Journal of Evaluation in Clinical Practice, **7**, 4, 373–385

Evaluation of clinical information systems. What can be evaluated and what cannot?

Thomas Bürkle,[1] Elske Ammenwerth,[2] Hans-Ulrich Prokosch[1] and Joachim Dudeck[3]
[1]Institute of Medical Informatics and Biometry, University of Münster, Germany
[2]Institute of Medical Biometry and Informatics, University of Heidelberg, Germany
[3]Institute of Medical Informatics, University of Gießen, Germany

Correspondence
Dr Thomas Bürkle
Institut für Medizinische Informatik und Biomathematik
Westfälische Wilhelms-Universität
Domagkstraβe 9
48129 Münster
Germany

Keywords: clinical systems, evaluation, methodology, outcomes, software development

Accepted for publication:
4 January 2001

Abstract

The evaluation of clinical information systems is essential as they are increasingly used in clinical routine and may even influence patient outcome on the basis of reminder functions and decision support. Therefore we try to answer three questions in this paper: what to evaluate; how to evaluate; how to interpret the results. Those key questions lead to the discussion of goals, methods and results of evaluation studies in a common context. We will compare the objectivist and the subjectivist evaluation approach and illustrate the evaluation process itself in some detail, discussing different phases of software development and potential evaluation techniques in each phase. We use four different practical examples of evaluation studies that were conducted in various settings to demonstrate how defined evaluation goals may be achieved with a limited amount of resources. This also illustrates advantages, limitations and costs of the different evaluation methods and techniques that may be used when evaluating clinical information systems.

Introduction

Clinical information systems support medical and nursing staff in their daily work by means of electronic data processing. They cover local systems, and departmental subsystems as well as hospital communication systems and hospital information systems in both inpatient and ambulatory care. The assessment of clinical information systems receives increased attention as more and more are used in clinical routine. This paper focuses on three questions for the evaluation of clinical information systems:

- what to evaluate;
- how to evaluate; and
- how to interpret the results.

'*Clinicians would be unwise to use any system unless it has been shown to be safe and effective*' (Van Bemmel & Musen 1997).

Evaluation is based on comparison. We may either compare the status after system introduction with the status before (or with the previous system) or we may agree on expected system effects and assess whether those effects have been established. Peterson & Gerdin Jelger (1988) describe three purposes of the evaluation process:

- To compare the results with the goals and expected effects of the system, for example concerning working conditions, service to the patients, integrity and security of data and finances (summative evaluation).
- To direct work towards the expected result with the help of formative evaluation during the development and the introduction of the system.
- To use the findings and outcomes of the evaluation process as an experience base for the next project.

Many different aspects of evaluation can be distinguished, such as the influence on the working environment of medical staff, the savings which may be achieved or even the effect a system may have on patient outcome. Evaluating clinical information systems remains an art (Heathfield *et al*. 1997).

Generally, two approaches to the evaluation of clinical information systems exist, an objectivist approach and a subjectivist approach (Friedman & Wyatt 1997; Van Bemmel & Musen 1997). The objectivist approach assumes an agreement on important system attributes that can be measured and interpreted. Here the preferred study design ('gold standard') would be a randomized controlled trial (RCT) (Tierney *et al*. 1994). Instead the subjectivist approach assumes that observation results are dependent on context and observer and that different individuals or groups may hold a different opinion about a systems value. In this approach qualitative data is preferred to quantitative results of controlled trials.

The objectivist approach is widely accepted in biomedicine and also scored positively in the evaluation of knowledge based systems (KBS). KBS may issue reminders or offer therapy advice based upon data of an individual patient and stored machine-readable medical knowledge. They may issue reminders for necessary diagnostic or therapeutic actions (McDonald 1976), they can check pathogen susceptibility to a given antibiotic (Evans *et al*. 1985; Evans *et al*. 1993), check for adverse drug reactions (Evans *et al*. 1995) or interpret laboratory values (Rind *et al*. 1992). Here, controlled trials (Johnston *et al*. 1994; Adlassnig & Horak 1995) have been conducted with good success and could prove the influence of KBS on process and occasionally on outcome quality. Today, several systematic reviews of evaluation studies exist (Johnston *et al*. 1994; Balas *et al*. 1996) and a stage has been reached where occasionally systems are directly compared with each other (Berner *et al*. 1994). There are several reasons why KBS can be more easily evaluated in a controlled trial than clinical information systems. Often such functions can be switched on and off in the background, thus allowing for blinded studies. Randomization is comparatively easy when single reminders are randomly either displayed or not. Success may be measured directly by checking if the clinician has started the activity that was recommended from the KBS.

However, the objectivist approach has seldom produced positive evaluation results when applied to complex clinical information systems (e.g. Van Gennip *et al*. 1994; Bürkle *et al*. 1999). Often it was found that evaluation of such systems consumed large amounts of resources and that the expected and measured objective parameters, such as time saving, did not exceed those of the control. Randomization is often difficult because one must randomize whole wards or departments of a hospital who do or do not receive the information system. As this is basically possible, the number of items (wards, etc.) needed to achieve sufficient power is often not available. In addition, it is often difficult to identify a comparable 'before' and 'after' situation with a steadily progressing technology. As Forsythe and Buchanan (1992) state: 'Real-world settings are not easily controlled'. Blinding is usually impossible as the information system is clearly visible for members of the intervention group. To define appropriate, objective parameters to measure is a demanding task. Very often the influence of such a system will not manifest in direct time savings, but rather in improved cooperation between departments, in increased quality of documentation, or in better patient care. No clear indicators exist to measure such effects. Heathfield *et al*. (1997) discuss some of these problems more extensively. We contend that clinical information systems evaluation based on RCTs seems to be difficult. Koch and Abel (1997) describe similar problems for the evaluation of surgical procedures and cite a study which indicates that over the years the percentage of RCTs in surgery could not be increased.

