Field trials of medical decision-aids: potential problems and solutions

Jeremy Wyatt, Medical Informatics, Stanford University, Stanford, CA 94305-5479
(from August 1st 1992: National Heart & Lung Inst., Fulham Rd, London SW3 6HP, UK)

David Spiegelhalter, MRC Biostatistics Unit, 5 Shaftesbury Road, Cambridge CB2 2BW, UK

Abstract

Only clinical trials can assess the impact of prototype medical decision-aids, but they are seldom performed before dissemination. Many problems are encountered when designing such studies, including ensuring generality, deciding what to measure, feasible study designs, correcting for biases caused by the trial itself and by the decision-aid, resolving the "Evaluation Paradox", and potential legal and ethical doubts. These are discussed in this paper.

Introduction

We are all aware of the problems of medical decision-making, and of the promise of medical decision-aids to alleviate some of these. Very few decision-aids have, however, passed into routine clinical use [1, 2]. One reason may be that this is a research area, with few workers building systems to address real-world problems. Another may be the reluctance of developers to subject systems to rigorous clinical evaluation because of the resources required or the complexity of such studies. Much has been written about the testing of decision-aids in laboratory settings [eg. 3, 4, 5], and such studies are now performed more frequently. This paper reviews the problems of field trials, and suggests some possible solutions.

Definitions

We define medical decision-aids as *"active knowledge systems which use two or more items of patient data to generate case-specific advice"*, in contrast to passive knowledge systems in which the user conducts the search through the system's knowledge [6]. A field trial of a decision-aid is *"a clinical trial in which users in their normal environment have access to the decision-aid's advice at a time when it can influence their decisions"*. A key feature of field trials is that users are under no obligation to use the system or to follow its advice. This contrasts with "laboratory" tests of performance in which the system is evaluated as if it were the decision-taker, or tests of the knowledge base or interface conducted using retrospective test data, often by the developers themselves [5]. Even a "formative" evaluation, in which potential users might rate the acceptability of sample advice [7], does not fall within our definition of a field trial.

The need for field trials

It is naive to assume that all decision-aids which show promise in laboratory tests are effective when used clinically. The development of drugs provide a useful analogy: many compounds are promising *in vitro* but fail to deliver their potential *in vivo* because of side-effects, imperfect absorption etc. Such mechanisms are relatively well understood, but still no-one would market a drug without conducting rigorous clinical trials. Even if a decision-aid was 100% accurate, it would only alter doctors' behaviour if they used it and if the advice influenced their decisions. We know very little about how advice influences doctors, so field trials of decision-aids are, if anything, more necessary than with drugs, and should aim to answer four fundamental questions:

1. Will doctors use the decision-aid in a clinical setting ?
2. Will its advice alter their decisions ?
3. Will altered decisions lead to changed behaviour ?
4. Will altered behaviour change patient outcomes ?

Potential problems of field trials

Field trials of decision-aids are open to a wide range of problems. In each of the subsections which follow we describe the problem, when it may prove troublesome, and possible methods for quantifying or eliminating it.

Ensuring that the setting is representative

It is unfortunate if the results of a trial cannot be generalised from the particular patients, doctors and decision-aid studied to similar settings elsewhere. This may happen if a trial is conducted in a tertiary referral centre, or if a better "system" is available than might be elsewhere because the developers provided users with extra support. To avoid these problems, the patients and users who are to be helped by the system must be identified, explicit admission criteria for the trial should be defined, and investigators should ensure that the site of the trial is appropriate. A multicentre trial including various types of centre is ideal, but is expensive and prone to administrative breakdown. To avoid developers providing extra support, the field trial should normally take place away from the development centre.

