An Inventory of Evaluation Studies of Information Technology in Health Care

Trends in Evaluation Research 1982-2002

E. Ammenwerth[1], N. de Keizer[2]
[1]UMIT – University for Health Sciences, Medical Informatics and Technology, Institute for Health Information Systems, Hall in Tyrol, Austria
[2]Department of Medical Informatics, Academic Medical Center, University of Amsterdam, The Netherlands

Summary

Objectives: During the last years the significance of evaluation studies as well as the interest in adequate methods and approaches for evaluation has grown in medical informatics. In order to put this discussion into historical perspective of evaluation research, we conducted a systematic review on trends in evaluation research of information technology in health care from 1982 to 2002.

Methods: The inventory is based on a systematic literature search in PubMed. Abstracts were included when they described an evaluation study of a computer-based component in health care. We identified 1035 papers from 1982 to 2002 and indexed them based on a multi-axial classification describing type of information system, study location, evaluation strategy, evaluation methods, evaluation setting, and evaluation focus.

Results and Conclusions: We found interesting developments in evaluation research in the last 20 years. For example, there has been a strong shift from medical journals to medical informatics journals. With regard to methods, explanatory research and quantitative methods have dominated evaluation studies in the last 20 years. Since 1982, the number of lab studies and technical evaluation studies has declined, while the number of studies focusing on the influence of information technology on quality of care processes or outcome of patient care has increased. We interpret this shift as a sign of maturation of evaluation research in medical informatics.

Keywords

Evaluation studies, health and medical informatics, technology assessment, information technology, health care, health information systems, inventory, review

Methods Inf Med 2005; 44: 44–56

Introduction

During the last years the significance of evaluation studies has grown in medical informatics. Evaluation studies are increasingly considered part of the planning, development, introduction and operation of information technology in health care [1-5].

Evaluation can be defined as the act of measuring or exploring some property of a system, the result of which informs a decision concerning that system in a specific context [6]. Evaluation of health information systems has to deal with the actors (the people), the artifacts (the technology), and the environment in which it is implemented as well as with their interaction [3].

Discussion in medical informatics addresses best methods and approaches, the need for an evaluation framework, and the quality of evaluation studies often deemed insufficient. This discussion has gone on for some years now (compare e.g. [3, 7, 8]).

In order to put this discussion into historical perspective of evaluation research, we conducted a systematic review of evaluation research of information technology in health care from 1982 to 2002. The aim of this review was to identify trends of evaluation research in this area in the last 20 years. We were, e.g., interested in the dynamics of studies in recent years, as a reflection of interest in evaluation of information systems. We wanted to learn what countries or journals dominate in evaluation research publications, supporting a focused search for evaluation studies. We were then interested to learn which type of information systems (e.g. CPOE) are predominantly evaluated, and whether there are shifts in recent years (e.g. we expected rising numbers of evaluations of telemedical systems reflecting a trend towards telemedical and cooperative information systems, compare [9, 10]). For the same reason, the location where evaluation studies took place was of interest for us (e.g. we expected a rising number of evaluations at the patient's home reflecting an increased significance of home-based care). Our inventory then emphasizes the focus of a study (e.g. software quality, effect on outcome quality) and the methods applied (e.g. qualitative versus quantitative methods), as methodological discussions have been going on for some years now in medical informatics (compare e.g. [5, 11-13]).

To our knowledge, no comparable analysis has yet been done. We will discuss differences to other approaches in the discussion section in more detail.

In this paper, we will present the results of our inventory of evaluation studies. We will answer the following questions:

- How does the number of published studies evolve?
- Which countries do the studies come from, in which languages have they been published?
- In which journals have they been published?
- Which types of information systems have been evaluated?
- In which location did the studies take place?
- Which evaluation strategy has been chosen?

Received: April 2, 2004; accepted: October 27, 2004

- Which evaluation methods have been used?
- In which setting did the studies take place?
- Which evaluation focus has been addressed?

In addition, we analyzed the correlations between the various aspects in an exploratory way to identify relationships (e.g. correlation of evaluation focus and methods applied, or of information system and evaluation strategy).

Methods

The inventory is based on a systematic literature search in PubMed. We decided to concentrate on the last 20 years, i.e. to only include studies published since 1982, to have a sufficient number of studies to analyze for each year, even if a few evaluation studies are older (such as the well-known study by Simborg [14]).

PubMed Search Strategy

We included papers on evaluation studies of health information systems. We defined a *health information system* as including all computer-based components which are used by health care professionals or the patients themselves in the context of inpatient or outpatient patient care to process patient-related data, information or knowledge. We defined an *evaluation study* as the systematic, empirical assessment of a component of a health information system. We did not include papers that only describe a study design, or that just contain system descriptions.

Based on those definitions, we decided to exclude medical-technical components (such as robotics or virtual reality systems) and all systems or methods which are only used to analyze images or signals (e.g. automatic image analysis system). We also excluded all computer-based training and education systems for health care professionals since these are not part of direct patient care. For the same reason we did not include studies on epidemiological systems or administrative systems. We also excluded isolated hardware evaluations and also evaluations of data capturing methods (e.g. touch-screen versus mouse) unless evaluated in direct patient care. Telemedical systems were only included if there was a clear indication of computer-based data transmission (not including videoconferencing tools or telephone-based telemedicine as we did not consider these as 'computer-based'). We also decided to exclude papers on drug-dosing algorithms (e.g. comparison of two algorithms) and on screening reminders sent to patients by normal mail. Finally, we excluded general surveys (e.g. on general computer knowledge or computer acceptance of a certain group).

We decided to focus our analysis on papers indexed in PubMed [15], as it makes the most important medical and medical informatics journals available. To make the inventory feasible, we decided to base it on the abstracts of the papers only, thus we only included papers where an abstract was available.

The first step in the search strategy in PubMed comprised a selection of papers between 1982 and 2002 based on the following three queries which were combined by "AND":

1a) Search for *health information systems* by searching in title words (e.g. computer, record, documentation, program, reminder, protocol, decision), in Major Mesh Heading (medical informatics), and in Minor Mesh Heading (e.g. computers, computer-assisted instruction, decision-support systems, hospital information systems, management information systems, medical record systems, microcomputers, radiology information systems, reminder systems, telemedicine, attitude to computers).

1b) Search for *evaluation studies* by searching in title words (e.g. impact, effect,

Table 1
Examples of titles of papers included or excluded in the inventory

Included Paper	*Excluded Paper (with short comment)*
An evaluation of telemedicine in surgery: telediagnosis compared with direct diagnosis	Ongoing developments in CCP4 for high-throughput structure determination (bioinformatics papers are not included)
A randomized, controlled trial of clinical information shared from another institution	Digital hearing aids and future directions for hearing aids (technical support devices are not included)
The effect of computer-assisted prescription writing on emergency department prescription errors	Computer-aided diagnosis of breast tumors with different US systems (CAD papers evaluating the algorithm are not included)
Patient attitudes toward using computers to improve health services delivery	Point-of-care testing of blood glucose in the neonatal unit (point-of-care tools are not included)
Effect of computerised evidence based guidelines on management of asthma and angina	Reproducibility of three different scoring systems for measurement of coronary calcium (evaluation of scoring systems is not included)
Systematic review of cost effectiveness studies of telemedicine interventions	Is there a role for computerized decision support for drug dosing in general practice? A questionnaire survey (general surveys are not included)
Feasibility of using a computer-assisted intervention to enhance the way women with breast cancer communicate with their physicians	Risk behaviours by audio computer-assisted self-interviews among HIV-seropositive (no evaluation study)
An evaluation of an automated anaesthesia record keeping system	Touch-screen system for assessing visuo-motor exploratory skills (evaluation of methods for data entry, e.g. touch-screen vs. paper form, are not included)

evaluation, meta-analysis, review), publication type (e.g. clinical trial, evaluation studies, meta-analysis, randomized controlled trial, review), or Minor Mesh Heading (e.g. comparative study, feasibility study, costs and cost analysis, interviews, questionnaires, program evaluation).

1c) Include only papers with an English *abstract* available.

