ELSEVIER

International Journal of Medical Informatics 69 (2003) 251–259

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

When and how to evaluate health information systems?

Jeremy C. Wyatt [a,*], Sylvia M. Wyatt [b]

[a] *Knowledge Management Centre, University College London, Gower Street, London WC1E 6BT, UK*
[b] *Future Health Network, The NHS Confederation, London, UK*

Abstract

Aims: Evaluating large scale health information systems (HIS) such as hospital systems can be difficult. This article discusses the reasons we need to evaluate these systems and a range of appropriate methods to carry out evaluations. It is written in non-technical language to assist health policy makers and others commissioning or implementing such systems, with references and a web site containing information for those wishing more detail (http://www.ucl.ac.uk/kmc/evaluation/index.html). *Methods:* A variety of questions relevant to HIS and qualitative and quantitative methods ranging from simple before–after to controlled before–after and fully randomised designs, are discussed. A running example—evaluating the impact of an order communications system on lab requests—is used to illustrate the potential problems, and how they can be resolved. *Results:* The main types of biases affecting impact studies and methods to reduce them are described. The article then discusses some trade-offs between the low cost, easily conducted before–after study with its unreliable results versus the more complex, expensive but much more rigorous randomised trial. *Conclusions:* As would be expected, the correct methods to evaluate depend not on what technology is being evaluated—whether an information system or a drug—but on the questions the study is designed to answer, and how reliable the answers must be. Only those commissioning an evaluation study can decide these.

Keywords: Health information systems; Evaluation methods; Technology assistant

1. Introduction: why, when and how to evaluate?

This article discusses evaluation methods for large scale health information systems (HIS) such as hospital information systems, focusing on how to evaluate information systems for clinicians, since these pose a typical range of evaluation problems. Clinical information systems include order entry and reporting systems, electronic patient records, telemedicine and decision support tools for health professionals, patients and the public. Such systems are a recent and expensive addition to the many health applications of information and communication technology (ICT), and decision makers in organisations around the world are under pressure from vendors, clinicians and the public to install

* Corresponding author. Present address: National Institute for Clinical Excellance, MidCity Place, 71 High Holborn, London WC1V 6NA, UK. Tel.: +44-20-7067-5800.
E-mail address: jwyatt@nice.nhs.uk (J.C. Wyatt).

1386-5056/03/$ - see front matter
doi:10.1016/S1386-5056(02)00108-9

them. However, information about the real benefits and drawbacks of such technology, necessary to make informed purchasing decisions, is seldom available [1]. One reason is that ICT systems are often technology, rather than real health need, driven. Another is that few ICT developers or advocates see the need to independently evaluate these systems. This leaves those responsible for spending money—whether public or private—on clinical ICT exposed, for the following reasons:

(1) ICT has a high cost, e.g., around 50M€ spread over 10 years for a full electronic patient record system for an average UK hospital (personal communication, David Hancorn, Silicon Bridge Consultants).

(2) Preparing for and installing ICT can be very disruptive for staff and organisations, since it can be a profound agent for change—as it is often intended to be.

(3) Despite good preparation and training, staff may simply not bother to use the system even when we believe it would help them.

(4) The majority of HIS have not been developed using the rigorous software engineering methods used in other safety–critical environments such as the airline or railway signalling industries [2]. An example of the problems caused was a subtle but serious error in software calculating Down's syndrome test results, which caused much anxiety [3].

(5) Even when correctly coded, ICT may have negative effects on clinical practice and patient care, e.g. delaying the arrival of dangerously abnormal lab results [4] or exposing an organisation to the risk of industrial action—in extreme cases by doctors (e.g. [5]).

(6) Health care and the organisations delivering it are very complex, so the benefits of even well designed ICT are unpredictable, and cannot be anticipated from a simple description of the system's structure or functions. In one case, it took a computer crash to reveal the sad truth that clinicians found it quicker and simpler to ask patients rather than use the computerised record [6].

It is now accepted that it is very hard to get information systems right first time [7,8], and successive rounds of developing prototypes, 'formative' evaluation, feedback of results to the development team and revision are necessary right from the start. Even after evolving a system to match its niche in this way, however, someone will still argue that the money could have been spent better elsewhere—on training, other quality improvement activities or direct patient care. What is needed at the conclusion of any substantial ICT implementation process is a 'summative' evaluation, to answer the question 'What is the impact of the new ICT system on the problem it was intended to address?' [9]. Such an evaluation may also document generic lessons—such as reasons for success and failure—or inform and justify future ICT purchasing decisions.