In this paper we will emphasize the necessity to evaluate clinical information systems, even under adverse conditions. We will present different methods of evaluation beside the RCT and demonstrate examples for the use of those methods in clinical settings. We will discuss strong points and weaknesses of each method. Before any evaluation is started, agreement on the evaluation goals is necessary. A systematic presentation of evaluation methods is illustrated using a set of practical examples. The examples are condensed for discussion.

Goals of an evaluation

For the evaluation of a clinical information system, the first question must be: What do we want to evaluate?

An agreement on the evaluation goals is required not only for the study protocol, but also for the selection of suitable evaluation parameters. Defining appropriate evaluation goals may be the first serious problem for an evaluation study (Bürkle *et al*. 1995b). If highly generic evaluation goals, such as the effectiveness of an information system (Hinson *et al*. 1994), the effects on medical or nursing activity (Spranzo Keller *et al*. 1992), or the effect on patient care (Petrucci *et al*. 1992), are mentioned, one has to check carefully whether the measured indicators allow a clear statement to be made for such a generic goal. At the end the researchers may answer a totally different question to the one they defined as their goal. Therefore, goals of an evaluation study should be well described and rather detailed so that the evaluation results may be easily compared with the goals. Preferably, the goals should have a close connection to the evaluated system, in order to assure that positive evaluation results may be definitely attributed to the system. A goal such as effect on patient care may fail this test because many other factors will influence this goal.

Methods of evaluation

Methods for the evaluation of clinical information systems are presented to answer the question: How do we evaluate?

Systematically, we may distinguish several phases of evaluation of software systems. Indeed, evaluation starts during program development and can be split into verification, validation, assessment of human factors and clinical assessment of clinical effect (Engelbrecht *et al*. 1995; Ohmann & Belenky 1995; Ohmann *et al*. 1998; see also Van Bemmel & Musen 1997).

Verification is an evaluation process that should be implemented during system design and development to answer the question 'Did we build the system correctly?'. Verification checks whether the system has been developed according to its specification and confirms consistency, completeness and correctness of the system. The methodology for verification is either a program proof (Schmitz *et al*. 1982) or a test strategy (Myers 1976, 1982). The program proof confirms total correctness of the program logic with mathematical methods, the test strategy confirms partial correctness of the program with given test cases.

Validation is performed later to answer the question 'Did we build the right system?'. Validation checks that the system performs the tasks for which it has been designed in the real working environment. Test strategies are the methodology of choice. Content validity compares program results with the expected results (for example a gold standard). Empirical validity checks whether the results of content validation remain stable when the system is under full workload. The system test examines the complete system in its working environment. Validation is especially important for KBS (e.g. Verdaguer *et al*. 1992).

Evaluation of human factors is the next phase of system evaluation. It answers the question: 'Will the system be accepted and used?'. Even if a system has been verified and validated, it may be so awkwardly designed that it cannot be used in real life, because using the system is either too cumbersome or consumes too much time. Imagine a car which has its driver seat in the back. It will drive correctly (verification) and it can be used on a normal road (validation). Nevertheless people will find it awkward to look over the heads of their passengers on the road. To determine usefulness and usability, one must select appropriate quality indicators. Usefulness of a system is often measured by examining user satisfaction. User satisfaction has system-dependent aspects, such as content satisfaction, interface satisfaction and organisation satisfaction, but also system-independent personal aspects such as individual dislike for computers (Ohmann *et al*. 1997a).

Observation of the system and its users in their working environment and asking the users is the appropriate methodology for assessment of usefulness and usability. We can distinguish observation studies, log studies, reaction studies and interviews or questionnaires. In an observation study, an external observer will record critique, comments and recom-

mendations of the user. In a log study one might want to check how often a certain part of the program has been used. In a reaction study, comments regarding the program are recorded directly, e.g. in an additional input window attached to the program. In a questionnaire study, the user is asked specific questions that serve as an indicator for usability. Questions may be either specific for the examined system or generic to assess the attitude of the users. To assess the latter, one can use evaluated questionnaire tools (Ohmann *et al*. 1997b). A clear distinction between validation and human factor assessment is sometimes difficult. Both are tested in a realistic environment. However, validation focuses on the system, and human assessment focuses on the user.

Evaluation of the clinical effect is the last phase of system evaluation. The appropriate question would be 'Which clinical effect has the system?'. We may drill down even further to the question 'How does the system affect patient outcome?'. If we want to derive causal conclusions from a certain therapy or intervention we should remember Donabedian (1982), who distinguishes three quality factors in patient care, namely structural quality, process quality and outcome quality. For each of those quality factors we may chose appropriate quality indicators such as financial resources, personnel resources or physical resources in the first case, gold standards and guidelines for correct behaviour in the second case and length of stay, morbidity or mortality in the last case to measure effects of our intervention upon the chosen quality factor. Careful consideration of facts dictates that a significant effect on one quality factor is only assumed when several indicators react positively.

The clinical effect is best measured in a field study using an RCT. The RCT attempts to eliminate or neutralize unwanted influence factors by randomization between control and intervention. However, as mentioned above, RCTs might be unsuitable for evaluation of clinical information systems. In such cases other methods, such as one-group studies, two-group studies measured after intervention (control group and intervention), two-group studies measured before and after intervention or cross-over studies, may be useful as well, although their power and sensitivity is not as good as that of RCTs (Van der Loo *et al*. 1994).