The amount of selected papers was then reduced by conducting the next two steps:

2) *Automatically* exclude all papers where we found indication that it was not of interest for our inventory, containing Mesh headings (e.g. animals, biology, DNA, plants, robotics, epidemiology) or certain title words (e.g. radiotherapy, brain, navigation, education, training). The remaining papers were stored in a database.
3) *Manually* exclude papers by checking title and abstract for our inclusion and exclusion aspects described above. This was done by the two authors together, based on detailed instruction for inclusion and exclusion. Any discrepancies were solved by discussion.

Table 1 presents some examples of papers included and excluded for our review.

The complete PubMed query is available upon request. To check the recall of our query, we chose review papers citing evaluation studies as a gold standard and controlled whether our query found the studies cited in those papers.

Multi-axial Classification

Each of the identified abstracts was then classified according to a multi-axial classification developed for this inventory purpose, to allow for a systematic historical analysis on the various aspects we were interested in. All included papers do have a title and abstract in English (even when the language of the paper itself is different). When the information in the abstracts was not sufficiently detailed on a given aspect (e.g. on evaluation methods used), this aspect was coded as "unclear or other". We also included review papers in our inventory, but classified them only by evaluation strategy and type of information system. The coding of each abstract based on the classification was done by both authors; any discrepancies were solved by discussion.

A review of the literature showed some earlier work which we used to built on our classification, especially work from Krobock [16], Sawyer and Chen [17], Grémy and Degoult [18], and van der Loo [19] (for a detailed analysis, see [20]). Our classification contained the following axes:

Type of Information System

We based this part of the classification on earlier work of Grémy and Degoulet [18] and van der Loo [19]. Grémy and Degoulet distinguished population-based systems, hospital information systems, clinical systems, clinical laboratories, consultation systems, training systems, and robotics. Choosing another approach, van der Loo distinguished diagnostic systems, therapeutic systems, nursing systems, supporting systems, and auxiliary systems. For our purpose, both approaches were not sufficiently detailed and we want to have mutually exclusive classes as much as possible. We therefore refined and extended them, based on several pre-tests with a subset of studies, to the following categories:

- patient information system (PIS; component primarily used by patients), physician order entry system (CPOE), radiology information system (RIS), picture archiving and communication system (PACS), laboratory information system (LIS), pharmaceutical information system (PHARM), surgery management system (OP; operation planning, management, documentation), anesthesia documentation system (ANAEST), patient data management system (PDMS; for patient monitoring in ICUs), nursing information system (NIS; nursing care planning and documentation), GP information system (GP; component for GPs and comparable outpatient care), teleconsultation system (TC; telemedical applications focussing on consultation, e.g. teleradiology), telemonitoring system (TM; telemonitoring of patients e.g. at home), expert system (XPS; specific knowledge-based components, including guideline-based systems and reminder systems), other or general clinical information system (CIS).

Location where the Evaluation Study Took Place

Our classification of the location of study was derived from a subset of evaluation studies and optimized during the course of the inventory to be able to code all papers as clearly as possible. The categories were:

- at the patient's home; at a general practitioner or in a health care center; in an outpatient unit or outpatient clinic; in an intensive care unit, emergency unit, or operation unit; in any other inpatient care unit; in a lab, pathology, pharmacy, or blood bank; in a radiology unit; in a trans-institutional setting (i.e. a system covering different clinical physical locations, e.g. telemedical systems); or others or unclear.

The location can, but need not to be correlated with type of information system (e.g., a PACS can be evaluated in a radiology unit or in an inpatient care unit).

Evaluation Strategy which Was Applied in the Study

Sawyer and Chen [17] presented a taxonomy distinguishing experimental, intensive/field-based (e.g. case studies), computational (e.g. simulation), and other studies. Based on their work and on general literature of research methods (e.g. [21]), we decided to distinguish exploratory studies (e.g., intensive case studies) and explanatory studies (e.g., controlled clinical trials) as main categories:

- Exploratory (explore a situation, generate hypothesis, find relationships, e.g. "Which side-effects occurred after the introduction of IT?", "Why do or don't users accept the information system component?"); explanatory (test hy-

potheses, e.g. "IT increases the quality or efficiency of care", "IT decreases costs of care"); review (only systematic reviews, overview papers were not included); mixed or unclear.

Methods Used in the Evaluation Study

Van der Loo [19] used a rather detailed classification to classify data collection methods (e.g. observation, interviews), and study design. As abstracts often do not contain sufficient information for such a detailed coding, we decided to concentrate on the following aspects:

- Predominantly quantitative methods (working with numbers, e.g. time measurements, quantitative user acceptance measurements, length of stay measurements, error rate scores); predominantly qualitative methods (working with text, e.g. focus group interviews, qualitative content analysis); and mixed or unclear methods.

Setting of the Study

As setting of a study, we use the following separation:

- Field study (i.e. study is executed in a realistic environment); lab study (i.e. environment is at least partly controlled, e.g. using test data or test user); and mixed or unclear setting.

Focus of the Evaluation Study

The classification of the focus of the evaluation study (in other words, of the evaluation criteria used) was more complex to define. Already in 1984, Krobock [16] developed a comprehensive hierarchical classification of evaluation questions that was, however, too detailed for our purpose. Van der Loo [19] structured the focus into structural focus (e.g. system performance, time consumption, user satisfaction), process focus (e.g. health care consumption, database use), and outcome focus (e.g. patient outcome, patient satisfaction). A slightly other approach was presented by Grémy and Degoult [18]. They separated evaluation criteria into technical level (e.g. response time, reliability), medical efficiency (e.g. completeness of record, adherence to protocols, patient compliance, and quality of care), economy (e.g. costs), sociological aspects (e.g. acceptability, usefulness, outside vision of the system), and philosophical aspects (e.g., human-computer relationships). Both classifications seemed quite interesting, however, detailed instruction and description of the individual categories were missing in both cases. We extended and clarified this earlier work and defined the following categories:

- *Structural quality:* 1) Hardware and technical quality (e.g. network performance, stability of a component, readability of transmitted data). 2) Software quality (e.g. correctness of algorithm, usability of software). 3) General computer knowledge or computer acceptance of users (not related to the evaluated component).
- *Quality of information logistics:* 1) Quality of documented or processed information (e.g. completeness or correctness of data). 2) Costs of information processing (e.g. costs for hardware, software, support). 3) User satisfaction with the component (e.g. user acceptance). 4) Usage patterns of a component (e.g. how often it was used; only coded if this was main study questions).
- *Effects on quality of processes:* 1) Efficiency of working processes (e.g. time or staff needed for a certain task, waiting times, time needed to access information). 2) Appropriateness of patient care (e.g. adherence to clinical guidelines, medication error rates, and clinical skills of clinicians). 3) Organization or social quality (e.g. quality of cooperation and communication in the health care professional team or between clinician and patient, changes in roles and responsibilities)
- *Effects on outcome quality of care:* 1) Outcome quality of patient care (e.g. morbidity, mortality, quality of life). 2) Costs of patient care (i.e. use of resources, such as drugs; length of stay). 3) Patient satisfaction with patient care (how the patient is satisfied with his or her care). 4) Patient-related knowledge or behavior (with regard to his or her illness).

A study could cover more than one study aspect.

Results

The PubMed search was done on April 25, 2003; an update was done on July 28, 2003. The first steps (1a-1c) in the PubMed search resulted in about 45,000 hits. After automatically excluding irrelevant papers (step

Table 2 Recall rates found by comparing the results of the chosen PubMed query with papers cited in six reviews

	Paper	*Type of cited evaluation studies in this paper*	*No. of studies cited which are relevant for this inventory and available with abstract in PubMed*	*No. of studies found by our query*	*Recall rate (%)*
1.	Balas [22]	Computerized information services	40 (of 107 cited studies)	32	80%
2.	Kaplan [23]	Decision-support systems	33 (of 46 cited studies)	27	82%
3.	Roine [24]	Telemedical systems	37 (of 50 cited studies)	33	89%
4.	Friedman [25]	Quantitative evaluation studies	23 (of 26 cited studies)	22	96%
5.	Hunt [26]	Decision-support systems	22 (of 32 studies cited in table 3)	16	73%
6.	Preselection of the section „Quality of Health" for the IMIA Yearbook of Medical Informatics 2003	Candidate studies for inclusion in the IMIA Yearbook 2003 [32]	7 (of 15 cited studies)	7	100%

