Developers and Evaluation of Interactive Health Communication Applications

Joseph Henderson, MD, John Noell, PhD, Thomas Reeves, PhD, Thomas Robinson, MD, Victor Strecher, PhD for the Science Panel on Interactive Communications and Health[a]

Background: Developers of Interactive Health Communication (IHC) are capable of providing great benefit by creating interactive programs that serve to protect and improve health. Conducting proper evaluation of these programs will ensure that they achieve these goals more successfully.

Conclusions: This article seeks to inform developers of IHC about which types of evaluation are most important to include as a part of the development process and to examine the ways in which such evaluation can be implemented to benefit the producers—and ultimately, the consumers—of IHC.

Medical Subject Headings (MeSH): computer communication networks, evaluation studies, (medical informatics), consumer participation (Am J Prev Med 1999;16(1):30–34)

Introduction

The use of evaluations can provide significant benefits to developers of interactive health communication (IHC) applications, as well as to the purchasers and users of their products. Developers are a vital link between the consumer and various sources of health-related information. As such, the developers of IHC applications are critically important to the future of health care and prevention. However, the world of interactive applications is rapidly becoming more complex; developers of these applications face numerous obstacles as they strive to create and disseminate effective and useful programs. Unfortunately, they often find themselves caught between a rapidly shifting marketplace (or lack of a current market) and the real-life constraints that influence program development. Economic success (or even survival) is difficult under such circumstances.

The ever more-frequent call for evaluations of interactive applications presents developers with a critical choice; such evaluations can be seen as yet another obstacle—or as a great opportunity. Some forms of evaluation are already an integral part of the development process, although they are often described in other terms, such as checking code against specifications, alpha testing, and beta testing. These can be grouped under categories of evaluation called formative and process. Unfortunately, developers do not typically conduct more difficult outcomes (also called summative) evaluations of actual impact on real users, yet this type of evaluation can be of profit: Impact and effectiveness are of great interest to purchasers; users like to know if others found the program accurate, useful, and enjoyable. When developers of IHC applications understand the opportunities inherent in evaluation and how to minimize the potential obstacles of such evaluation, they can maximize their—and the end users'—benefit.

Unfortunately, there is little doubt that the power of interactive applications can also result in harm, such as that done by providing inaccurate or poorly framed information. To the extent that developers can assure all, including themselves, that their applications produce benefits—and do no harm—the marketplace will

From Dartmouth Medical School, Dartmouth, NH (Dr. Henderson), Oregon Center for Applied Science, Inc. and Oregon Research Institute, Eugene, OR (Dr. Noell); University of Georgia, Athens, GA (Dr. Reeves), Stanford University School of Medicine, Palo Alto, CA (Dr. Robinson); and University of Michigan Comprehensive Cancer Center, Ann Arbor, MI (Dr. Strecher).

Address correspondence to: Thomas R. Eng, Office of Disease Prevention and Health Promotion, HHS, 200 Independence Ave., SW, Room 738G, Washington, DC 20201

Address reprint requests to: Mary Jo Deering, PhD, Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services, 200 Independence Ave., SW, Washington, DC 20201.

[a]Other panel members and staff: Linda Adler, MPH, MA, National Member Technology Group, Kaiser Permanente, Oakland, CA; Farrokh Alemi, PhD, Cleveland State University, Cleveland, OH; David Ansley, Consumer Reports, Yonkers, NY; Patricia Flatley Brennan, RN, PhD, FAAN, School of Nursing and College of Engineering, University of Wisconsin-Madison, Madison, WI; Molly Joel Coye, MD, MPH, The Lewin Group, San Francisco, CA; Mary Jo Deering, PhD, Thomas R. Eng, VMD, MPH, Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services, Washington, DC; David Gustafson, PhD, University of Wisconsin-Madison, Madison, WI; Holly Jimison, PhD, Oregon Health Sciences University, Portland, OR; Albert Mulley Jr, MD, MPP, Massachusetts General Hospital, Boston, MA; and Kevin Patrick, MD, MS, San Diego State University, San Diego, CA.