Selection of measures or end-points

The ultimate intention of medical decision-aids is not to aid in decision-taking, but to assist users in maximising patient health per unit of resource consumed. Thus, although changes in diagnostic accuracy or test ordering may suggest benefit, the impact of these changes on patient outcome should also be assessed. The measures which are of most interest depend on the specific role of

0195-4210/91/$5.00 © 1992 AMIA, Inc.

the decision-aid, but investigators should ensure that they are clinically valid and repeatable, and address the three key aspects of healthcare activity [8]:

Structure: How well does the decision-aid fit into its environment, do users find it helpful ?

Process: What effect does the system have on healthcare processes, such as the accuracy of decisions or the number of treatments given ?

Outcome: Are the effects on healthcare process reflected in patient outcomes, or in "surrogate outcomes" such as control of hypertension ?

Feasible trial designs

Although controlled trials with patients randomised are usual for assessing therapeutic interventions, for other interventions randomisation of doctors, departments or even hospitals has advantages [9]. If the decision-aid is introduced and then withdrawn in a sequential design, this gives greater statistical power and may avoid the difficulty of matching control and decision-aid doctors. However, this design may antagonise staff who believe that the decision-aid is beneficial, so the withdrawal can be synchronised with staff changeover in a full crossover design, or the decision-aid can be incrementally enhanced with the introduction of new facilities or coverage of further diseases. If the fragment introduced at each stage is randomised and there is a final control period, this further strengthens the design, especially if a "dose-response" relationship is sought. An alternative, which may appear attractive if data is already available, is to use historical controls [eg. 10], but changes in casemix and patient management may confound the introduction of the decision-aid. This risk may be reduced by including a final control period.

When planning trials, it is wise to remember that the availability of patients is often overestimated, even by a factor of 10 (Lasagna's Law). Recruitment may be a greater problem in trials of decision-aids since they are not simply prescribed like drugs. Also, the "compliance rate" is likely to be lower than in drug trials: for example, in a trial of the Leeds abdominal pain system, doctors entered patient data in only 45% of cases [11]. Careful engineering of the decision-aid to its clinical niche may improve usage rates; doctors used a paper-based system requiring only one minute to complete in 96% of patients [10].

Correcting for biases caused by the trial

To eliminate bias requires that the control and decision-aid groups are matched. This means that the management of all patients must be identical apart from the availability of advice in the decision-aid group. This is most easily achieved by randomisation, which can be stratified to ensure equal distribution of patients to clinically significant subgroups. However, other biases can still arise (see figure 1).

Recruitment and allocation biases: In a trial where patients are randomised and they or the doctors have a preference for the decision-aid, several biases may arise. Doctors may cheat the randomisation method and allocate more difficult cases to the decision-aid group (allocation bias), they might use the decision-aid illicitly in some control cases [eg. 12], or they might fail to recruit difficult cases to the trial if they know in advance that they will be allocated to the control group (recruitment bias). These biases would all underestimate the system's value. In a trial where doctors or departments are randomised, bias will arise if the doctors' enthusiasm for the decision-aid is correlated with their clinical competence. Thus, inexperienced or insecure doctors might drop-out of the trial less often if allocated to the decision-aid group, and would then use the system more frequently than more competent doctors. These effects would dilute the benefit of the decision-aid with the doctors' incompetence, reducing the system's apparent benefits.

The solution to these problems is to define the population of patients or doctors eligible for the trial, screen them strictly for eligibility, and to randomise patients as late as possible before the system is used. It is wise to check whether doctors can cheat the randomisation method (eg. if envelopes are not completely opaque), to quantify the numbers of patients or users who were *not* recruited, and to analyse the study according to the "intention to provide advice" principle (see below).

Global Hawthorne Effect: The Hawthorne Effect is the tendency for humans to improve their performance if they know it is being studied, discovered by psychologists observing workers at the Hawthorne factory in Chicago [13] during an investigation of the effect of ambient lighting on productivity. Productivity increased as the illumination level was raised, but when the level was accidentally reduced, the workers' productivity again increased, suggesting that it was the study itself rather than changes in illumination which caused the increase. During a trial of a medical decision-aid, the Hawthorne Effect can lead to an improvement in performance of all decision-makers, regardless of access to a decision-aid. This "global" Hawthorne Effect is particularly likely to occur when doctors' performance is low because they lack some simple knowledge or insight which is easy for them to correct [eg. 12], and would lead to the benefit of the system's advice being under-estimated. To quantify a global Hawthorne Effect requires a preliminary, low-profile, prospective study of the performance of decision-makers, before any large-scale interventions are made. Disguising its true intention may take some ingenuity, but is a necessary evil if the Hawthorne Effect is not to contaminate this "baseline" study too.