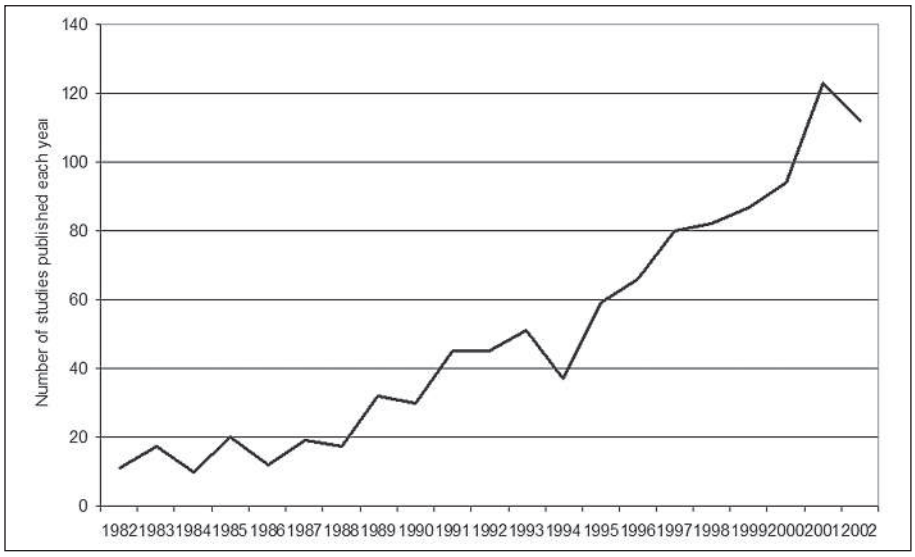

Fig. 1 Number of annually published evaluation studies 1982-2002 (n = 1035 studies)

2), about 15,500 hits remained. After manually controlling those papers (step 3), 1035 papers were left which met our inclusion criteria and which were stored in a database. As planned, the two authors did all manual steps together.

Precision and Recall of the Search Query

From the 15,500 papers automatically selected by our query, about 1035 papers were of interest for our study. The precision of our query was, therefore, about 7%. The recall rates determined by comparing our results with review papers from various areas as gold standards were around 80% and higher (Table 2). For example, from the 107 papers cited by Balas [22], 40 were eligible for our study (the others e.g. had no abstracts, were not in PubMed, or did not meet our inclusion criteria). From those 40 papers, 32 (= 80%) were found by our query. Similar rates (80-90%) were found by checking the papers cited in the reviews by Kaplan [23] and Roine [24].

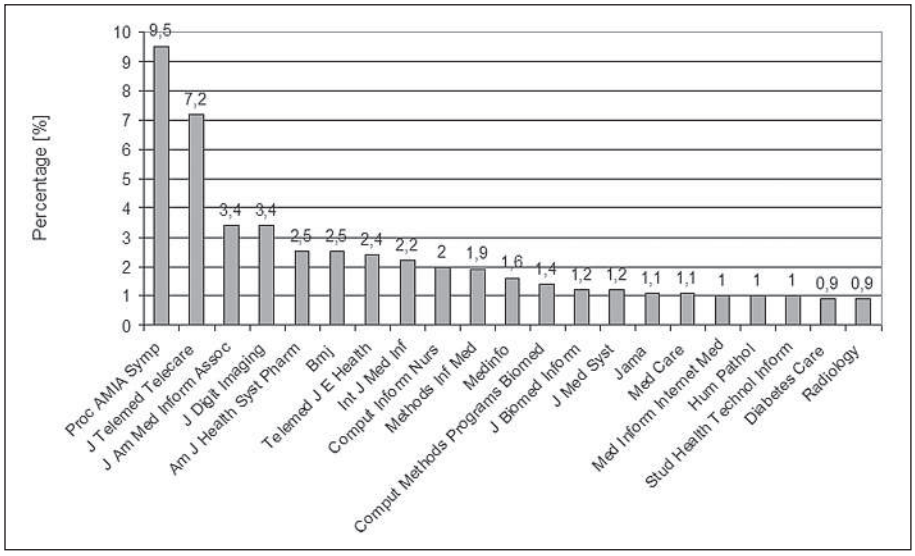

Fig. 2 Percentage of studies published in the top 21 journals 1982-2002 (n = 1035)

However, it should be noted that those first three review papers (Balas, Kaplan, Roine) had also been used as a training set for our query, i.e. the query had been optimized when the recall rate was considered too low. After finishing the optimization and finalizing the query, we once again checked the recall based on three newly selected papers (Friedman [25], Hunt [26] and an IMIA Yearbook preselection list). Here, we found recall rates between 73 and 100% (Table 1) and thus decided that the query was now ready to be used.

Number of Published Studies

The overall number of published studies per year is steadily increasing (Fig. 1). The percentage of evaluation studies compared to the overall number of papers with the Major Mesh Term "medical informatics" in PubMed nearly doubled from 0.6% in 1982-1984 (38 papers out of 6496) to 1.0% in 2000-2002 (326 out of 31,390).

Origin of First Author

More than half of the first authors (n = 541) came from North America (mostly USA, n = 502), about one third from Europe (n = 356, 34%). The dominance of papers from the USA was stable over the last 20 years. 93% (n = 963) of all papers were in English (all papers had an abstract in English), being the dominant language for the last 20 years (the lowest percentage of English papers was 80% in 1992).

Distribution of Journals

The 1035 studies have been published in 343 journals. Half of all studies (n = 509) have been published in only 21 different journals (each of them containing from 9 up to 98 evaluation studies between 1982 and 2002). The number of publications in those journals is given in Figure 2. Please note that not all journals exist since 1982, that some journals changed their name in between, and that some of the 'journals' are in fact proceedings.

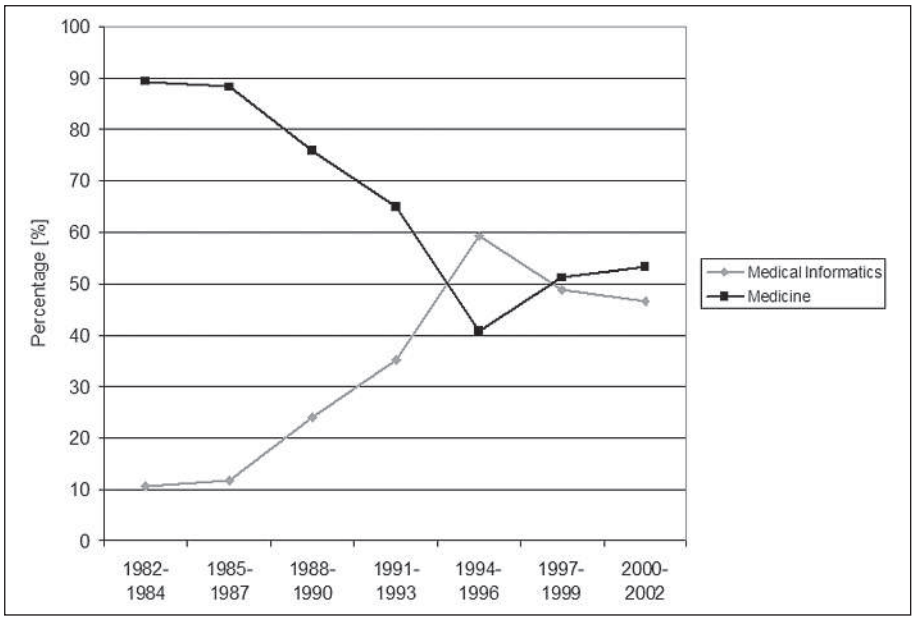

Fig. 3 Percentage of studies published in medical informatics and in medical or other journals 1982-2002 (n = 1035)

A detailed analysis of the journals showed that some (medical) journals were dominant in the nineteen eighties but became less prominent as sources of evaluation papers later. For example, Am J Health Syst Pharm published 18% (n = 7) of all evaluation studies in the period between 1982 and 1984 and only 1% (n = 3) in the period 2000-2002. A comparable development can be found e.g. for Med Care (1982-1984: n = 5, 13%; 0% since 1991). On the other side, new (medical informatics) journals were published. For example, J Telemed Telecare was founded in 1995 and published 13% (n = 44) of all evaluation studies in the period 2000-2002. Comparable developments can be found for J Am Med Inform Assoc which was founded in 1994 and published 4-5% of all studies since then. Proceedings publications in the Proc AMIA Symp were rather dominant in the early 90's but then became less prominent (1994-1996: n = 33, 21%; 1997-1999 n = 26, 11%; 2000-2002: n = 21, 6%).