0749-3797/99/$–see front matter
PII S0749-3797(98)00106-8

grow more rapidly. The health care consumers and providers who are potential buyers of IHC applications are becoming more skeptical of their effectiveness as the number of available applications expands. Increasingly, potential buyers are heard to say, "It may be flashy, but how well does it work?" They are inquiring also about potential liabilities, "Is there any danger if we use this program? Is it accurate, reliable, and safe?"

It is becoming essential to have data to support claims of efficacy and safety. However, the constraints of tight budgets and schedules can force developers to make tough choices regarding program evaluation. Therefore, it is important to identify the factors that matter most, the best approaches to the measurement of such factors, and the most appropriate standards of evidence. Developers can be better prepared to include evaluations as part of the development process, which will strengthen their products and consumer confidence. Indeed, including *formative* evaluations during development can increase greatly the chances that the final product will achieve its intended purpose and minimize the potential for costly revisions. *Process* evaluations usually are used to keep projects on-schedule and on-track. For developers, process evaluations most often take the form of alpha and beta testing. Finally, *outcome* evaluations often are the most important to potential purchasers. This third type of evaluation helps answer such key questions as, "Does it work in terms of impact on the end-user?" All three types of evaluation are described subsequently.

Before describing the advocated approach (which will provide a framework for developers to better evaluate their applications), it seems appropriate to identify what is *not* being suggested: A multi-year, randomized trial with a large subject sample is *not* the level of evaluation befitting most situations. The technology is far too dynamic, and lessons learned would have limited relevance, especially in proportion to the effort involved in the trial. In addition, the uniform adoption of a scientific standard, e.g., a "$P \leq 0.05$" level of statistical significance, should not be seen as a basic necessity. Indeed, determining the effect magnitude (i.e., of program impact) usually is more meaningful, and often more useful, than reporting significance levels. In other words, the fact that there is an effect decreases in importance if the actual impact of that effect is minor.

A balanced approach to program evaluation is advocated here, one that takes into consideration the need to conserve development resources (e.g., cost and time) while still protecting the safety of users and ensuring that the program is effective. This approach also includes a realistic appraisal of the advantages associated with being able to assure potential buyers that a particular program *has* been evaluated—and found to be effective, accurate, and safe.

Benefits of Evaluation

If developers are to adopt evaluation as an integral part of program development, they first must see the value in doing so. There are numerous potential benefits to the general public, some of which are identified in other articles in this series. Here are some for developers themselves:

- *Sales are likely to increase.* Many consumers and purchasers tend to perceive evaluated products to be of better quality (i.e., more desirable) than those that have not been examined. For example, products that receive high ratings from independent consumer organizations tend to sell much better than those that are either not rated highly or not evaluated at all.
- *Profit margins may be higher.* Consumers often are willing to pay more for an evaluated product that has been reviewed favorably.
- *Increased market share.* Evaluated products are likely to be seen as more trustworthy. For example, health plans and other large purchasers of IHC applications tend to be interested in products that are likely to be cost-effective for their organization. Products that have been evaluated and shown to have this advantage are much more likely to be purchased in mass quantities than products that have not been evaluated.
- *Improved effectiveness, utility, and reliability of the product.* By incorporating evaluation methods throughout the development and implementation process, the developer can gain valuable feedback from end users to improve product design and ensure a more attractive, effective, and user-friendly application.
- *Evaluations may decrease potential liability for harm caused by a product.* Developers who have evaluated their product thoroughly to minimize any associated health risks may be less likely to be found negligent if an individual claims the product resulted in some harm.
- *Evaluations may minimize or prevent government regulation of these products.* Without a standard by which all products are evaluated in order to prevent the release of potentially harmful programs, it is more likely that some programs will result in severe health consequences. If such situations do occur, it is likely that the public will call for government regulation of the industry. Voluntary compliance of developers to ensure the quality and effectiveness of their products through routine evaluation may forestall such a situation and any resulting government intervention.