Examples for the evaluation of clinical information systems

In the next section we present a set of examples for the evaluation of clinical information systems that have either been published or conducted by the authors themselves. We will describe each study, present the obtained results and discuss power and disadvantages of the used methodology.

First example: evaluation of user satisfaction

(A) Study design and environment For the evaluation of user satisfaction we cite a study which was published in 1997 by Ohmann and colleagues from Düsseldorf University Hospital, Germany (Ohmann *et al*. 1997a, 1997b; Boy & Ohmann 1999). In this study medical doctors' satisfaction with a documentation module of a commercial hospital information system was evaluated.

A questionnaire study was chosen. The observation objects (164 hospital physicians) were selected randomly. For the evaluation of user satisfaction two standardized instruments were employed: the User Information Satisfaction tool (UIS) with 13 items (Ives 1983) and the Questionnaire of User Information Satisfaction tool (QUIS) with 27 items (Chin 1988). Both tools have been evaluated in a pilot phase, resulting in reliability values of 0.77 for the UIS and 0.90 for the QUIS. The tools have a slightly different focus: the UIS covers quality of information with five items, available data processing staff and service with five items, and knowledge and involvement of user with three items. Its focus is therefore on the contents of the system and the organization. Each item is rated on a scale from –3 to + 3. The QUIS covers overall reaction with six items, screen layout with four items, terminology and system information with six items, system use learning with six items and system capabilities with five items. It focuses on the user interface. Each item is rated on a scale from 1 to 9.

(B) Measured results In this study the basic medical documentation module (MCW) of the commercial SMS-Dataplan HIS used at Düsseldorf University Hospital was evaluated. The study suffers from a low response rate (58%) and only 24% of the question-

Table 1 Overview of examples

	Goals of evaluation	*Type of evaluation*	*Study design*	*Data collection method*	*Effect measure*
Ohmann *et al.* 1997	Personal attitude concerning a basic medical documentation module	Human factors	Non-controlled study (one group after measurement)	Questionnaire	User satisfaction
Bürkle *et al.* 1994	Effects of a hospital information system and a nursing information system	Clinical effects, human factors	Controlled study (one group pre and after measurement)	Work sampling, questionnaire	Time distribution change, user satisfaction
Ammenwerth *et al.* 2000	Effects of a computer-based nursing documentation system	Clinical effects, human factors	Randomized controlled study (two groups pre and after measurement)	Time measurement, questionnaire, interview, chart review	Time savings, documentation quality, user satisfaction
Ammenwerth *et al.* 1999	Application scenarios and user acceptance of mobile information processing tools	Clinical effects, human factors	Simulation study (one group after measurement)	Observation, questionnaire, interviews	User opinion

naires could be evaluated; the other physicians hadn't used the system yet. Therefore, the results, although interesting, remain questionable. On a scale between –3 and +3, users rated quality of information (0.37) and support by data processing staff (0.51) as positive, and involvement of users as neutral (0.03), giving a total of 0.34 for the User Information Satisfaction tool (UIS). On a scale between 1 and 9, users rated the system above average (5.06) on the Questionnaire of User Information Satisfaction tool (QUIS). Junior doctors had a better opinion about electronic data processing than senior medical staff.

(C) Power of used methodology This study attempts to measure the effect of a clinical information system on user satisfaction with quantitative tools. It is characterized by the use of proven and evaluated questionnaire tools to study the attitude of clinical staff towards the information system, support by data processing staff and overall user satisfaction. The use of such standardized tools should allow comparability with international literature, which, however, is not performed in the accessible publications. The study could be conducted with a reasonable evaluation effort (translation and redesign of questionnaires, randomization of medical staff, sending out questionnaires, collecting and analysing questionnaires). In principle, studies like this one can quickly deliver results.

(D) Problems of used methodology The intervention (installation of the clinical information system) was not blinded. It is not clear from the published papers if the users who did not have the system were used as a control group. The distribution of questionnaires among intervention group and control group was not quoted as randomized. Environmental factors have not been controlled. The measurement using standardized tools was objective, but the study suffers from a low return rate. The published literature does not give hints as to which parts or functions of the evaluated HIS module were found useful.

(E) Assessment This study is a typical example of a simple straightforward approach to measure the impact of a clinical information system. In its present

status such a study at its best can only deliver a rough positive or negative opinion about the system. It will not show which details work well and which do not. Therefore other authors (e.g. Kushniruk *et al*. 1997) have proposed much more complex observation studies for a subjectivist approach to find out in detail what works well and what does not. Those complex designs avoid also the principal problem of questionnaire studies being of limited validity if the return rate is too low.

Second example: evaluation of structure and process quality and user satisfaction

(A) Study design and environment The second study was performed by one of the authors at Giessen University Hospital, Germany. In a stepwise approach the impact of a hospital information system and a nursing information system on nurses working environment was examined (Bürkle *et al*. 1994; Bürkle *et al*. 1995a; Bürkle *et al*. 1999). A prospective intervention study with before and after measurements was performed on two medical wards, using control parameters to assure similarity of workload during control and intervention periods. The study design was incremental, with two consecutive interventions. The study was predominantly objective, using time motion observation of nurses (work sampling). Repeated non-standardized questionnaires, with up to 50 items, were used. Work sampling is a statistical method to measure time distribution on certain tasks (Tippett 1935; Sittig 1993). A person or area is observed at fixed time intervals and a note is taken of which task is performed in the moment of observation. For preparation of a work sampling study an analysis of performed work is required. The authors analysed organization structure, forms, data flow, work flow, communication and weakness, to derive 76 detailed work packages (Bürkle *et al*. 1994). For the sampling period this was condensed to 23 work packages in three nursing categories. The questionnaires tool was developed in-house and boasted an incremental design. For the historic control a question might ask if a specific computer function was considered necessary. Later, when the function was introduced, the user was asked if he had used the function.