In most countries there is a statutory duty on public officials to account for their use of public money. For example, NHS officials have had to explain themselves four times to the House of Commons Public Accounts Committee following successive National Audit Office reports on ICT in the NHS (e.g. [10]). Each time they found to their discomfort that they could not simply assume that ICT would bring the claimed health benefits. The opinion of the committee was that either the benefits and risks must already be documented, or a rigorous evaluation should be carried out.

1.1. Why is evaluating HIS difficult?

It is much easier to evaluate a discrete, localised ICT system than one which is diffuse and implemented according to a schedule that cannot be changed. This and other challenges posed by evaluating HIS are listed in Table 1, together with responses.

Table 1
Challenges posed by evaluating organisation-wide information systems, and possible responses to these

Challenge	Response
HIS are usually implemented as part of a complex change/re-engineering process	Carry out pragmatic, not explanatory studies—but do not claim that the HIS itself caused the benefits
HIS are used by many different professional groups	Seek evaluation questions from each group; answer as many as feasible
Associated with many different impacts	Weight impacts and sum to measure overall impact
Serve real patient care needs	Often implemented incrementally, which provides opportunity for designs such as the externally controlled before–after study (see text)
HIS are ubiquitous, we cannot disrupt live system	Exploit natural experiments—system maintenance, network outages (monitor helpline calls, workload)
It is usually impossible to implement HIS in all parts of an organisation simultaneously	Randomise wards, departments etc. to early or late implementation and make measurements when half are implemented (see text)
Carry over or contamination of HIS benefits from staff training sessions, cross cover, rotations...	Time evaluations to coincide with new junior staff arrival
HIS are multi-functional: admission, discharge, transfer log; order communications, reminders...	Isolate each important function for detailed study
HIS are usually tailored to each organisation	Design evaluation studies to answer the question '*Did this system help here*?' since it is harder to answer the question '*Can HIS help in general*?'

So, evaluation of HIS is necessary, and, despite the considerable challenges, it is quite possible as long as one is realistic about the questions to be addressed.

To make this more concrete, imagine the following scenario. As chief executive of St Elsewhere's hospital, your laboratory director reminds you of the ever-increasing number and cost of blood tests ordered, and unnecessary days patients spend in hospital while junior doctors investigate all the spuriously abnormal results. She has already modified request forms to force doctors to write in the names of tests rather than simply ticking off a menu [11], but test numbers continue to rise. She suggests purchasing a 3M€ laboratory order communications system which, by issuing reminders to doctors as they request inappropriate tests, should save money and also lead to less unnecessary follow-up testing [12]. You agree, but only in the context of a careful evaluation to justify this expenditure and learn more about its benefits and risks.

2. What to measure, and how?

The first issue in any evaluation is, what are the key questions? You need to cover all relevant perspectives—here, the organisation, the staff and the patients—and address a range of questions and concerns [9]. These questions, and in turn measurements and associated methods, usually fall out in discussion with the relevant stake holders, who can be encouraged to formulate their concerns as questions that can be answered. Typical questions for the above scenario are shown in Table 2.

In every evaluation, questions proliferate while the resources needed to answer them are fixed. As chief executive, you must balance the priorities of the various interest groups and decide which few questions will help you to judge if the system is a success. You decide that the actual number of tests ordered per patient is the key issue (influencing expenditure, patient waiting time etc.) but that other

Table 2
Possible evaluation questions about an order communication system, from four perspectives

Perspective	Questions
Organisation	Does the system save money? Does the system reduce risk exposure? Does the system impact patient satisfaction?
Clinical staff	Is the system easy to use? Is the system quicker than paper forms? Is it clinically safe? Does the system impact clinical freedom?
Laboratory staff	Does the system encourage more appropriate blood test orders? Does the system halt the rise in orders?
Patients	Do I wait less time for tests and results? Does my care suffer as a result?

concerns will also be addressed if resources allow.

