will convince them about the impacts that the resource is or is not having. Many such stakeholders will wish to see quantified indices of benefits or harms from the resource, for example the number of users and daily uses, the amount the resource improves productivity or reduces costs, or perhaps other benefits such as reduced waiting times to perform key tasks or procedures, lengths of hospital stay or occurrence of adverse events. Such data are collected through objectivist studies as discussed earlier. Other stakeholders may prefer to see evidence of perceived benefit and positive views of staff, in which case staff surveys, focus groups and unstructured interviews may prove the best evaluation methods. Often, a combination of many methods is necessary to extend the investigation from understanding what impact the resource has to why this impact occurs – or fails to occur.

If the investigator is pursuing objectivist methods, deciding which of the possible effect variables to include in an impact study and developing ways to measure them can be the most challenging aspect of an evaluation study design. (These and related issues receive the attention of five full chapters of a textbook by the authors of this chapter (Friedman and Wyatt 2005).) Investigators usually wish to limit the number of effect measures employed in a study for many reasons: limited evaluation resources, to minimize manipulation of the practice environment, and to avoid statistical analytical problems that result from a large number of measures.

Effect or impact studies can also use subjectivist approaches to allow the most relevant "effect" issues to emerge over time and with increasingly deep immersion into the study environment. This emergent feature of subjectivist work obviates the need to decide in advance which effect variables to explore, and is considered by proponents of subjectivist approaches to be among their major advantages.

In health care particularly, every intervention carries some risk, which must be judged in comparison to the risks of doing nothing or of providing an alternative intervention. It is difficult to decide whether an information resource is an improvement unless the performance of the current decision-takers is also measured in a comparison-based evaluation. For example, if physicians' decisions are to become more accurate following introduction of a decision-support tool, the resource needs to be "right" when the user would usually be "wrong" This could mean that the tool's error rate is lower than that of the physician, or its errors are in different cases, or they should be of a different kind or less serious than those of the clinician, so as not to introduce new errors caused by the clinician following resource advice even when that advice is incorrect – "automation bias" (Goddard et al. 2012).

For effect studies, it is often important to know something about how the practitioners carry out their work prior to the introduction of the information resource. Suitable measures include the accuracy, timing, and confidence level of their decisions and the amount of information they require before making a decision. Although data for such a study can sometimes be collected by using abstracts of cases or problems in a laboratory setting (Fig. 11.2), these studies inevitably raise questions of generalization to the real world. We observe here one of many trade-offs that occur in the design of evaluation studies. Although control over the mix of cases possible in a laboratory study can lead to a more precise estimate of practitioner decision making, ultimately it may prove better to conduct a baseline study while the individuals are doing real work in a real practice setting. Often this audit of current decisions and actions provides useful input to the design of the information resource, and a reference against which resource performance may later be compared.

When conducting problem impact studies in health care settings, investigators can sometimes save themselves much time and effort without sacrificing validity by measuring effect in terms of certain health care processes, rather than patient outcomes (Mant and Hicks 1995). For example, measuring the mortality or complication rate in patients with heart attacks requires data collection from hundreds of patients, as complications and death are (fortunately) rare events. However, as long as large, rigorous trials or meta-analyses have determined that a certain procedure (e.g., giving heart attack patients streptokinase within 24 h) correlates closely with the desired patient outcome, it is perfectly valid to measure the rate of performing this procedure as a valid "surrogate" for the desired outcome. Mant and Hicks demonstrated that measuring the quality of care by quantifying a key process in this way may require one tenth as many patients as measuring outcomes (Mant and Hicks 1995).

Appendix B: Other Views of Evaluation that Bear on Informatics

The field of evaluation continues to evolve. We describe briefly below two perspectives on evaluation that reflect the goals of making evaluation relevant and useful. Biomedical informatics has inherited from the culture of biomedical research a default vision of evaluation that reflects the fully randomized clinical trial as the gold-standard for determining the truth expressed as cause and effect relationships, and which, in the parlance of this chapter, puts objectivist comparison-based studies on a pedestal. In the limited space of this chapter, while devoting the most space to objectivist studies, we have introduced the complementary subjectivist approaches. Overall, objectivist comparison-based studies are limited by the time and expense of conducting them. This can be particularly problematic in a field

like informatics where the information resources themselves change very rapidly and yet, paradoxically, have to be "frozen" for the full duration of an objectivist study if the study is going to be internally valid. Readers interested in addressing this challenge are encouraged to read further about the two methods described below, and other emerging alternatives – if only to consider for themselves whether and how these approaches might apply their work, and perhaps dismiss them.