Of course, evaluation is not solely for the benefit of developers. Ultimately, evaluations are intended to ensure that consumer health information systems meet the needs of consumers, health care providers, and policy-making bodies. The history of health care is full of examples of fraudulent or mistaken claims about the efficacy of medical products. Consumer health infor-

mation products may also be subject to similar fraud or errors of judgment, and evaluation is one of the few tools available to combat these tendencies. In addition to issues of self-interest, developers have a moral and ethical responsibility to engage in evaluation that maximizes safety and avoids injury or error.

Evaluation Criteria

There are six key criteria for evaluation that can be applied to most programs: accuracy, appropriateness, usability, maintainability, bias, and efficacy.

- *Accuracy* of content is of paramount importance. However, as discussed subsequently, it is often not obvious how to ensure the accuracy of the content. Accuracy includes a number of components, including currency and validity. Sometimes there is a tension between the two, in that the newest information may not hold up through the tests of time and broader experience. In addition, some information can be accurate and still be misleading. This, too, needs to be considered.
- *Appropriateness* of content includes two factors: applicability and intelligibility to the user. Not all programs are intended for all people. First it must be made explicit exactly *who* the intended users are. Then it is important to make sure that the content actually is applicable to all such users, and that they can understand the content.
- *Usability* is a measure of how easily a user can get the program to do what it is intended to do. This is where the interface design is critical. A flashy interface may look good at first glance, but actually make an application harder or more intimidating to use. Another component of usability is acceptability. Developers must take care that the interface, or elements thereof, not antagonize the end user.
- *Maintainability* is important because both content and users are likely to shift over time, therefore requiring modifications to the program. It is important to consider who is to make those changes and how easily (i.e., by what means and at what cost) they can be done.
- *Bias* is a factor that developers of a program often overlook. Clearly, bias cannot be eliminated completely because the perception thereof is dependent on the individual user. Nevertheless, it is important to be sensitive to and aware of both potential and actual biases. For example, if a program incorporates an assumption that alternative medicine is good (or that alternative medicine is bad), it can be both limiting (e.g., to whom it has sales appeal) and dangerous (e.g., in terms of liability to both developer and provider).
  Although conflicts of interest do not necessarily lead to bias, it often is nearly as important to avoid the *appearance* of bias as it is to avoid it in reality. Therefore, it is incumbent on the part of the evaluators to avoid any potential conflicts of interest, if at all possible. Where it is not possible, it is essential to use the most objective criteria (those most resistant to bias) available.
- *Efficacy* is a measure of the extent to which a program actually has its intended impact, e.g., on behavior change (does the program actually help more people to stop smoking?) or on decision making (does the program provide adequate, reliable information for the consumer to make an informed decision?). A similar concept is effectiveness. Technically, efficacy refers to a program's impact under controlled (i.e., experimental) conditions, while effectiveness is the program's impact under real-world conditions. (Sometimes a program may have efficacy but not, ultimately, be effective.)

In today's highly dynamic information and technologic environment, development—and accompanying evaluations—can never really be thought to have an end-point. Information will become outdated, new information will become available, and the ways in which information is presented to the individual end-user will evolve as delivery methods evolve. For many IHC applications, development involves a long-term commitment to a process of updating and revision, with on-going quality-assurance evaluations.

Types of Evaluations

The basic questions of evaluations usually are straightforward: What are the goals and objectives of the application? Are we heading in the right direction to accomplish those goals and objectives? Did we get there (i.e., does the application do what it is supposed to do)? Few fields are as complex, or as filled with philosophical variety, as the field of evaluation. However, developers can conduct meaningful, valid, and illuminating evaluations without becoming lost in complexities and philosophical issues.

Evaluation can range from informal to formal and from simple to complex. The three basic types of evaluation are formative, process, and outcome (or summative). *Formative* evaluations determine what program(s) to create, and how they will look and work when finished. Measures of user satisfaction are also considered part of the process of formative evaluation. (Although the end-user is involved in such evaluation, it is not considered summative since it can be measured at early or intermediate stages to improve the product.) *Process* evaluations look at the process of developing, testing, or implementing an application: Is (each phase of) the project still on time? Is it on track? Is implementation proceeding according to plan? *Outcome* evaluations are those that are conducted at the end of a

development phase, and are used to determine if the goals of the program have been met. Simply put, what *are* the outcomes—how well does it work?