The second issue in an evaluation is, how to make the measurements? Two broad categories of methods are used to make measurements:

Objective or quantitative methods: (chapters 4–6 in Ref. [9]), used to collect objective data such as patient waiting times, the number of lab tests ordered per patient, how many patients are seen per clinic room hour or staff satisfaction on 1–5 scale.

Subjective or qualitative methods: (chapters 8–9 in Ref. [9]), such as interviewing, focus groups or participant observation. These are used to explore and describe staff or patient motivations, hopes and fears or to document stories about what worked or did not, and why. These methods can be combined with document analysis techniques to pick out themes emerging in the minutes of meetings, complaint letters written by patients—or even from lawyers!

The goal of qualitative methods is to create a rich, compelling description, so subjectivity is expected (see articles by Marc Berg and Joan Ash, in this issue). With numerical or quantitative measures, however, the evaluator needs to make sure that each of the methods is reliable (obtains the same result whoever is using it) and valid (measures what you intend it to measure, not something else). This is not easy, and it often takes ingenuity to devise a reliable, valid measurement method that is also feasible, given limited resources. Sometimes you are lucky, and routinely collected data can be employed. Here, the number of tests ordered per patient can be calculated from routine laboratory figures. Often, the best strategy is to find measurement methods such as survey questionnaires that others have developed and tested—see chapter 5 in Ref. [9] for a list and the article by Friedman et al. in this issue. If you do not use validated measurement methods, be warned that any improvement or worsening due to the new system may be hidden by random variation due to poor measurement. Another risk is using an insensitive measure, such as total laboratory expenditure including staff salaries and other fixed costs, that would probably fail to reflect real changes taking place.

Once you have chosen your portfolio of measurements and appropriate methods to make them, you are ready to select your study design. Here is another choice: will a simple before–after study, with all its biases, suffice, or should you anticipate the inevitable criticisms and invest in a more complex study design? If so, do you need the rigour of a randomised trial, or can you compromise on a controlled before–after design?

3. Choosing a study design

Each study design has its pros and cons, some of which are shown in Table 3. Note that the minimum detectable effect size and the overall study cost depend markedly on the number of cases and centres studied, so they

Table 3
Characteristics of three major types of evaluation study design

	Simple before–after	Controlled before–after	Randomised trial
Typical use	Local audit	Regional decisions	National policy setting
Study role	To describe what happened	To suggest the cause	To determine cause and size of the effect
Approximate minimum detectable effect	Large (> 50% change)	Medium (> 30% change)	Small (> 10% change)
Chance of bias	Very high	Medium	Low if well designed
Scale	Within a single organisation	Within 2–5 organisations	The more organisations the better
Estimated lowest cost	Low: < €5 k	Low/medium: €5 k–€20 k	Medium/high: €30 k +

are very broad estimates. However, the ranking of the three study designs along both these dimensions is robust.

It can be seen that the choice of study design depends on the type of question and how reliable the answer needs to be. If you are just curious about what happened, a simple before–after study is enough, but if you are responsible for HIS policy in a group of hospitals, or even in a whole country, you will need a reliable estimate of the size of benefit attributable to various kinds of HIS. This needs evidence from a more rigorous study, such as a randomised trial.

The next three sections discuss the pros and cons of these designs in more detail.

3.1. The simple before–after design, and its defects

For the senior manager, a before–after evaluation in their own organisation reveals a great deal. The 'before' or baseline measurement documents the nature and frequency of the problem and its consequences, helping both to specify the new system and acting as a baseline for later comparison. Measurements made after the new system is implemented, staff are trained and the system is in routine use tell them how much the ICT 'solution' has alleviated the problem, helping them justify the expenditure and judge the likely value of future ICT expenditure.

Let us say that in the scenario described, the before–after study results are a disaster, showing that the number of tests ordered per patient increased by 35% after installing the order communications system. Your helpful lab director suspects that the poor 'after' results are due to a new group of insecure junior doctors or maybe an epidemic of infectious disease, so you immediately launch an investigation to confirm this and promote her to Director of Quality Improvement.

However, now imagine what you would have done had the results been positive, showing a *reduction* in the number of tests by 35%. Everyone would be happy, and no one would seek to discover whether the 'after' results actually appeared to be reduced compared with baseline figures that were temporarily inflated. Three months later you discover by accident that test ordering was indeed inflated a few months ago, due to a new group of insecure junior doctors or maybe an epidemic of infectious disease, just at the time when your lab director had asked for the new OCS. You therefore sack her.