An interesting development can be found when classifying the 343 journals either as "medical informatics" (32 journals[a]) or as

[a] Proc AMIA Symp; J Telemed Telecare; J Am Med Inform Assoc; J Digit Imaging; Telemed J E Health; Int J Med Inf; Comput Inform Nurs; Methods Inf Med; Medinfo; Comput Methods Programs Biomed; J Biomed Inform; Stud Health Technol Inform; J Clin Comput; J Clin Monit Comput; J Clin Monit; MD Comput; Med Biol Eng Comput; Med Inform; Top Health Inf Manage; Med Inform Internet Med; Artif Intell Med; Biomed Sci Instrum; Biomed Tech; BMC Med Inform Decis Mak; Comput Med Imaging Graph; Healthc Inform; IEEE Trans Biomed Eng; Int J Clin Monit Comput; Int J Technol Assess Health Care; Comput Biol Med; Comput Healthc; Med Decis Making.

Table 3 Percentage of studies on specific type of information systems 1982-2002

	1982–1984	*1985–1987*	*1988–1990*	*1991–1993*	*1994–1996*	*1997–1999*	*2000–2002*	Overall percentage	Overall number of studies
Number of papers published in each period	38	51	79	134	159	248	326	100%	1.035
	Percentages of specific type of information systems per time period								
XPS: Expert system	31.6	51.0	48.1	30.6	25.5	15.3	16.9	24.2	250
TC: Teleconsultation system	2.6		1.3	6.0	17.6	30.2	28.8	20.0	207
CIS: General/other clinical information system	26.3	13.7	12.7	15.7	16.4	14.9	13.5	15.0	155
PACS: Picture archiving and communication system			7.6	7.5	6.9	10.5	7.4	7.4	77
PIS: Patient information system		5.9	3.8	7.5	7.5	6.5	7.7	6.7	69
CPOE: Computerized order entry system	5.3	11.8	3.8	3.7	5.0	3.6	6.7	5.3	55
GP: General practitioner system	10.5	3.9	5.1	6.7	6.9	4.0	4.6	5.3	55
NURSE: Nursing information system			3.8	7.5	5.7	7.3	2.5	4.6	48
PHARM: Pharmaceutical information system	21.1	9.8	5.1	3.7	3.1		1.5	3.1	32
TM: Telemonitoring system				2.2	0.6	0.8	4.6	2.0	21
ANAEST: Anaesthesia information system			1.3	4.5		2.4	0.9	1.5	16
RIS: Radiological information system			3.8	0.7	1.3	2.0	1.5	1.5	16
LAB: Laboratory information system	2.6	3.9	1.3	0.7	1.9	0.4	1.5	1.4	14
PDMS: Patient data management system			1.3	3.0	1.3	1.2	1.5	1.4	15
OP: Operation information system			1.3		0.6	0.8	0.3	0.5	5

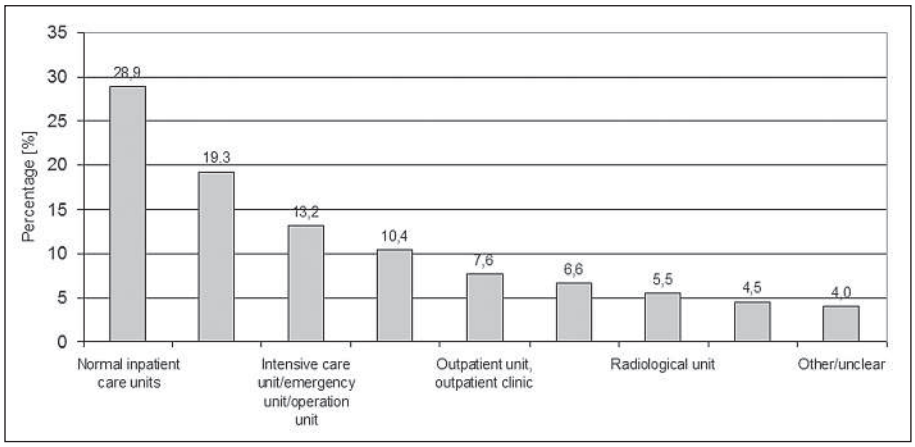

Fig. 4 Percentage of location of evaluation studies 1982 -2002 (n = 983, reviews excluded)

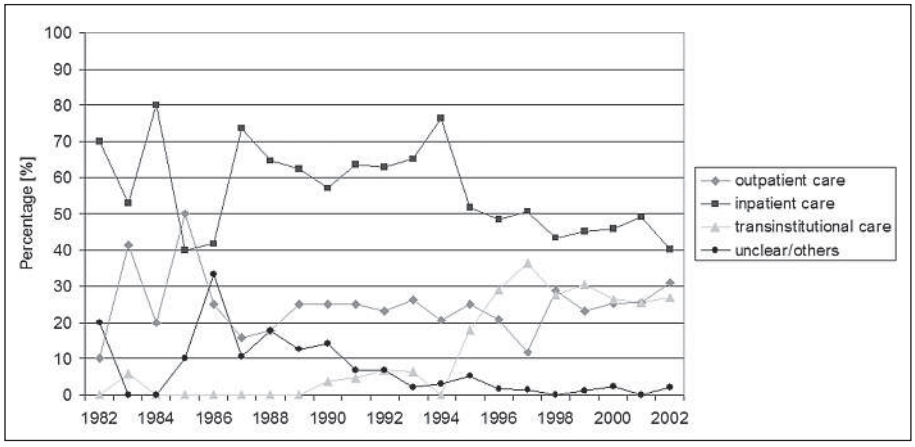

Fig. 5 Percentage of type of location of evaluation studies 1982-2002 (n = 983, reviews excluded)

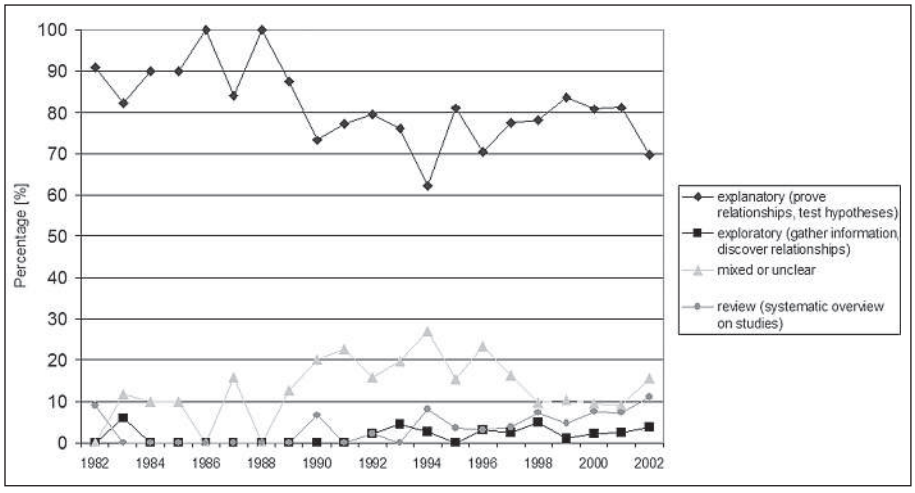

Fig. 6 Percentage of evaluation strategy of evaluation studies 1982-2002 (n=1035)

"medical or other" journals (321 journals). The number of studies which are published in medical informatics journals rose from 11% (n = 4) in the period between 1982 and 1984 to 59% (n = 94) in the period between 1994 and 1996 and is around 50% since then (2000-2002: n = 152, 47%) (Fig. 3). Overall, since 1982, 43% (n = 443) of all studies have been published in medical informatics journals.

Type of Information System

Evaluation studies of expert systems (24% of all studies), teleconsultation systems (20%) and general clinical information systems (15%) cover about 60% of all studies (Table 3).

An analysis over the last 20 years shows mostly stable percentages for the different types of information systems with two exceptions: In the mid 80's, the evaluation of expert systems (XPS) counted for half of the evaluation studies. Since then, their percentage has strongly fallen. Secondly, in the 1990s, evaluation of teleconsultation systems (e.g. teleradiology) started to get more prominent. They count for about one third of all studies between 2000 and 2002.