(B) Measured results The time motion study measured the distribution of nurses work time into 23 work packages. Grouped results indicated that nurses spent 15% of their time on general nursing care (feeding, washing, etc.), 35% on specific care (medication, catheters, dressing change, etc.) and 30% on administrative activities (charting, appointments, etc.). Twenty per cent remained for non-classified tasks. The measured nursing workload control parameters (staff count, admission and discharge count, required nursing care minutes) did not affect time distribution between those work areas. The same was found for the two intervention steps. Neither the hospital information system nor the nursing information system affected time distribution beyond statistical variation. Time spent at the computer was between 1.5% and 2.8% of the total. As a side-effect, time distribution was, however, significantly affected from organizational differences between the two wards.

The questionnaires confirmed a positive attitude towards clinical data processing. Several basic HIS program functions scored high in acceptance and matched the expressed need for introduction. Other previously desired program functions were not used as often as anticipated. Organization of nursing work was beyond optimum, but the two examined systems did not cure this problem. The concluding results indicated a positive influence of the hospital information system on nurses' working environment, while the nursing information system was not well accepted.

(C) Power of used methodology This study presents a mixed, predominantly objective evaluation approach. Compared with other studies that measure only the time used for manual or computer-based performance of single activities, the study focuses on the total work time and its distribution upon activities. Thus it could in principle detect if time savings led, for example, to reinforced nursing activities. The study is 'partially controlled'. Workload might change during the different control and intervention periods and influence time distribution. Thus workload is controlled by measuring several workload indicators that in case of significant difference would have been used to derive a weighting factor.

Compared with other published studies, for the pre-intervention measurement a 'clean' situation

with no computers was measured. The methods mix delivers a large amount of detailed results regarding work time distribution and the usability of specific HIS functions. Several of the measured parameters, such as time distribution, can and have been compared with international literature.

(D) Problems of used methodology The intervention *per se* was not randomized. Only the observation objects (nurses) have been randomly selected each day, thus eliminating specific influence of certain nursing roles on time distribution. For the second intervention (the nursing information system) we considered random intervention on one ward only but did not succeed due to organizational changes on the wards. Other designs, such as cross-over designs, might have been better although more costly. The control of changing work with indicator parameters did not influence time distribution. Although organizational differences between the wards could be measured, it is possible that those control parameters were not optimal. The intervention was not blinded; observation effects (Hawthorne effect) are likely both in control and intervention periods. The evaluation represents a snapshot in time in a changing system environment. There is no rule regarding when to perform the evaluation.

The study was designed to measure clinical effects. As it seemed difficult to measure effects on outcome quality, the study concentrated on process quality. Work sampling statistics proved a problem as time distribution on single work packages is highly dependent from time for other work packages. A statistically acceptable comparison between interventions was only possible on the level of grouped work packages. Similar to many other studies, the questionnaire tools were not standardized, so that comparison with other researchers is difficult. Most essentially, however, it must be emphasized that studies of this kind require a tremendous amount of time and resources. This study lasted for over 2 years and other researchers have reported a similar duration (Van Gennip *et al*. 1994). The equivalent of 3–4 man years was spent on this evaluation, not counting all the preliminary preparatory work.

(E) Assessment With this study we present an objective approach to the evaluation of clinical information systems. External influence is partially controlled even without randomization. The study measures influence on process quality and applies statistically sound methodology to the evaluation. It turns out that the presumed quality indicator, time distribution, is not appropriate on the given statistical level for measuring of influence. We are able to say this because one of the two measured clinical systems was at the time an approved and accepted standard in our hospital. Positive results in this study derive from the questionnaire parts where nurses state a positive attitude and confirm usefulness of certain functions of the HIS.

Third example: evaluation of process quality and user satisfaction

(A) Study design and environment A computer-based nursing documentation system (PIK) was compared with the existing paper-based nursing documentation system (Ammenwerth *et al*. 1999). The study was conducted on one ward of the Department of Psychiatry of the Heidelberg University Medical Centre. Nurses used PIK for care planning and nursing documentation for the 30 patients in the intervention group, and they used the existing paper-based documentation system for care planning and nursing documentation for the 30 patients in the control group. The objective was to evaluate the differences between both patient groups concerning time effort for nursing documentation and quality of nursing documentation. Furthermore, the effects of computer-based documentation on acceptance of nursing process and computers in nursing, on user satisfaction and on co-operation with other professional groups, were assessed.

The study was conducted as a 2-month, randomized controlled trial with 60 patients. Differences in time used for nursing documentation were assessed by time measurements, carried out by the nurses themselves. Differences in quality of documentation between the PIK group and control group were assessed by documentation analysis and interviews. Differences in user satisfaction were assessed at the beginning and at the end of the study by questionnaires and interviews. Most of the questionnaires were based on available standardized questionnaires (for example Bowmann *et al.* 1983; Nickell & Pinto

1986; Chin 1988; Lowry 1994; Ohmann *et al*. 1997a). Nevertheless, they had to be adapted to the specific environment of the study.

(B) Measured results Both objective and subjective results showed advantages of computer-based nursing documentation. Time needed for care planning was much lower in the PIK group than in the control group (16.4 min vs. 43.3 min). Time for nursing documentation was significantly higher in the PIK group (4.4 min for documentation of tasks, 6.2 min for report writing) than in the control group (1.8 min and 3.5 min).