Extra Work because the trial is in progress: The fact that a trial is in progress often places extra work on users of the decision-aid, who fill-out forms or book extra tests for their patients, reducing their time for decision-making. This may lead to an under-estimate of the value of the system's advice. To eliminate this bias, control doctors should perform the same extra work.

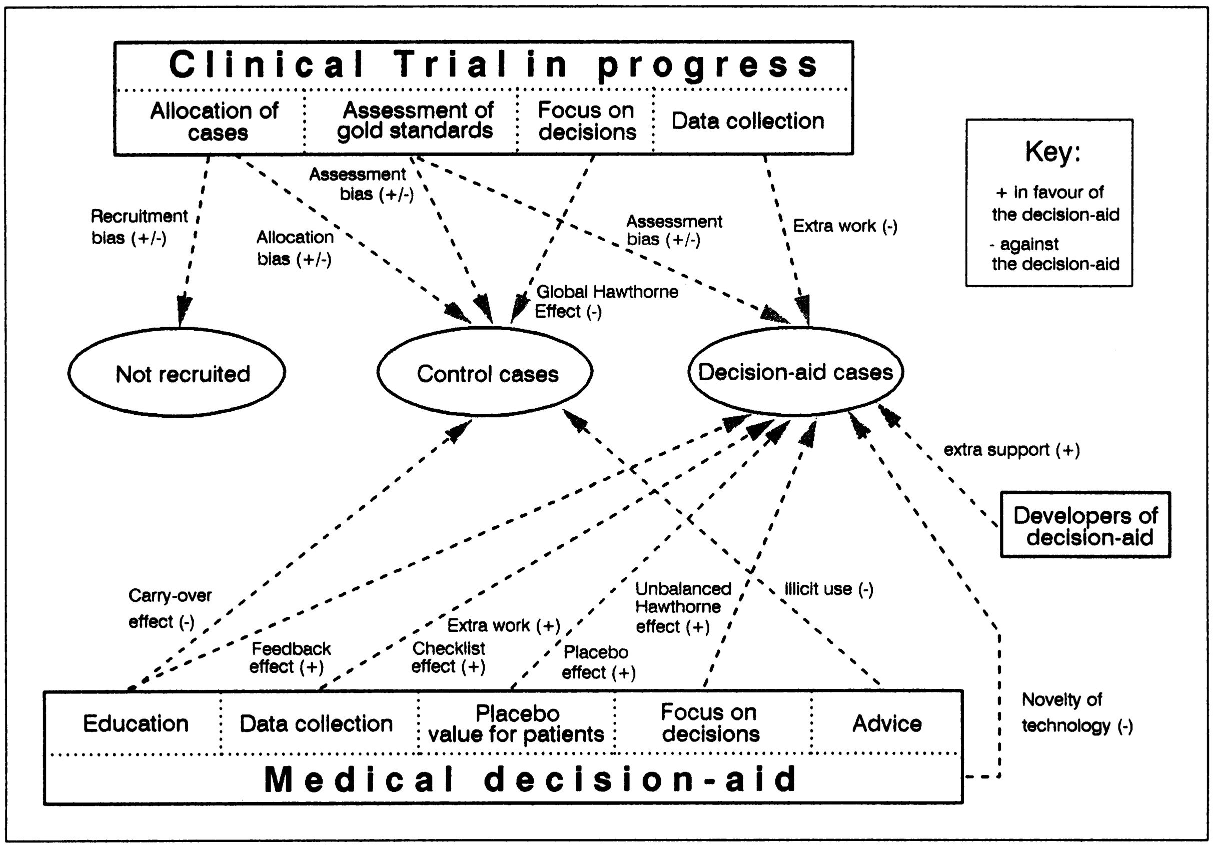

Figure 1: Sources of bias in field trials of medical decision-aids

Assessment bias and bias in follow-up: If the users of a decision-aid also collect data needed for assessing gold standards, they might collect extra data to prove themselves right and the system wrong. Even worse, if the system users provide the actual gold standards, evaluators will be completely reliant on their open-mindedness about the decision-aid. Bias may also affect those assessing whether the doctors' diagnosis or management was correct: if they know in which patients the decision-aid was used, they might be prejudiced in their assessment. This bias is particularly likely if assessors have strong preconceptions about the value of the system, for example if they participated in its construction, and when the criteria used for assessing correct management or outcome are subjective. To eliminate these biases, everyone involved in the collection of follow-up data and assessment of gold standards must be blinded to whether the decision-aid was used. This is only practical if patients are followed-up by a second group of clinicians after an independent person removes any evidence of decision-aid use from the notes. If evaluators use objective measures collected in all cases to classify patients this would alleviate the problem but, unfortunately, such measures are rare in medicine.

Correcting for biases due to the decision-aid

In a trial of a medical decision-aid, an important question is: *"What is the effect of each component of the system ?"*, where possible components include education, data collection, a placebo effect on patients and a focus on decisions, as well as the advice itself (figure 1). To answer it, investigators will need to understand the mechanisms by which decision-aids may influence doctors and patients. These are discussed in the subsections which follow.