Study Location

The dominant location of evaluation was the inpatient care unit (e.g. a ward in a non-emergency, non-surgery department), followed by transinstitutional studies (mostly evaluations of telemedical applications), studies in intensive care units, and studies at a general practitioner (Fig. 4).

An analysis over the last 20 years showed rather stable trends. Only, the number of transinstitutional studies has risen from 1% (n = 1) between 1988 and 1990 up to 26% (n = 78) between 2000 and 2002. Similar developments can be found for studies in the intensive care and emergencies areas – their number rose from 5% (n = 2) between 1982 and 1984 to 14% (n = 42) between 2000 and 2002. The number of studies in inpatient care units have decreased during this time from 49% (n = 18) between 1982 and 1984 to 23% (n = 68) between 2000 and 2002, and studies in a laboratory from 11% (n = 4) between 1982 and 1984 to 2% (n = 7) between 2000 and 2002.

Summarizing outpatient care, inpatient care, and transinstitutional care, Figure 5

shows the development of locations over time, showing a clear increase in transinstitutional studies. Since the mid-1990's, nearly 25% (n = 242) of all studies took place in outpatient care areas (patient's home, GP, outpatient unit), 52% (n = 512) in inpatient care areas (intensive care, inpatient care, lab/pharmacy, radiology), 19% (n = 190) focused on trans-institutional care, and 4% (n = 39) were unclear/others.

Evaluation Strategy

Since 1982, most of the studies (79%, n = 815) were clear explanatory studies, i.e. they test a pre-defined hypotheses. Only 2% (n = 23) were exploratory studies, i.e. they explore an unknown area and try to generate hypotheses. Five percent (n = 52) of the papers were systematic reviews, their number slowly rising in the last years. Figure 6 shows the developments of evaluation strategies used over time.

A detailed analysis showed that an exploratory strategy was overrepresented when evaluating organizational and social aspects (10 out of 49 = 20% of such studies on organizational aspects were exploratory studies), usage patterns (6 out of 73 = 8% of such studies were exploratory), and user satisfaction (17 from 215 = 8% of such studies were exploratory).

Evaluation Methods

Throughout the last 20 years, quantitative evaluation methods dominated in evaluation studies: overall, 83% (n = 820) of all studies focused on quantitative methods, only 5% (n = 44) on qualitative methods. The others used combinations of quantitative and qualitative methods or the abstract does not clearly describe the method used. Figure 7 shows the development of evaluation methods over time.

Qualitative methods were mostly used in exploratory studies – 61% (14 out of 23) of all exploratory studies were dominated by qualitative methods – but only 1% of explanatory studies (9 out of 815) use predominantly qualitative methods.

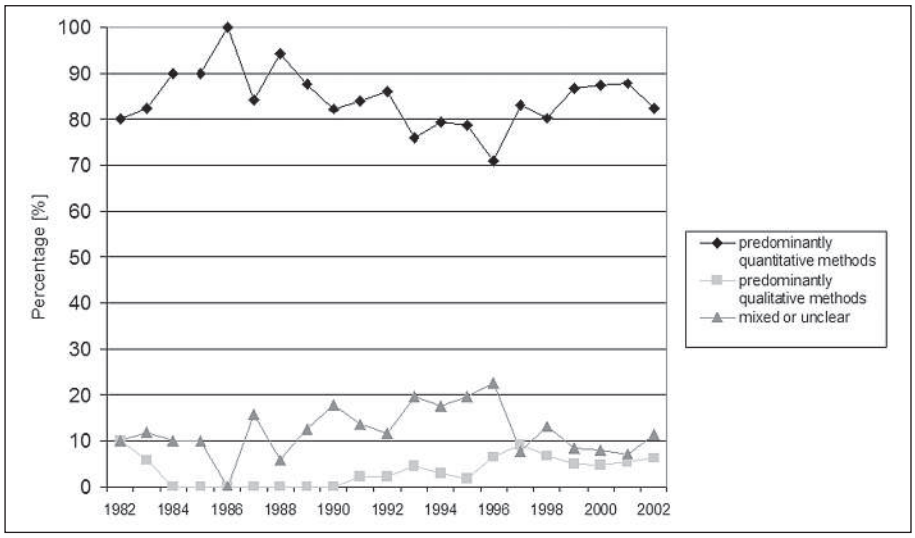

Fig. 7 Percentage of evaluation methods of evaluation studies 1982-2002 (n = 983, reviews excluded)

A detailed analysis showed that qualitative methods were often applied when focusing on the organizational and social impacts of IT (17 out of 49 = 35% of those studies primarily used qualitative methods), on user satisfaction (27 out of 215 = 13%), and on computer knowledge/attitudes (3 out of 28 = 11%).

Study Setting

75% (n = 741) of all evaluation studies (from n = 983, reviews excluded) were field studies, 22% (n = 214) were lab studies. The number of field studies has been slowly rising since the late 1990's from 67% (n = 34) in 1985–87 to 84% (n = 197) in 1997-1999 and to 76% (n = 225) in 2000-2002.

Lab studies predominate in the evaluation of expert systems (114 out of 238 = 48% of such studies were lab studies) and in the evaluation of teleconsultation systems (50 out of 190 = 26% were lab studies). Lab studies were rarely used for the evaluation of GP systems, pharmaceutical systems, or patient information systems (less than 5%

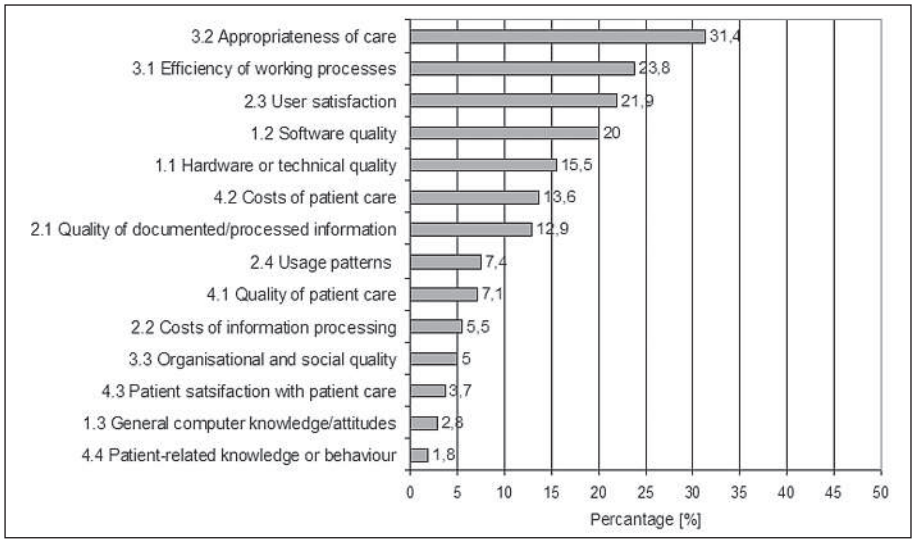

Fig. 8 Percentage of evaluated aspects in evaluation studies 1982-2002 (n = 983, reviews excluded). A study can cover more than one aspect; therefore the sum of percentages is > 100%).

Table 4 Percentage of evaluated aspects in studies 1982-2002 (n = 983, reviews excluded). A study can cover more than one aspect; therefore the sum of percentages is > 100%).

	1982–1984	*1985–1987*	*1988–1990*	*1991–1993*	*1994–1996*	*1997–1999*	*2000–2002*	Overall percen-tage	Overall number of studies
Number of papers published in each period	38	51	79	134	159	248	326	100%	983
	Percentages of evaluated aspects per time period								
1.1 Hardware or technical quality	2.7			12.8	19.1	25.5	15.1	15.5	152
1.2 Software quality	21.6	33.3	44.2	30.1	23.7	11.1	12.1	20.0	197
1.3 General computer knowledge/computer attitudes	5.4	0.0	1.3	2.3	4.6	3.8	2.0	2.8	28
2.1 Quality of documented or processed information	5.4	7.8	10.4	18.0	11.8	13.6	13.1	12.9	127
2.2 Costs of information processing	5.4	5.9	9.1	0.8	9.9	5.5	4.4	5.5	54
2.3 User satisfaction with the component	16.2	21.6	15.6	22.6	26.3	21.3	22.1	21.9	215
2.4 Usage patterns of the components	2.7	5.9	2.6	3.8	9.2	8.9	9.1	7.4	73
3.1 Efficiency of working processes	27.0	15.7	22.1	21.8	23.0	25.5	25.2	23.8	234
3.2 Appropriateness of care	43.2	35.3	27.3	24.8	28.9	30.2	35.6	31.4	309
3.3 Organisational and social aspects	2.7	0.0	1.3	5.3	4.6	7.2	5.4	5.0	49
4.1 Outcome quality of patient care	5.4	2.0	1.3	6.0	9.2	6.4	9.7	7.1	70
4.2 Costs of patient care	10.8	11.8	9.1	12.0	13.8	12.8	16.8	13.6	134
4.3 Patient satisfaction with patient care	2.7	0.0	2.6	3.8	1.3	3.8	5.7	3.7	36
4.4 Patient-related knowledge or behaviour	0.0	0.0	2.6	0.8	1.3	1.7	3.0	1.8	18

of those studies were lab studies). Lab studies were also rarely used in exploratory studies (1 out of 23, 4%), but more frequently in explanatory studies (203 out of 815, 25%).