The overall quality of the content of documentation was judged as equal in both groups by two external nursing experts (2.4 vs. 2.3 on a 5-point scale). Some formal aspects of quality (such as completeness and legibility) were strongly higher in the PIK group. User acceptance of computers in nursing and of computer-based nursing documentation systems grew significantly during the study. In general, the acceptance of PIK by nurses and by physicians was high. PIK is now used routinely on two wards and will be introduced on other wards at Heidelberg University Medical Centre.

(C) Power of used methodology This was an RCT where patients had been randomly distributed in intervention and control groups for nurse care planning on one single ward. Therefore, influences due to patients' data (e.g. type of disease) or due to organizational structures (e.g. documentation workflow) were minimized. The study permitted comparison directly between paper-based and computer-based nursing process documentation. To our knowledge no other randomized study exists which compares two nursing documentation systems in a clinical environment. Another advantage of the study design was that every nurse on the ward got experience both with the 'old' and the 'new' documentation system. Quantitative methods such as time measurements were combined with qualitative methods such as interviews and questionnaires. This multimethod approach delivered a multitude of results and improved the robustness of its results (e.g. Kaplan 1995). An evaluation should be formative ('constructive') and provide guidance for further system development (Kaplan 1997; Brender 1998). This study followed the formative evaluation concept. All program malfunctions and user proposals for software improvement have been reported and discussed with software developers and users. This motivated the nurses, giving them feedback that their reports could directly influence software quality.

(D) Problems of used methodology In this study the time measurement for nursing process documentation was carried out by the nurses themselves. Self-assessment of effort, however, can be biased, for example due to incomplete data. However, external observation 24 h a day for several weeks would have been too expensive. Statistical approaches such as work-sampling (Sittig 1993) were not adequate as the time used for nursing process documentation was very low (the estimates were about 5% of the overall time for the study ward). Probably there are missing values in the self-assessment, and the absolute time needed for care planning or documentation may be too low. However, this should not influence comparison between intervention and control groups. In addition, all subjective results (from questionnaires and interviews) confirm the results of the self-measurements. It remains an open question how the time savings may influence patient care.

The study was not blinded, observation effects (Hawthorne effect) are likely both in the control and intervention. The evaluation represents a snapshot in time in a changing system environment. There was no rule as to when to perform the evaluation. The study was designed to measure clinical effects. As it seemed difficult to measure effects on outcome quality, the study concentrated on process quality. Documentation quality was difficult to measure. No validated tools to measure documentation quality exist. Therefore, the two external nursing experts were asked to use a self-constructed checklist to judge documentation quality. Good interobserver validity was noted, both experts judged nearly equal, but those results remain partially subjective. Starting with standardized tools, the researchers finally used some non-standardized questionnaires and checklists, so that comparison with other researchers is difficult. This study did require a high amount of time and resources.

(E) Assessment The study shows successful randomization with significant results in a clinical environment, using a multiple-method approach. Full RCT level is achieved. The study measures influence on structure and process quality and applies statistically sound methodology to the evaluation. The quality indicator time saving is positive for nurse care planning and slightly negative for care documentation. The indicator of documentation quality is unchanged. User acceptance is high. However, the study results are representative only for very similar environments (e.g. a psychiatric ward with equally IT-skilled staff) and cannot be easily compared with other, especially non-psychiatric, departments.

Fourth example: subjectivist evaluation of human factors

(A) Study design and environment Mobile information and communication systems in clinical routine have the potential to improve communication, facilitate information access, eliminate duplicate documentation, and increase quality of patient care in the long run. Mobile computers can support documentation, information access and communication. A prototype version of such a mobile, multifunctional personal assistant was evaluated in a 1-week study in the Heidelberg University Medical Centre (Ammenwerth *et al*. 2000)

The mobile personal assistant was implemented on Apple, Newton 2000, and networked by the German mobile telephone network D1. The prototypes offered functions such as access to a simulated patient database, access to medical knowledge, diagnosis documentation and coding, electronic examination requests, a personal organizer and speech and text communication. The aim of this study was to test the prototype in a close-to-reality environment with genuine users. A variety of research questions were processed in this study, such as suitability for routine use, usefulness of mobile communication, mobile information needs of the staff, and scenarios for mobile documentation. The study was designed as a simulation study. Thirty-one test users (mainly physicians and nurses) were observed during one week while working with the prototypical technology in a close-to-reality environment. Simulated patient cases were prepared to be 'treated' by test users, parallel to their usual daily work and in their usual environment. Special problems such as data security were examined under direct intervention in the simulation environment (i.e. by deliberately disturbing communication pathways). Statements and opinions of the users concerning the research questions were captured during user observation, intensive interviews (before and after the study) and questionnaires.

Simulation studies do not permit a complete evaluation, but rather aim to obtain design proposals for a new technology from the prospective user. As a test method, they combine elements of the laboratory test and the field study (Roßnagel *et al*. 1999; Kumbruck & Schneider 1995). They are an example of subjectivist, formative evaluation.

(B) Measured results A majority of the participants felt that mobile communication was useful in assisting them in their daily hospital routine. However, opinions on the usefulness of computer-aided mobile information processing were diverse. They depended, among other things, on the activity and on the general attitude of a physician towards this new technology. A positive rating was given for mobile access to medical knowledge, to patient data and to general information (such as telephone books). Some users saw potential use for mobile documentation during physician's rounds. Overall, most of the 31 test persons noticed need for mobile computer implementation in clinical routine. Most of them felt that mobile computers can complement the functionality of the clinical workstation in the areas of mobile communication and mobile information processing. However, some users saw no need for mobile computers at all, as long as basic functions could be performed at the clinical workstation. In contrast to the more enthusiastic mobile personnel, the more sceptical users were less mobile in their work. Based on these results, a multidevice mobile computer architecture was conceived that combines stationary clinical workstations and different types of mobile computers.