Realist Evaluation

"Realist" or "Realistic" evaluation is based on Pawson and Tilley's work (Pawson and Tilley 1997) and has started to influence the design and interpretation of a handful of studies in biomedical informatics. This approach is based on a subset of the philosophical school of realism called scientific realism, which asserts that both material and social worlds are 'real', in the sense that they can have real effects. The aims of realist evaluation are thus to work towards a better understanding of material and social elements which can cause change, to acknowledge that change can occur in both material and social dimensions, and that both are important. Some of the insights made by Pawson and Tilley which underlie the realist approach to evaluation include:

- Many interventions are an attempt to address a social problem – that is, to create some level of social change.
- Many interventions, such as information resources, work by enabling participants to make different choices, although these choices are usually constrained by participants' previous experiences, beliefs and attitudes, opportunities and access to resources.
- Making and sustaining different choices requires a change in participant's reasoning (for example, values, beliefs, attitudes, or the logic they apply to a particular situation) and/or the resources (e.g. information, skills, material resources, support) they have available to them. This combination of "reasoning and resources" is what causes the impact of the intervention and is known in realist evaluation as the intervention "mechanism".
- Interventions work in different ways for different people, i.e. Interventions can trigger different change mechanisms in different participants.
- The contexts in which interventions are delivered often makes a difference to the outcomes they achieve. Relevant contexts may include social, economic and political structures, organizational context, participants, staffing, geographical and historical context, and so on.
- Some factors in the context may enable particular mechanisms to be triggered, while other factors may prevent this. There is always an interaction between context and mechanism, and that interaction is what creates the intervention's impacts or outcomes: Context + Mechanism = Outcome.
- Because interventions work differently in different contexts and through different change mechanisms, they cannot simply be replicated from one context to another and automatically achieve the same outcomes. Good understanding about "what works for whom, in what contexts, and how" is, however, portable.
- Therefore, one of the tasks of evaluation is to learn more about "what works for whom", "in which contexts particular interventions do and don't work", and "what mechanisms are triggered by what interventions in what contexts".

It is important to note that Realist Evaluation is derived from – and is largely applied to – social and educational programs, where the context (rather than the intervention) is likely to be a much more important determinant of the outcome. We believe that this rarely applies in the study of biomedical informatics and information resources. In addition, the message of realist evaluation – that each study's results can only be applied in the context in which they were derived – will seem rather pessimistic, even deconstructivist, to most scientists. This is because the aim of science is to progressively develop better grounded theories that we can confidently use to predict the impact of interventions in a wide range of – though not necessarily all – contexts. Arguably, if Pawson and Tilley's realist

approach applied throughout biomedical informatics, we could not confidently generalize about the impact of any intervention from the results of any evaluation studies. This in turn would make biomedical informatics a discipline in which progress based on the work and findings of others was difficult, if not impossible. This is manifestly untrue – as this book amply demonstrates. However, there may be some biomedical informatics settings in which the context is more important – and more variable – than the intervention, so Realist evaluation methods would then be more appropriate.

Utilization Focused Evaluation

Based on pragmatism rather than theory, Utilization-Focused Evaluation begins with the assumption that evaluations should be judged by their utility and the actual use and impact of the results (Patton 1999). The implication is that evaluators should facilitate the evaluation process and design any evaluation with careful consideration of how everything that is done, *from beginning to end*, will affect the use of the results. Use concerns how real people in the real world apply evaluation findings and experience the evaluation process. Therefore, the *focus* in utilization-focused evaluation is on intended use by intended users.

Since no evaluation study is entirely value-free, utilization-focused evaluation addresses the question, "*Whose values should frame the evaluation* ?" by working with clearly identified, primary intended users, who in turn have the responsibility to apply the evaluation findings and implement the recommendations. Any study based on the principles of utilization-focused evaluation is thus highly personal and situational. The evaluation facilitator develops a working relationship with intended users of the results to help them determine what kind of evaluation they need. This requires negotiation with the evaluator offering a menu of possibilities within the framework of established evaluation methods, approaches and principles.

As a result, utilization-focused evaluation does not advocate any particular evaluation content, model, method, theory, or even use. Instead, it is a process for helping primary intended users select the most appropriate model, methods, theory, and uses for their particular situation. The need to respond to the situation guides the interaction between the evaluator and primary intended users. A utilization-focused evaluation can therefore include any evaluative purpose (e.g. formative, summative, developmental), any kind of data (e.g. quantitative, qualitative, mixed), any kind of design (e.g., naturalistic, experimental), and any kind of focus (e.g. processes, outcomes, impacts, costs, and cost-benefit). Utilization-focused evaluation is a process for helping evaluators to make decisions about these issues in collaboration with an identified group of primary users of the study results, focusing on the intended uses of the evaluation. Collaborative evaluation is a further development of this approach (Rodriguez-Campos 2012).