The highest percentage of lab studies could be found for studies on hardware quality (54 out of 152 = 36% of all hardware studies were lab studies) and software quality (123 out of 197 = 62% were lab studies), while outcome quality of patient care was nearly never evaluated in lab studies (1 out of 70 = 1%).

Focus of Evaluation Study

Over the last 20 years, most publications addressed a single evaluation focus (52%, n = 507). About 23% (n = 294) addressed two, 14% (n = 141) three, and only 4% (n = 41) addressed more than three evaluation foci (i.e., analyzing more than three evaluation criteria). Recently, the mean number of aspects evaluated in a study is slowly rising: from 1.5 between 1982 and 1984 (range: 1-3) to 1.8 between 2000 and 2002 (range: 1-6).

Since 1982, evaluation focused predominantly on appropriateness of care (n = 309) with one third of all studies, followed by efficiency of work processes (n = 234), user satisfaction (n = 215) and software quality (n = 197). Effects on outcome quality of patient care were only considered in 7% (n = 70) of all studies (Fig. 8).

A detailed analysis showed that the percentage of software quality studies fell from a high of 44% before 1990 to 12% today. During this time, hardware quality increased to 15-25% of all studies, mostly as part of studies of telemedical systems. Since the early 90's, appropriateness of care stayed continuously high and is now the most evaluated aspect (about 35% of all studies) (Table 4).

A detailed analysis revealed that hardware quality was often considered during the evaluation of PACS (n = 16, 21% of all PACS studies considered hardware quality) and of teleconsultation systems (n = 120, 63% of all teleconsultation studies). Software quality was often evaluated for expert systems (n = 132, 56% of all expert system studies). Appropriateness of care was important in the evaluation of CPOE (n = 22, 41% of all CPOE studies), GP systems (n = 23, 46%), pharmaceutical information systems (n = 15, 52%), patient information systems (n = 28, 42%), telemonitoring systems (n = 13, 65%), and expert systems (n = 96, 40%). Efficiency of care was dominant in the evaluation of nursing information systems (n = 17, 36% of those studies), PACS (n = 44, 58%), pharmaceutical systems (n = 20, 69%), and RIS (n = 13, 81%). Effects on outcome quality of patient care is frequently considered in the context of evaluation of patient information systems (n = 15, 22% of those studies) and for telemonitoring systems (n = 6, 29%).

When we analyze the developments in the evaluation focus since 1982, we can see that studies on outcome quality of patient care (comprising effects or quality and costs of patient care, patient satisfaction, and patient behavior) rose from around 15% in the 1980's to 35% today, and studies on quality of processes rose from 50% to 65% today (Fig. 9).

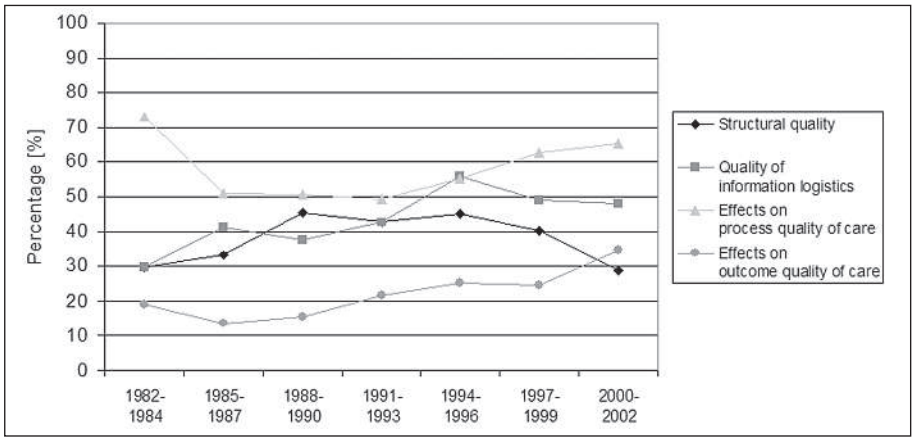

Fig. 9 Percentage of evaluation focus 1982-2002 (n = 983, reviews excluded). A study can cover more than one aspect; therefore the sum of percentages is >100%).

Web-based Inventory of 1035 Evaluation Studies

We decided to make the result of this inventory, i.e. the 1035 identified and classified evaluation studies, available for other researchers. Our classification should allow easily identifying a study, e.g., on a given type of information system, or using defined methods. The database of evaluation studies is available on the web at http://evaldb.umit.at. It allows searching for studies based on the following selection criteria: country, language, type of information system, study location, evaluation strategy, evaluation methods, setting, and evaluation focus. It also allows searching for free text in author, title, journal, and abstracting (Fig. 10).

Discussion

Discussion of Methods

There have been several partly comparable literature reviews in the last years. Moorman [27] analyzed PubMed publications on Electronic Patient Records published between 1991 and 2002 with regard to journal, origin and MeSH term, but he did not focus only on evaluation studies. Sawyer [17] analyzed publications on information systems between 1990 and 2001 with regard to evaluation construct (information, technology, or people), evaluation level (team or institution), and evaluation strategy (experimental/explanatory, intensive/exploratory, or computational). However, his research did not concentrate on IT in health care, but on information system literature in general.

Our systematic review is mostly modeled after van der Loo [19] from 1994. He analyzed 108 evaluation studies with regard to type of system, study design, data collection, economic evaluation type, and type of effect measure. Our analysis differs in several ways: It was done nine years after van