The unfairness here is striking: when studies generate the result we want, we take them at face value, and fail to register even major defects in study design. However, when they

go against us, we launch investigations to find out why, and to uncover all possible biases. Surely the rational approach is to start out with a more rigorous, bias free study design, so that whatever the result, we are satisfied with it and not left with lingering doubts about the real cause of the findings?

With a simple before–after study, even if the number of tests ordered after system installation is genuinely less than the number before, we still cannot be sure that the ICT itself was responsible. The improvement may be due to training increasing staff awareness about excessive lab orders, or business process re-engineering associated with the introduction of the system, rather than to the technology itself. It would be hard to motivate staff to participate in these activities without the promise of an order communications system, but it is still misleading to conclude that the system *itself* brings the benefit. Introducing the system in another hospital without staff training and process redesign probably would not work.

Perhaps most worrying, though, is that the reduction in test ordering may actually have been a pure co-incidence, due to some 'non-specific' factor like a local initiative to reduce all test orders. A chief exec may take the pragmatic view that, if the problem is solved, it does not matter now. But can that chief exec honestly recommend this expensive ICT system to a colleague in another organisation, with a different set of local initiatives? What is missing here is reliable attribution of cause and effect. All we know is that we had a problem, we installed ICT, and now it is gone. This leaves open the question, did the ICT *cause* the problem to disappear, or was it something else—a bias? Evaluators have two tactics to resolve this: the controlled before–after design and the randomised trial.

3.2. The controlled before–after design

To increase our confidence that the ICT caused the observed change, the first step is to add an *external control* to the before–after design. To do this, the evaluator needs to locate a similar local organisation, subject to the same initiatives, case mix etc., but which will not install the ICT—an order communication system, in this case. Finding a suitable external control site may pose serious difficulty in itself, especially since the other hospital needs to collect data twice on your behalf, coinciding with the baseline and post implementation data collections at the site where the ICT is installed. If there is no change in the external site but there is a change in your site, that does suggest that the ICT was responsible. But maybe the organisation you chose as an external control was not as similar to yours as you thought, is not subject to the same non-specific factors, or ignored the local initiative on reducing test orders. What is really needed is a second, *internal control*, within your own hospital.

Adding an internal control considerably strengthens the before–after study. It means that you measure not only the before–after difference in the problem you targeted (the number of blood tests), but also the before–after difference in something else subject to the same non-specific factors but which you do not expect the system to influence. In our scenario, this could be the number of tests ordered on histopathology or bacteriology specimens, as long as they are not yet included in the order communications system. If the number of blood test orders falls but the number of bacteriology and histopathology test orders continues to rise, this is strongly suggestive that the order communication system is responsible for the changes and not non-specific factors affecting all lab orders. If this happens while blood test orders continue

to rise in a local control hospital—see the figure for idealised results—it is hard to come up with a plausible explanation other than the order communications system.

However, real world studies rarely provide such neat and tidy results as those shown in Fig. 1, especially when you are looking for only small changes, such as a reduction of 20% in orders, which would still be worthwhile. If you are looking for a definite result, a very reliable result or for small differences, then you should carefully consider the most rigorous, bias-free design: the randomised trial.

3.3. The randomised controlled trial

The randomised controlled trial is the established method for testing new drugs, surgery and other procedures, and is being used increasingly in other sectors such as education, social care and criminal justice [13]. It is the only method that allows us to estimate reliably the size of small but worthwhile benefits attributable to an intervention of any kind, and also allows us to estimate the frequency and severity of side effects. You can even combine qualitative and quantitative evaluation methods in such a design, as explained later.

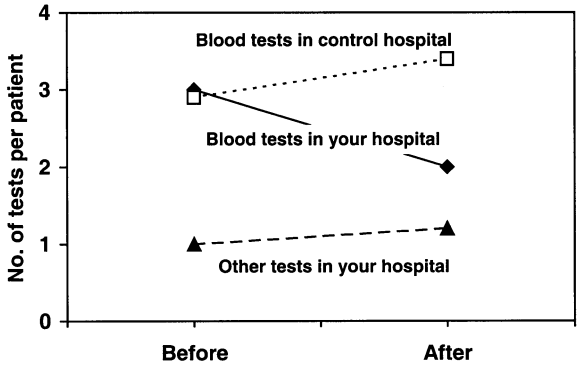

Fig. 1. Idealised results of a before–after study on an order communications system with external and internal controls.