(C) Power of used methodology The simulation study is an appropriate method to obtain close to reality a safe assessment of prototype technology. Most test persons welcomed the possibility to partic-

ipate in the design and evaluation of a new technology before its introduction into routine. A variety of users from different professional backgrounds participated, thus different attitudes, ranging from the sceptical to enthusiastic, were involved. From the interviews, ideas and suggestions could be extracted that would have been difficult otherwise. These valuable statements and suggestions would surely not have been uncovered through a pure laboratory study.

(D) Problems of used methodology Simulation studies have some disadvantages. They do not allow complete and summative evaluation of a new technology. Instead, they deliver hints and opinions to guide future development. No comparison with a control group is possible, the results are only valid for the chosen simulation environment. Preparing and processing a simulation study requires much effort and work, both for researchers and test users. Several test users reported that processing the test cases parallel to the normal work was more work than expected. This can diminish motivation of the test user. Some users complained that working with simulated patient data was missing practice with real situations. Overall, it seems difficult to weigh up the results of the study against the amount of work needed.

(E) Assessment This study aimed at a formative, mostly subjectivist evaluation of new mobile technology. No controlled environment was used. A lot of valuable statements from future users could be extracted, but the amount of work both for researchers and test users was high. Overall, simulation studies should only be carried out in areas where no earlier experiences are available, and where field studies are not yet possible (for example due to unstable or potentially harmful technology).

Discussion

We have now presented four different examples of evaluation of clinical information systems.

How to interpret the results?

As we have shown, evaluation is dependent on available resources, goals of the evaluation and type of technology that is to be examined. None of the four approaches solves all the evaluation problems, a generic solution does not exist. All examples are centred around the third and fourth evaluation phase, namely evaluation of human factors and evaluation of clinical effect. The examples range from rather primitive questionnaire studies through complex combined approaches up to simulation studies.

We have to accept, that with a given amount of resources only a restricted evaluation is possible. Therefore, simple questionnaire studies retain an essential position in evaluation strategy. They are especially suitable for the evaluation of human factors and can be quickly accomplished with limited resources. When conducting such a study one should look out for validated tools to measure human factors in order to make results comparable with the work of others. However, often those tools will need local adaptation. One should try to randomize observation objects, e.g. when a certain system is to be assessed, one might select half of the questionnaire recipients randomly from system users and the other half from a comparable non-system-user clientele. A pure questionnaire study will often suffer from low return rates. Repeated reminders can ease this problem but will increase the amount of needed resources.

To obtain detailed information on an observation area a simple questionnaire study is not sufficient. In this case a mixed approach must be selected that combines the measurement of several indicators. The combined picture of all indicators will be clearer than each result alone. We notice that those mixed approaches require an effort which is at least tenfold that of a simple questionnaire or log study. Therefore for such a study it is essential to have an idea of what to expect from system introduction in order to define clear and appropriate evaluation goals. Many researchers attempted to measure time saving as a process quality indicator for the system and failed to get positive results. Other quality indicators such as documentation quality should be combined with human factor assessment to derive a conclusion about system impact. The formative character of such studies should be emphasized if possible by influencing and improving system design based on study results. Therefore, questionnaires and observation

will not be restricted to attitude towards computers, but may have to be specified depending on the functions that the observed computer system has. This explains why standardized questionnaire tools are often *per se* not sufficient because they lack questions which are specific for the examined system and working environment.

For many quality indicators such as time or quality, different measuring methods have been developed. We have shown two examples for time measurement, direct measuring and the statistical work sampling method. The first method quickly delivers exact time-spans (if external observers are used) but is often difficult to accept in the hospital environment. The second needs a higher preparatory and measuring effort but may be more acceptable in hospital (people do not feel they have to work against the stopwatch) and can show how the time is used instead. The latter may be essential as there are hints in the literature that saved time is not used for patient care. Similar thoughts are required for other quality indicators, e.g. documentation quality, and often the measurement remains somewhat subjective.

Mixed observation studies are often non-randomized, due to organizational facts (e.g. a system being implemented in a certain part of the hospital) and the problem of not having enough observation objects for a sensible randomization. This fact must be accepted, but good study designs will allow for a certain control of unwanted effects, e.g. by measuring and weighting those effects, by implementing a matched pairs study or with cross-over design (ward A, first intervention then control; ward B, first control then intervention). Occasionally randomized controlled trials can be implemented, e.g. when patients can be randomly assigned to intervention and control. Most cases of clinical information system evaluation are not suitable for blinding. Therefore, certain unwanted effects, e.g. the Hawthorne effect, have to be accepted that, however, with some luck are equally distributed between the control group and intervention group. Nevertheless, we may note that evaluation technologies for clinical information systems today slowly become mature enough to make study results comparable, even across country borders. It must be noted, however, that the results of an evaluation study are only representative for a similar environment. This environment should be clearly defined in each study to allow for comparison.

Simulation studies are a rather new field for evaluation of medical information systems. They follow a subjectivist approach where qualitative data are preferred and no facts or indicators are measured. The simulation study is an appropriate method to obtain a safe assessment of prototype technology close to reality. This study type accepts and welcomes different user attitudes and derives its value from ideas and suggestions of the participants. However, simulation studies do not allow the complete and summative evaluation of a new technology. There is no control group and the results are only valid for the chosen simulation environment. Preparation and processing a simulation study requires much effort and work, both for researchers and test users. Simulation studies are appropriate in areas where no earlier experiences are available, and where field studies are not yet possible, for example due to potentially harmful technology.