The Carry-over Effect: This is a contamination of the management of control patients or the competence of control doctors by doctors who have access to a decision-aid. It is most likely to occur with decision-aids which have an educational effect and will reduce the apparent value of the decision-aid. To eliminate the Carry-over Effect, it is probably best to raise the size of the sampling unit [14], randomising doctors instead of patients [eg. 15], or departments instead of doctors [eg. 11]. This randomisation by group has implications for statistical analysis [16]. To quantify the carry-over, investigators may conduct a study with alternating decision-aid and control periods [eg. 17], which allows carry-over after the decision-aid is withdrawn to be quantified, as long as time-related trends can be ignored.

The Feedback Effect: This is similar to the Carry-over Effect, and occurs if the decision-aid has an audit function, enhancing users' performance by informing them of their failures and successes. It may lead to the value of the advice itself being over-estimated, but is fortunately easy to eliminate by giving the same feedback to control doctors as to decision-aid doctors. Alternatively, it can be quantified by randomising some doctors to a "feedback only" group [eg. 11].

The Checklist Effect: This is an improvement in decision-making due to the more complete and better-structured collection of patient data when paper questionnaires or computers are used. Its impact on decision-making can equal that of the advice [11], and it must therefore either be controlled for or quantified. It is most likely to occur when junior doctors collect complex data under critical time pressures. To control for the Checklist Effect, the same data is collected in the same way in control and decision-aid cases, but the system's advice is only available in the latter group [eg. 12]. To quantify it, a randomly selected "data collection only" group of patients can be recruited [eg. 11].

Extra Work before using a decision-aid: This occurs if the doctors who use the decision-aid perform extra work before using it. For example, users of the Leeds Abdominal Pain system are required to record their own diagnosis before using the system, but during the control period of a field trial, failed to record it in 14% of cases, which were counted as errors [11]. The need to record a diagnosis is a strong incentive to consider clinical evidence carefully, and would over-estimate the value of the advice. To eliminate this bias, control doctors should perform the same extra work, while it could be quantified by recruiting an "extra work only" group of doctors.