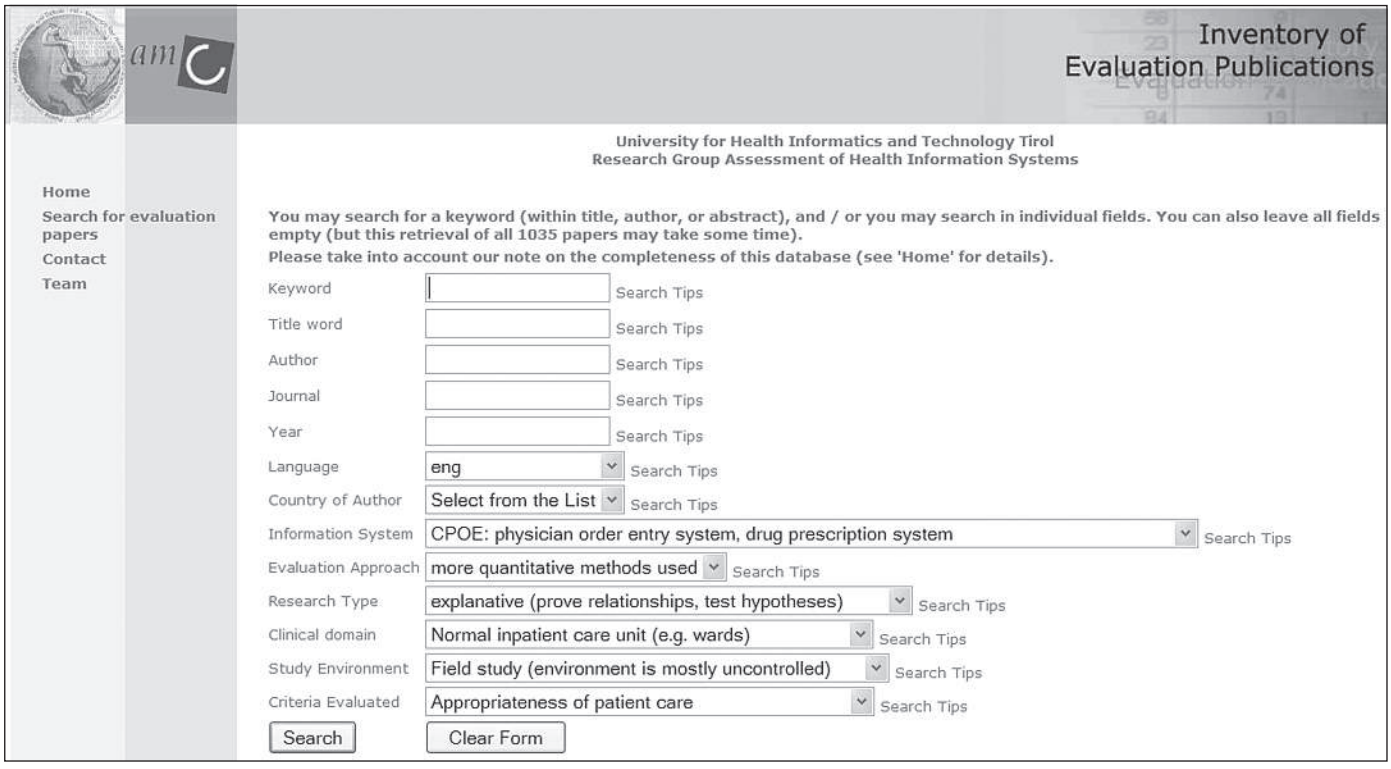

Fig. 10 Screenshot of the search module of the web-based inventory of evaluation studies (http://evaldb.umit.at)

der Loo, thus including the most recent work. We used a more detailed classification of information system type and an extended list of evaluation criteria, for a more detailed analysis. We also added information on evaluation strategy and methods used. We found a larger number of studies than van der Loo, as we used broader inclusion criteria (e.g. including lab studies). We also emphasized trends in evaluation research in the last 20 years.

Our inventory concentrated on those attributes of evaluation studies which could be identified based on the abstract. Other than van der Loo, we did not try to describe study design, study methods or results in detail, or check the quality of studies. Instead we analyzed major trends in evaluation research and therefore favored a restricted classification e.g. methods (quantitative vs. qualitative) or setting (field vs. lab study). In most cases, we found that this information can be found even when only looking on the abstracts (and not on the full papers). However, it is the aim of a subsequent study to take a random sample of the studies, to further investigate in depth based on full papers which evaluation methods and aspects are used (similar to Fletcher and Fletcher [28]), and what the quality of the study is (similar to Hunt [26]).

Our classification of information systems was based on existing literature and was tested and refined while doing the inventory. As definitions and terminology in medical informatics may have changed over the last 20 years, we invested heavily in properly defining and testing the categories. Any inclusion/exclusion as well as classifying was done together by both authors. However, our classification may not always be conclusive. For example, qualitative and quantitative methods are extremes on a continuum of methods rather than clear-cut classification criteria. A formal test of consistency across classes was not conducted, as our decisions on inclusion and on classifying were checked and revised several times when discrepancies arouse, each time refining the instruction and definitions of the classes. We repeated this until no discrepancies between both authors were left.

We cannot be sure that we found all available studies in our search. First, potential variation of terms in title, abstract or MeSH-headings is much too diverse to cover all possible combinations describing evaluation studies. Second, terminology in medical informatics changed over the last 20 years (e.g., talking about automated systems or computer-based protocols in the 1980's). Third, the indexing of medical informatics studies in PubMed seems at times insufficient. Fourth, we tried to optimize our query to get maximum recall; however, this resulted in a rather low precision of only around 7%, increasing the danger of overlooking evaluation papers during the manual check. Finally, many evaluation studies may not have been published at all. This so-called publication bias may have also led to an over-representation of experimental studies in our inventory (compare e.g. discussion in [29]).

Due to often missing abstracts in older papers within PubMed, we may have missed especially older studies. We finally decided to start our search in 1982 as the first year where we were able to identify at least ten studies having an abstract – fully conscious of the fact that by choosing this limit, we excluded some well-known earlier studies such as Simborg [14] or McDonald [30].

We checked the recall rate of our search by taking other review articles as gold standards. Here we found rates of 73% and higher. In many cases, particularly in older papers, we found that the abstracts were too vague or incomplete to decide on the inclusion criteria, in which case the paper was excluded. In any case, it must be taken into account that we just included published evaluation studies that have an abstract. Therefore, all results must be taken with care and cannot easily be generalized to evaluation research in medical informatics in general.

Discussion of Results

The number of studies identified increased in the recent years. This could indicate a rising significance of evaluation studies. However, evaluation studies still cover only around 1% of all medical informatics publications indexed in PubMed. The drop in the number of evaluation studies we found in 2002 may be caused by the incomplete documentation of 2002-studies in PubMed at the time of our search.

Nearly 50% of all 1035 identified studies have been published in only 21 journals. In recent years, there has been a strong shift from medical or other journals to medical informatics journals which is not too surprising as many medical informatics journals only started to appear in the 1990's. The most prevalent publication medium was the AMIA-Proceedings (with 9.5% of all studies). Here we must take into consideration that these conference proceedings publish a large number of papers annually. After a boom in the early 1990's the prevalence of proceedings such as AMIA or Medinfo is now declining. The other dominant journal was the Journal of Telemedicine and Telecare (7.2% of all studies), indicating a rising dominance of telemedical studies in recent years (nearly one third of all studies in recent years were telemedical studies). The dominating language of all studies is English (reflecting the dominance in PubMed, and the dominance of US-American papers).

Explanatory research (90% of all studies) dominated evaluation studies in the last 20 years, most of them relying on quantitative data (90% of those studies). The dominance of quantitative methods in explanatory studies is not surprising, as the verification of statistical hypotheses needs quantitative data. On the other hand, 60% of exploratory studies relied on qualitative methods. Not surprisingly, qualitative methods were mostly used for the evaluation of organizational and social impacts of IT. It should be noted that our classification was based on the predominant methods used – however, most quantitative studies include some qualitative methods (e.g. free text fields in a standardized questionnaire), and the other way round.

The bulk of evaluation studies still address inpatient care (around 40%). Since the mid-1990's, the number of trans-institutional studies increased to 30% today, reflecting the trend towards cooperative and shared care. The number of field studies has been steadily rising (currently around 75% of all studies). Of all exploratory studies, 96% were done in the field, which is not surprising, as the detection of unknown rela-

tionships and the generation of new hypothesis only make sense in a realistic environment.

The most common focus of evaluation studies was appropriateness of patient care, efficiency of patient care, user satisfaction, and software quality. Outcome quality of patient care was only evaluated in around 7% of all studies but have been increasing in recent years and cover 10% of all studies in 2000-2002. Organization and social aspects were evaluated in around 5% of all studies most recently, with rising tendency.

The evaluation of teleconsultation systems and expert systems often addresses hardware and software quality. Evaluation of CPOE, GP information systems, telemonitoring systems, and patient information systems often focus on outcome quality of patient care. Here, the researcher seems to anticipate a direct influence on the quality and costs of patient care. For CPOE, costs are frequently evaluated, and for telemedical systems quality and costs are considered important, while in patient information systems changes in the behavior of the patients are of greatest interest. Apart from those findings, user satisfaction, efficiency of patient care and appropriateness of patient care are the most frequently addressed evaluation criteria.

There is a slowly growing trend towards evaluation studies covering more than one evaluation aspect (e.g. one paper reports a study addressing e.g. software quality, user acceptance, effects on efficiency of care, and effects on outcome of care). Whether the studies themselves really get broader in their focus, or whether just the publication strategy changes (summarizing several substudies in one paper), cannot be analyzed based on the given data.

It is interesting to compare our results with van der Loo [19]. He analyzed 108 studies which have been published between 1967 and around 1995. With regard to the evaluation criteria, he found "performance of user" in 29% of all studies – this seems to be comparable to appropriateness of care for which we found 31%. He found time savings ("time personnel + time processes") in 29%, whereas we found efficiency of patient care in 24%. He found "costs of patient care" in 15%; we found this in 14% of all studies. Larger differences can be found when comparing "user satisfaction" (he found it in 11%, we found it in 22%), and "patient outcome" (he found it in 21% of all studies, we found it in 7%). Reasons for those differences are not clear and unexpected since we observed a rising attention to patient outcome in the recent years and our study includes more recent evaluation studies than van der Loo.