All six key criteria listed can be used in formative, process, and outcome evaluations. Indeed, all six *should* be used for most programs. The relative importance of each will depend on the program and the targeted audience.

Specific evaluation techniques, such as surveys, direct observations, focus groups, and interviews, are not the primary purview of this article and will not be discussed at any length. The exact needs of the developer must be considered before choosing any particular technique. As with the criteria listed previously, however, all of these techniques can be used as part of formative, process, or outcome evaluations. The *Evaluation Reporting Template*, proposed by the Science Panel on Interactive Communications in Health, includes several questions concerning evaluation and serves as a general guide for developing evaluations.[1]

Formative Evaluations

The purchasers and end-users of a program are more likely to ask about outcome evaluations; however, formative evaluations, if conducted carefully, can have tremendous influence on a potential buyer. Additionally, formative evaluations are often of the greatest direct value to the developer. It is through the use of formative evaluation approaches, such as interviews and focus groups, that the developer can be assured that the final product will work *and* have a market. There are many examples of developers who have produced programs they were certain were in demand, only to find that others did not share their view—and did not buy. Nearly all of us have used programs that contained elements that were confusing or seemed illogical. Formative evaluations can provide the ounce of prevention that avoids the necessity for applying a pound of [expensive] cure after the program was supposed to be finished.

The *Evaluation Reporting Template*[1] can help in developing a formative evaluation plan. The first questions usually have to do with the specific goals and objectives of the program: What is it intended to do? For whom is it intended? What do they need to use the program? The next set of questions gets more detailed: What is the required reading level (if applicable)? What is average time required for use? How long does it take to train someone to use the program?

A key issue that is addressed in the formative evaluation stage is accuracy, especially that of content. It is axiomatic that if information is to be valuable, it must be accurate overall and accurate for the individual user. An important key to accuracy of content is good staffing, committed for the life of the project (initial development, revision, and maintenance phases). The use of content experts at this formative stage of development can make a huge difference, saving many hours of lost effort. The development team should recruit the appropriate experts at the project's inception.

The basic purpose of formative evaluation in this context is to protect the developer from wasting time and energy creating a program that is not needed, that will not be purchased or used, and/or that, in the worst case, will do harm to individuals. If the program lacks an audience, or does not fill a perceived need, the developer may end up wasting time and money. If the program does harm, the developer has much more serious concerns.

Process Evaluations

Process evaluations look specifically at processes such as development, program implementation, or even evaluation. Developers usually are familiar with process evaluations, at least in the forms of alpha and beta testing. However, it is often important to track the process of program implementation, especially to ensure that an application has been used appropriately, that users understand and appreciate it, or that its evaluation is being carried out properly.

Outcome Evaluations

There are two basic questions in most outcome evaluations: [How well] does it work? and, Does it do anything it shouldn't? Obviously, the critical aspects of these questions are in the details. In what specific ways does it work or not work? As shown on the evaluation template, the first questions often asked are, Did the users like the program? and Did they find it helpful and useful? Questions that should be asked at every phase of development. Unless people like it or find it useful, it is not likely to be used (even if they *should* use it).

Equally important to many purchasers and users are questions about how well it works to increase knowledge, change beliefs or attitudes, or change behaviors. If it fails in these areas, a program is not likely to have a positive effect on targeted situations, such as quality-of-life, morbidity, mortality, resource utilization, or organizational culture. Programs that do not affect one or more of these domains are unlikely to be successful economically. Prudent developers will conduct some sort of outcome evaluations (or arrange to have them done by others) to increase sales and marketability, while limiting liabilities.

Once the questions, Does it work? and Does it do anything it shouldn't? have been asked (and answered), the next question is usually, Can I believe the results of the evaluation? Answering this question is a function of applying the statistical measures that comprise standards of evidence.

Standards of Evidence

Many of the debates in the field of evaluation have to do with standards of evidence. Two central concepts are the reliability and validity of the evaluation. There are a number of techniques for measuring reliability and validity, and a methodologic or statistical consultant can be helpful in this area.