References

Adlassnig K.P. & Horak W. (1995) Development and retrospective evaluation of HEPAXPERT-I: a routinely-used expert system for interpretive analysis of hepatitis A and B serologic findings. *Artificial Intelligence in Medicine* **7**, 1–24.

Ammenwerth E., Buchauer A., Bludau B. & Haux R. (2000) Mobile information and communication tools in the hospital. *International Journal of Medical Informatics* **57**, 21–40.

Ammenwerth E., Eichstädter R., Kochenburger L., Pohl U., Rebel S. & Haux R. (1999) Systematic evaluation of computer-based nursing documentation system. *Medizinische Informatik, Biometrie und Epidemiologie* **85**, 286–290.

Balas E.A., Austin S.M., Mitchel J.A., Ewigmann B.G., Bopp K.D. & Brown G.D. (1996) The clinical value of computerized information services. A review of 98 randomized clinical trials. *Archives of Family Medicine* **5**, 271–278.

Berner E.S., Webster G.D., Shugerman A.A. *et al.* (1994) Performance of four computer–based diagnostic systems. *New England Journal of Medicine* **330**, 1792–1796.

Bowmann G., Thompson D. & Sutton T. (1983) Nurses' attitudes towards the nursing process. *Journal of Advanced Nursing* **8**, 125–129.

Boy O., Ohmann C., Aust B. *et al*. (2000) Systematische Evaluierung der Anwenderzufriedenheit mit einem Krankenhausinformationssystem – Erste Ergebnisse In: *Medical Infobahn for Europe – Proceedings of MIE 2000 and GMDS*. pp. 518–522. IOS Press, Amsterdam.

Brender J. (1998) Trends in assessment of IT-based solutions in healthcare and recommendations for the future. *International Journal of Medical Informatics* **52**, 217–227.

Bürkle T., Kuch R., Passian A., Prokosch H.U. & Dudeck J. (1995a) The impact of introducing computers on nursing work patterns. Study design and first results. *Medinfo 95* **2**, 1321–1325.

Bürkle T., Kuch R., Prokosch H.U. & Dudeck J. (1995b) Evaluation von Medizinischen Informationssystemen: Methoden, Ziele, Ergebnisse. In *GISI 95 – Herausforderungen Eines Globalen Informationsverbundes für die Informatik* (eds F. Huber-Wäschle, H. Schauer & P. Idmayer), pp. 662–668. Springer, Berlin.

Bürkle T., Kuch R., Prokosch H.U. & Dudeck J. (1999) Stepwise evaluation of information systems in a university hospital. *Methods of Information in Medicine* **38**, 9–15.

Bürkle T., Schmitz M., Prokosch H.U. & Dudeck J. (1994) *A Systematic Approach for Evaluation of Nursing Work in a University Hospital. Proceedings 12th Medical Informatics, Europe, 1994*. EFMI, Lisbon.

Chin J. (1988) *Development of a Tool Measuring User Satisfaction of the Human–Computer Interface. Chi'88 Conference. Proceedings: Human Factors in Computing.* Association for Computing Machinery, New York.

Donabedian A. (1982) *The Criteria and Standards of Quality*. Health Administration Press, Ann Arbor, Michigan.

Engelbrecht R., Rector A. & Moser W. (1995) Verification and validation. In *Assessment and Evaluation of Information Technologies* (eds E.M.S.J. Van Gennip & J.L. Talmon), pp. 51–66. IOS Press, Amsterdam.

Evans R.S., Gardner R.M., Bush A.R., Burke J.P., Jacobson J.A., Larsen R.A., Meier F.A. & Warner H.R. (1985) Development of a computerized infectious disease monitor (CIDM). *Computers and Biomedical Research* **18**, 103–113.

Evans R.S., Pestotnik S.L., Classen D.C. & Burke J.P. (1993) Development of an automated antibiotic consultant. *MD Computing* **10**, 17–22.

Evans R.S., Pestotnik S.L. & Gardner R.M. (1995) Evaluating the impact of computer-based drug monitoring on the quality and cost of drug therapy. In *Hospital Information Systems: Design and Development Characteristics*; *Impact and Future Architectures* (eds H.U. Prokosch & J. Dudeck), pp. 201–220. Elsevier, New York.

Forsythe D.E. & Buchanan B.G. (1992) *Broadening our Approach to Evaluating Medical information Systems. Proceedings 15th Annual Symposium on Computer Applications in Medical Care, 8–12*. McGraw Hill, New York.

Friedman C.P. & Wyatt J.C. (1997) *Evaluation Methods in Medical Informatics*. Springer, New York.

Heathfield H.A., Peel V., Hudson P., Kay S., Mackay L., Marley T., Nicholson L., Roberts R. & Williams J. (1997) *Evaluating Large Scale Health Information Systems: From Practice Towards Theory. Proceedings of the 1997 AMIA Annual Fall Symposium*. Hanley & Belfus, Philadelphia.

Hinson D.K., Huether S.E., Blaufuss J.A., Neiswanger M., Tinker A., Meyer K.J. & Jensen R. (1994) *Measuring the Impact of a Clinical Nursing Information System on One Nursing Unit. Proceedings Seventeenth Annual Symposium on Computer Applications in Healthcare SCAMC*. McGraw Hill, New York.

Ives B. (1983) The measurement of user information satisfaction. *Communications of the ACM* **26**, 785–793.

Johnston M.E., Langton K.B., Haynes R.B. & Mathieu A. (1994) Effects of computer-based clinical decision support systems on clinician performance and patient outcome. *Annals of Internal Medicine* **120**, 135–142.