Placebo Effect on patients: In drug trials, the Placebo Effect may be more powerful than the drug effect, and may even obscure a total lack of therapeutic benefit. In a trial of a medical decision-aid, if some patients notice that their doctors consult an impressive workstation while others have no such experience, this could unbalance the groups and overestimate the value of the decision-aid. This is most likely to arise when the measurements are highly subjective (patient's mood, satisfaction with therapy etc.) and when the computer is used in front of the patient. One remedy is to arrange that all doctors briefly leave the patient to visit another room, where some would use the decision-aid. It is better, however, to make objective measures of the patients' condition that are immune to the placebo effect.

Unbalanced Hawthorne Effect: In trials where doctors are randomised, the fact that one group of doctors uses a decision-aid would be a constant reminder to them that their decisions are under close scrutiny, which control doctors would lack. This could result in an *unbalanced* Hawthorne Effect, greater in the decision-aid than the control group, and an over-estimate of the decision-aid's benefit. To eliminate this, investigators would need to provide control doctors with a system which produced placebo advice, indistinguishable to them from genuine advice. However, it would be difficult to provide doctors with patient-specific advice which they did not detect as placebo, and did not contribute to patient care. However, this bias is easy to eliminate in a trial comparing the impact of one decision-aid with another.

Illicit use of the decision-aid in control cases: This has already been discussed under allocation biases. To eliminate it requires that access to the decision-aid is only granted after users enter the patient's study number, which must correspond to a decision-aid case.

Novelty of the technology: The introduction of a decision-aid will often coincide with doctors' first exposure to information technology and a need to learn keyboard skills. The benefit of the system may thus be obscured by a learning curve, or even by resentment of the technology. This problem is most likely to affect decision-aids based on novel software, requiring use of unfamiliar input devices, or in institutions without a hospital-wide information system. It will gradually ameliorate as more doctors become familiar with computers and more institutions install information systems, but until then may seriously reduce the benefit of a decision-aid. Once such information systems are installed, the marginal benefit of adding decision-support using data already captured will be much easier to assess.

Analysis by "intention to provide advice"

In field trials of decision-aids it is likely that the system will not always be used in intended cases, and that its advice will often not be followed. There is a close analogy with drug trials: in which group should one include patients who were randomised to the new drug but failed to take it, or who took it but were found to have diminished absorption ? The problem is that if one discards cases who did not take the drug from the analysis, the *average* benefit of giving the drug to patients described by the trial entry criteria will be overestimated. Thus, when analysing drug trials, patients are included in the group to which they were randomised, the *"intention to treat"* principle. The same argument applies to decision-aids: the aim is to detect the *average* impact on patients when the system made available to doctors, not its maximum potential for benefit: indeed, this is the motivation for conducting field trials. Thus, we should not exclude patients in whom doctors failed to use the decision-aid or ignored its advice, but must analyse the trial according to the *"intention to provide advice"*. Control patients in whom the decision-aid was illicitly used are a possible exception to this: but if the doctor was sufficiently uncertain to consult the decision-aid, they might have sought advice from elsewhere had it not been available.

The Evaluation Paradox

In a field trial, doctors will be reluctant to act on the advice of a decision-aid until it has been shown to improve decisions; however, to quantify the system's impact on decision-making, its advice has to be acted on. This paradox is most likely to apply to "black box" decision-aids, which provide little insight into the reasons for their advice [18], and would lead to the benefit of the system being under-estimated. Although one solution might be to deliberately mislead doctors about the benefits of the system, we prefer to provide users with an honest account of the decision-aid's scope and performance in laboratory tests, the differences between its reasoning method and the data it uses compared to those used by doctors, and examples of cases where doctors' decisions outranked the decision-aid's and vice versa. This should encourage users to treat the system as an aid, not a black-box dictator. There is no place for instructing users to follow the advice implicitly, as this will certainly not be the case when the decision-aid is released for general use.