Reliability can be seen as repeatability: If I asked the same question of the same people again, would I get the same answer? When the wrong questions are asked, or the questions are asked incorrectly, people can give answers that have little to do with the intent of the questions, or they can give answers that vary from one time to another. Therefore, it is important to be sure that what one is asking is understood fully by the people who are being asked, and that they actually can provide dependable (i.e., reliable) answers.

The validity of evaluation findings can be viewed as the truthfulness of the findings. Are the findings correct, or are they an aberration? Are they meaningful in this context? There are two types of validity: internal and external. Internal validity is the validity of the findings within the study itself. External validity is the validity of applying the findings to other situations. External validity often is referred to as generalizability. If the people on whom we tested the program liked it, will everyone else who uses it have the same overall reaction? Can the results obtained with the study sample be generalized to other groups as well? Generalizability can be critically important because in some situations developers may want to rely on the findings or results obtained by others. For example, if tailoring has improved message impact for others, it may be more appropriate for a developer simply to adopt a proven approach rather than conduct additional evaluations.

Often, there is great emphasis on the statistical significance of the outcome findings. While this may be an appropriate concern, it can be over-emphasized. The key notion underlying statistical significance is to what degree are the results an accurate reflection of reality, and not due to chance? Is there really a connection between use of the program and the outcomes? What are the odds that the outcomes really are due to the program, rather than due to chance and chance alone? The traditional metric of scientific studies is, $P \geq 0.05$, which simply means that no more than 5% of the time, or 1 in 20 times, would you expect a given result to occur by chance. In other words, there is at least a 95% probability that the outcomes occurred because of the effectiveness of the program, and not by chance. However, effect size often is a more important concern.

Effect size is the term used to describe the magnitude of impact the program has on its users. If a program is supposed to encourage people with diabetes to monitor their blood sugar more carefully, just *how much more* (or less) carefully do they do it after using the program? If a program is designed to decrease utilization of a service, *to what extent* do users of the program use that service less (or more) than people who did not use the program? While the statistical significance of the evaluation's results is important, it may be more important to know how strongly it affected the users.

Obviously, everyone will want to be assured that the results they see are not due to some random factor. Therefore, it becomes important to test programs enough times with enough people and to choose test subjects appropriately. While determining just how best to test a program can be quite complicated, common sense is usually the best guide. The evaluation template, mentioned previously, offers a number of basic things to think about when developing outcome evaluations.

Summary

Developers, especially those with limited resources (e.g., time and money), are likely to find that the timely use of evaluations, rather than diminishing those resources, can provide significant assistance in developing effective, profitable programs. The use of evaluations also can potentially reduce liabilities. Finally, voluntary use of evaluations can reduce the probability of government intervention and regulation.

The six key criteria listed here (accuracy, appropriateness, usability, maintainability, bias, and efficacy) can form the basis for developing appropriate evaluations. Taken together with the specific questions included in the *Evaluation Reporting Template*,[1] these criteria can be used as a roadmap to success as a developer of interactive applications. The developer community is capable of providing great benefit to the community at large by creating interactive programs that ultimately serve to protect and improve health. Conducting proper evaluation of these programs will ensure that they achieve these goals more successfully.

The authors are grateful to Paul Kim; Andy Maxfield, PhD; Anne Restino, MA; and John Studach, MA; for their contributions to the panel's work, and to Maria-Teresa Wilson and Linda Friedman for assistance with copy editing. In addition, the authors thank the liaisons to the Science Panel on Interactive Communication and Health, especially the following persons who offered valuable suggestions for improving this manuscript: Loren Buhle, PhD; David Cochran, MD; Connie Dresser, RDPH, LN; Alex Jadad, MD; Craig Locatis, PhD; Ed Madara; Georgia Moore; Kent Murphy, MD; Scott Ratzan, MD, MPA, MA; Helga Rippen, MD, PhD; and Christobel Selecky.

Reference

1. Robinson TN, Patrick K, Eng TR, Gustafson D, for the Science Panel on Interactive Communication and Health. An evidence-based approach to interactive health communication: A challenge to medicine in the information age. JAMA 1998;280:1264–9.