To detect such small but useful benefits means making measurements on many individuals to average out individual differences. In the case of clinical information systems, you often need to measure the effects on patients or clinicians. Carrying out a simple before–after study in one patient or doctor, or even a handful, will not allow us to conclude much except that the benefit of the ICT varies. To get a better estimate of the likely overall benefit of the ICT we need to assemble a larger group of individuals and average the results obtained. How many individuals are needed depends on three factors:

1) How much the measurement (e.g. number of tests ordered) varies between individuals (often assessed by the standard deviation)
2) The minimum benefit needed, taking account of cost and disruption, that would reassure us that the ICT is useful
3) How accurately we need to estimate this benefit, in terms of statistical significance (usually fixed at $P = 0.05$) and the power of the study to defect it (usually 0.8)

Given this information, a statistician can carry out a precise sample size calculation. Let us say you need 200 patients to check if your new order communication system reduces the number of tests ordered from an average of 4 to 3 per patient. You could check the number of tests ordered in 100 patients before the system is running and 100 after. However, this before–after study has all the opportunities for bias discussed above. Somehow we need to ensure that the two groups of patients we study (those managed with and without the new system) are as identical as possible in all characteristics, known and unknown, that might influence the number of tests ordered.

Otherwise, we will always have a lingering doubt about the real cause of the observed changes, and cannot reliably attribute the small reduction in number of tests to the information system.

Fortunately, there is a simple way to ensure that the two groups of patients are comparable in all respects: we randomise them. This means taking every one of 200 consecutive patients as they arrive and using a random number table (e.g. Appendix in [14]) to allocate each to one of two groups: use of the system allowed, or not allowed. We then have two groups of patients who differ only because they were randomly allocated, not because, say, the second group arrived during a flu epidemic, with fewer staff on duty. Any measurements we obtain from these patients will then reflect only the effects of the order communications system, and give us a bias-free estimate of the improvement it causes. Usually, clients or patients are randomised, but sometimes it is necessary to randomise health care workers, hospital departments, general practices or even whole hospitals [15,16].

Randomised trials of different kinds of clinical information systems uncover different biases, each of which demand specific measures to counter them [17–21]. Computer-based tools such as Design-a-Trial (see http://www.ucl.ac.uk/kmc/DaT/index.html for information) can help evaluators to write the necessary trial protocol and calculate sample sizes. Trials have an unfair reputation for being expensive but can actually be carried out quite economically, especially if much of the data needed is already collected routinely. For example, a recent trial of educational visits in 25 hospitals studied 4 recorded clinical practices in 4500 pregnant women but cost just 40 000€ thanks to careful planning and study design [15].

4. Summary and conclusions

There is increasing pressure on health services, care organisations and health policy makers to not only spend money wisely, but also to demonstrate it was wisely spent. Nowhere is this more acute than with clinical information systems, which have enormous potential to improve health [22], but also much potential to disrupt it. Perhaps worse, although usually viewed as part of the modernisation process, information systems are often left untouched for years after installation, leading to technology lock-in and loss of flexibility [1]. Even if a system is effective when installed, it may rapidly lose its edge as the health system around it changes, making repeated evaluation necessary, to take account of the changing health context.

While the simple before–after design will be the mainstay of many local evaluations, using external or internal controls (preferably both) can help compensate for its significant defects. The results will never be as convincing as a randomised trial, but will often be sufficient for local use. Such studies can often be written up for wider dissemination, as long as the potential biases are explored.

However, when a health policy maker wishes to make large scale recommendations, they should first seek sound evidence. Sound evidence requires at least one randomised trial and preferably a systematic review combining the results of all relevant trials [23]. For an example review which identified and combined the results of 66 RCTs of the impact of decision support systems on clinical actions such as test ordering and patient outcomes, see Hunt et al. [24]. Although trials and systematic reviews cost more than simpler studies and take months to conduct, they consume only a tiny fraction of the costs of implementing such technology in one health institution. Incorrectly assuming that a clin-

ical information system—or any other health technology—will always help, when it sometimes makes matters worse could be much more expensive. Further advice and articles discussing these issues can be found on the UCL Knowledge Management Centre evaluation web site (http://www.ucl.ac.uk/kmc/evaluation/index.html).