Potential legal and ethical difficulties

System developers and doctors may be concerned about the possible legal implications should a patient sue a doctor who had access to a decision-aid during a field trial. This is a complex topic [19], but in summary, both would probably be immune from negligence claims if they could show that:

1. The system had been carefully evaluated in laboratory studies
2. The system provided its user with explanations, well-calibrated probabilities, or the opportunity to participate in the decision-making process
3. No misleading claims had been made for the system
4. Any error was in the design or specification rather than in the coding or hardware
5. Users had been adequately trained and had not modified the system

A second concern is whether it is ethical to make a system available to doctors without the approval of the patients whose management it may influence. It seems wise to request the approval of an institutional review body before starting the trial, as well as the informed consent of patients whose doctors are participating.

Summary and conclusions

We have argued in favour of rigorous field trials of medical decision-aids, and discussed the many potential difficulties to be overcome. Despite these difficulties, there is evidence that medical decision-aids can have a beneficial impact on both the processes and outcome of medical care. However, if they are to become more widely accepted in clinical medicine, such field trials will need to become the rule, not the exception.

Acknowledgements

JCW is receiving a UK Medical Research Council Travelling Fellowship.

References

1. Shortliffe EH. Computer programs to support medical decisions. JAMA 1987; 258: 61-6.
2. Lundsgaarde HP. Evaluating medical expert systems. Soc Science in Med 1987; 24: 805-19
3. Gaschnig J, Klahr P, Pople H, Shortliffe E, Terry A. Evaluation of expert systems: issues and case studies; in Hayes-Roth F, Waterman DA & Lenat D (eds), Building expert systems, Wokingham, Bucks: Addison Wesley, 1983.
4. Miller PL. The evaluation of artificial intelligence systems in medicine. Comp. Meth. Prog. Biomedicine 1986; 22: 5-11.
5. Wyatt J, Spiegelhalter D. Evaluating medical expert systems. Medical Informatics 1990; 15: 205-217
6. Wyatt J. Improving clinical access to medical knowledge. The Lancet, 1991 (to appear)
7. de Bliek R, Friedman C, Blashke T, France C, Speedie S. Practitioner preferences for patient-specific advice from a therapy monitoring system. In: Greenes R (ed). Proc 12th SCAMC; IEEE Press, 1988: 225-8
8. Donabedian A. Evaluating the quality of medical care. Millbank Mem Quart 1966;44:166-206
9. Buck C, Donner A. The design of controlled experiments in the evaluation of non-therapeutic interventions. J Chron Dis 1982; 35: 531-8
10. Fenyo G. Routine use of a scoring system for decision-making in suspected acute appendicitis in adults. Acta Chir Scand 1987; 153: 545-551
11. Adams ID, Chan M, Clifford PC et al. Computer aided diagnosis of acute abdominal pain: a multicentre study; Brit Med J 1986; 293: 800-804.
12. Wyatt J. Lessons learned from the field trial of ACORN, an expert system to advise on chest pain. In: Barber B, Cao D, Qin D, eds. Proc. Sixth World Conference on Medical Informatics, Singapore. Amsterdam: North Holland 1989: 111-115
13. Roethligsburger FJ, Dickson WJ. Management and the worker. Harvard Univ. Press, Cambridge, Massachusetts, 1939
14. Spiegelhalter DJ. Evaluation of clinical decision aids, with an application to a system for dyspepsia. Statistics in Medicine 1983; 2: 207-216.
15. Pozen MW, d'Agostino RB, Selker HP Sytkowski PA, Hood WB. A predictive instrument to improve coronary care unit admission in acute ischaemic heart disease. N Engl J Med 1984; 310: 1273-8
16. Cornfield J. Randomisation by group: a formal analysis. Am J Epidem 1978; 108: 100-102
17. Murray GD, Murray LS, Barlow P et al. Assessing the performance and clinical impact of a computerised prognostic system in severe head injury. Stats in Med 1986: 5: 403-410
18. Hart A, Wyatt J. Evaluating black boxes as medical decision-aids: issues arising from a study of neural networks. Medical Informatics 1990; 15: 229-236
19. Brahams D. Wyatt J. Decision-aids and the law. Lancet 1989: 2: 632-4