Kaplan B. (1995) An evaluation model for clinical information systems. Clinical imaging systems. *Medinfo 95* **2**, 1087.

Kaplan B. (1997) Addressing organizational issues into the evaluation of medical systems. *Journal of the American Medical Informatics Association* **4**, 94–101.

Koch A. & Abel U. (1997) Zur Aussagekraft nichtrandomisierter Studien. *Medizinische Informatik, Biometrie und Epidemiologie* **82**, 391–395.

Kumbruck C. & Schneider M.J. (1995) Simulation Studies, a New Method of Prospective Technology Assessment and Design. Darmstadt Project Group 'Verfassungsverträgliche Technikgestaltung' (provet) e. V. http://www.provet.org/publikat.htm.

Kushniruk A.W., Patel V.L. & Cimino J.J. (1997) *Usability Testing in Medical Informatics: Cognitive Approaches to Evaluation of Information Systems and User Interfaces. Proceedings of the 1997 AMIA Annual Fall Symposium*. Hanley & Belfus, Philadelphia.

Lowry C. (1994) Nurses' attitudes toward computerised care plans in intensive care. Part 2. *Intensive and Critical Care Nursing* **10**, 2–11.

McDonald C.J. (1976) Protocol-based computer reminders, the quality of care and the non-perfectability of man. *New England Journal of Medicine* **295**, 1351–1355.

Myers G.J. (1976) *Software Reliability – Principles and Practices*. John Wiley & Sons, New York.

Myers G.J. (1982) *Methodisches Testen Von Programmen*. Oldenbourg-Verlag, München.

Nickell G. & Pinto J. (1986) The Computer Attitude Scale. *Computers in Human Behaviour* **2**, 301–306.

Ohmann C. & Belenky G. (1995) Leitfaden zur Evaluierung von wissensbasierten Systemen. *Medizinische Informatik, Biometrie und Epidemiologie* **80**, 417–420.

Ohmann C., Boy O. & Eich H.P. (1998) Arbeitskreis Evaluation im MEDWIS-Programm. Leitfaden zur Evaluierung von wissensbasierten Systemen. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **29**, 77–83.

Ohmann C., Boy O. & Yang Q. (1997a) A systematic approach to the assessment of user satisfaction with health care systems: constructs, models and instruments. *Studies in Health Technology and Informatics* **43** Part B, 781–785.

Ohmann C., Boy O., Yang Q. & Eich H.P. (1997b) Evaluierung der Benutzerzufriedenheit mit einem Krankenhausinformationssystem: Theoretische Aspekte und klinische Anwendung. *Medizinische Informatik, Biometrie und Epidemiologie* **82**, 31–34.

Peterson H., Gerdin Jelger U. (1988) Evaluation: a means to better results. In *Nursing Informatics* (eds M.J. Ball *et al.*), pp. 64–77. Springer, New York.

Petrucci K., Petrucci P., Canfield K., McCormick K.A., Kjerulff K. & Parks P. (1992) *Evaluation of UNIS: Urological Nursing Information System. Proceedings Fifteenth Annual Symposium on Computer Applications in Healthcare SCAMC*. McGraw Hill, New York.

Rind D.M., Safran C., Phillips R.S., Slack W.V., Calkins D.R., Delbanco D.L. & Bleich H.L. (1992) *The Effect of Computer-Based Reminders on the Management of Hospitalized Patients with Worsening Renal Function. Proceedings Annual Symposium on Computer Applications in Healthcare SCAMC.* McGraw Hill, New York.

Roßnagel A., Ammenwerth E., Buchauer A. & Bludau H.B. (1999) Simulation study for the evaluation of security technology. In *Multilateral Security for Global Communication* (eds G. Müller & K. Rannenberg), pp. 547–562. Addison-Wesley, Bonn.

Schmitz P., Bons H. & van Megen R. (1982) *Software-Qualitätssicherung-testen im Software-Lebenszyklus*. Vieweg, Braunschweig.

Sittig D.F. (1993) Work-sampling: a statistical approach to evaluation of the effect of computers on work patterns in healthcare. *Methods of Information in Medicine* **32**, 167–174.

Spranzo Keller L., McDermott S. & Alt-White A. (1992) *Effects of Computerized Nurse Careplanning on Selected Health Care Effectiveness Measures. Proceedings Fifteenth Annual Symposium on Computer Applications in Healthcare SCAMC*. McGraw Hill, New York.

Tierney W., Overhage J. & McDonald C. (1994) A plea for controlled trials in medical informatics. *Journal of the American Medical Informatics Association* **1**, 353–355.

Tippett L.C.H. (1935) A snap reading method of making time studies of machines and operations in factories: a survey. *Journal of Textile Institute* **26**, 51–70.

Van Bemmel J.H. & Musen M.A. (1997) *Handbook of Medical Informatics*. Springer, Heidelberg.

Van der Loo R.P., van Gennip E.M.S.J., Bakker A.R., Hasman A. & Rutten F.F.H. (1994) *Evaluation of Automated Information Systems in Health Care: an Approach to Classify Evaluative Studies. Proceedings 12th Medical Informatics Europe, 1994*. EFMI, Lisbon.

Van Gennip E.M.S.J., Kramer H., Enning C.J., Klaassen-Leil C.C., Stokman R.A.M. & Van Valkenburg R.K.J. (1994) *VISTA: Study of Effects of an Integrated Nursing Information System in Three Dutch Hospitals Set-up and Intermediate Results. Proceedings 12th Medical Informatics Europe, 1994*.

Verdaguer A., Patak A., Sancho J.J., Sierra C. & Sanz F. (1992) Validation of the medical expert system PNEUMON-IA. *Computers and Biomedical Research* **25**, 511–526.