References

[1] J. Keen, J. Wyatt, Back to basics on NHS networking, BMJ 321 (2000) 875–878.

[2] M.F. Smith, Are clinical information systems safe?, BMJ 308 (1994) 612.

[3] P. Wilkinson, Down's test leaves 150 women in abortion fear. The Times 31 May (2000) 1, 3.

[4] E.S. Kilpatrick, S. Holding, Use of computer terminals on wards to access emergency results: a retrospective audit, BMJ 322 (2001) 1101–1103.

[5] L. Sears-Williams, Microchips vs. stethoscopes: Calgary hospital MDs face off over controversial computer system, Can. Med. Assoc. J. 147 (10) (1992) 1534–1547.

[6] I. Krakau, C. Fabian, Who needs all that information collected in computerised medical records? A computer crash shows that to ask the patient is often simpler and quicker, Lakartidningen 96 (1999) 4032–4034 (in Swedish).

[7] M.F. Smith, Prototypically topical: software prototyping and delivery of health care information systems, Br. J. Healthcare Computing 10 (6) (1993) 25–27.

[8] J.C. Wyatt, Clinical data systems, Part III: developing and evaluating clinical data systems, Lancet 344 (1994) 1682–1688.

[9] C. Friedman, J. Wyatt, Evaluation methods in medical informatics. New York: Springer 1997 (ISBN 0-387-94228-9), reprinted 1998 (review: Faughnan J. BMJ 1997; 315: 689. Available from http://www.bmj.com/cgi/content/full/315/7109/689).

[10] Comptroller and auditor general, The NHS Executive Hospital Information Support Systems Initiative, HMSO, National Audit Office, London, 1996.

[11] I. Durand-Zaleski, J.C. Rymer, F. Roudot-Thoraval, J. Revuz, J. Rosa, Reducing unnecessary laboratory use with new test request form: example of tumour markers, Lancet 342 (1993) 150–153.

[12] J.C. Wyatt, Knowledge for the clinician 9. Decision support systems, J. Roy. Soc. Med. 93 (2000) 629–633.

[13] A. Oakley, Experimentation and social interventions: a forgotten but important history, BMJ 317 (1998) 1239–1242.

[14] D. Altman, Practical Statistics for Medical Research, Chapman and Hall, London, 1991.

[15] J. Wyatt, S. Paterson-Brown, R. Johanson, D.G. Altman, M. Bradbum, N. Fisk, Trial of outreach visits to enhance use of systematic reviews in 25 obstetric units, BMJ 317 (1998) 1041–1046.

[16] K. Herbst, P. Littlejohns, J. Rawlinson, M. Collinson, J.C. Wyatt, Evaluating computerised health information systems: hardware, software and human ware. Experiences from the Northern Province, South Africa, J. Public Health Med. 21 (1999) 305–310.

[17] J. Wyatt, Evaluating electronic consumer health material, BMJ 320 (2000) 159–160 (commentary).

[18] A.D. Randolph, R.B. Haynes, J.C. Wyatt, D.J. Cook, G.H. Guyatt, Users' guides to the medical literature: XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system, JAMA 282 (1999) 67–74.

[19] J.C. Wyatt, Measuring quality and impact of the World Wide Web, BMJ 314 (1997) 1879–1881.

[20] J. Wyatt, Evaluation of clinical information systems, in: J.H. van Bemmel (Ed.), Handbook of Medical Informatics (Chapter 30), Springer, 1997, pp. 463–469.

[21] J.C. Wyatt, Telemedicine trials: clinical pull or technology push?, BMJ 313 (1996) 1380–1381 (commentary).

[22] J.C. Wyatt, Possible implications of information and communications technologies for NHS buildings in 2020. In: R. Glanville, S. Francis (Eds.), Health care building for tomorrow: developing a 2020 vision (ISBN 1 902089 242 1), Nuffield Trust, London, 2000, pp. 21–24.

[23] M. Egger, G. Davey Smith, D.G. Altman (Eds.), Systematic Reviews in Health Care, second ed., BMJ Books, London, 2001.

[24] D.L. Hunt, R.B. Haynes, S.E. Hanna, K. Smith, Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review, JAMA 280 (1998) 1339–1346.