It must also be taken into account that we used PubMed exclusively, which is dominated by journals from the positivistic or objectivistic research tradition. Restricting our analysis to PubMed may also have produced a bias with regard to U.S. publications. Thus, we might find other studies or a larger number of exploratory or qualitative studies if we include other databases (e.g. Embase or Social Science Index). It is planned, as a next step, to analyze other databases to extend the inventory and to verify the trends we have detected so far.

What are the practical applications of our results? First, our inventory presents a comprehensive overview of evaluation papers in the last 20 years, therefore updating and extending earlier work (e.g. by van der Loo). Second, a deeper insight into historical developments in published evaluation studies is an important input into the ongoing discussion on best methods and approaches within evaluation research. It e.g. shows that even when qualitative methods and exploratory approaches get more and more accepted, this is not yet reflected in the literature of published studies. Third, we believe that our multi-axial classification which found its basis in existing literature, has an additive value in describing evaluation studies. Finally, classifying the identified abstracts did not only help to detect those historical trends, but can also – supported by the available database – help researchers search for certain studies, learn from earlier work, and avoid replicating studies which are already available.

Conclusion

We found interesting developments in evaluation research in the last 20 years. For example, there has been a shift from medical or other journals to medical informatics journals. Also, the number of systematic reviews is steadily rising. In addition, the evaluation of expert systems decreased in prevalence, while that of telemedical systems increased, reflecting rising interest towards cooperative, shared care. The number of field studies is also steadily rising, and the published studies slowly become broader in focus.

On the other hand, we also identified rather stable trends. For example, explanatory research and quantitative methods dominated evaluation studies in the last 20 years which does not yet reflect the rising awareness of the benefits of qualitative methods. How much those findings are biased due to publication bias cannot be analyzed based on the available data.

In order to help other researchers accessing evaluation studies, and to reproduce our review, we developed a web-based interface to allow searching our database. It is available at http://evaldb.umit.at without charge.

Rigby [31] noted that the focus of evaluation of an information system changes during its life cycle: While during the implementation phase, evaluation addresses technical aspects, it later shifts to impacts on patient care, and then to impacts on the overall organization during routine use of an information system. We found comparable developments in medical informatics: Since 1982, the number of lab studies and of studies focusing on technical aspects has declined, while studies focusing on quality of care processes and patient outcomes have increased. We interpret this shift as a sign of maturation of evaluation research in medical informatics

Acknowledgment
We thank Frieda Kaiser and Megan Stepushyn for their contribution to the website and the analysis. A short version of this paper was presented at Medinfo 2004.

References

1. Beynon-Davies P, Lloyd-Williams M. When health information systems fail. Top Health Inf Manage 1999; 20 (1): 66-79.
2. Forsythe DE, Buchanan BG. Broadening our approach to evaluating medical information systems. In: Clayton P, editor. 15th Annual

Symposium on Computer Applications in Medical Care. New York: McGraw-Hill; 1992. pp. 8-12.
3. Friedman C, Wyatt JC. Evaluation Methods in Medical Informatics. New York: Springer; 1997.
4. Heathfield H, Peel V, Hudson P, Kay S, Mackay L, Marley T, et al. Evaluating Large Scale Health Information Systems: From Practice Towards Theory. In: Masys D, editor. AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus; 1997. pp. 116-20.
5. Moehr JR. Evaluation: salvation or nemesis of medical informatics? Comput Biol Med 2002; 32 (3): 113-25.
6. Ammenwerth E, Brender J, Nykänen P, Prokosch H-U, Rigby M, Talmon J. Visions and strategies to improve evaluation of health information systems – reflections and lessons based on the HIS-EVAL workshop in Innsbruck. Int J Med Inf 2004; 73 (6): 479-91.
7. Anderson JG, Aydin CE, Jay SJ, editors. Evaluating Health Care Information Systems – Methods and Applications. London, New Delhi: Sage Publications; 1994.
8. van Gennip E, Talmon J, editors. Assessment and evaluation of information technologies in medicine. Studies in Health Technology and Informatics vol. 17. Amsterdam: IOS Press; 1995.
9. Whitten P, Mair F, Haycox A, May C, Williams T, Hellmich S. Systematic review of cost effectiveness studies of telemedicine interventions. BMJ 2002; 324: 1434-7.
10. Haux R, Ammenwerth E, Herzog W, Knaup P. Health Care in the Information Society: A Prognosis for the Year 2013. Int J Med Inf 2003; 66: 3-12.
11. Tierney W, Overhage J, McDonald C. A Plea for Controlled Trials in Medical Informatics. J Am Med Inform Assoc 1994; 1 (4): 353-5.
12. Heathfield H, Buchan I. Current evaluations of information technology in health care are often inadequate. BMJ 1996; 313 (7063): 1008.
13. Kaplan B. Evaluating informatics applications – some alternative approaches: theory, social interactionism, and call for methodological pluralism. Int J Med Inform 2001; 64: 39-56.
14. Simborg D, McDonald L, Liebman J, Musco P. Ward information management system – an evaluation. Comput Biomed Res 1972; 5 (5): 484-97.
15. PubMed. National Library of Medicine. http://www.ncbi.nlm.nih.gov.
16. Krobock JR. A taxonomy: hospital information systems evaluation methodologies. J Med Syst 1984; 8 (5): 419-29.
17. Sawyer S, Chen T. Conceptualizing Information Technology in the Study of Information Systems: Trends and Issues. In: Myers M, Whitley E, Wynn E, DeGross J, editors. Global and Organizational Discourse about Information Technology. London: Kluwer; 2002. pp 1-23.
18. Grémy F, Degoulet P. Assessment of health information technology: which questions for which systems? Proposal for a taxonomy. Med Inform (Lond) 1993; 18 (3): 185-93.
19. van der Loo R. Overview of Published Assessment and Evaluation Studies. In: van Gennip EMSJ, Talmon JS, editors. Assessment and evaluation of information technologies. Amsterdam: IOS Press; 1995. pp. 261-82.
20. Ammenwerth E. Die Bewertung von Informationssystemen des Gesundheitswesens [The evaluation of health information systems]. Publication Series No 1 of UMIT. University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Austria. Aachen: Shaker; 2004.
21. Bortz J, Döring N. Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Berlin: Springer; 2002.
22. Balas EA, Austin SM, Mitchell JA, Ewigman BG, Bopp KD, Brown GD. The clinical value of computerized information services. A review of 98 randomized clinical trials. Arch Fam Med 1996; 5: 271-8.
23. Kaplan B. Evaluating informatics applications – clinical decision support systems literature review. Int J Med Inf 2001; 64: 15-37.
24. Roine R, Ohinmaa A, Hailey D. Assessing telemedicine: a systematic review of the literature. Canad Med Assoc J 2001; 165 (6): 765-71.
25. Friedman C. Is Medical Informatics a mature science? Int J Med Inf 2003; 69: 261-72.
26. Hunt D, Haynes R, Hanna S, Smith K. Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes. JAMA 1998; 280 (15): 1339-45.
27. Moormann P, van der Lei J. An Inventory of Publications on computer-based medical records: an update. Methods Inf Med 2003; 42: 199-202.
28. Fletcher RH, Fletcher SW. Clinical research in general medical journals: a 30-year perspective. N Engl J Med 1979; 301 (4): 180-3.
29. Tierney W, McDonald C. Testing Informatics Innovations: The Value of Negative Trials. J Am Med Inform Assoc 1996; 3 (5): 358-9.
30. McDonald C. Protocol-based computer reminders, the quality of care and the non-perfectability of man. 1976; 295 (24): 1351-5.
31. Rigby M. Health informatics as a tool to improve quality in non-acute care – new opportunities and a matching need for a new evaluation paradigm. Int J Med Inf 1999; 56: 141-50.
32. Haux R, Kulikowski C. Yearbook of Medical Informatics (annual edition). Stuttgart: Schattauer; Since 2001. http://iig.umit.at/yearbook.

Correspondence to:
Ass.-Prof. Dr. Elske Ammenwerth
UMIT – University for Health Sciences,
Medical Informatics and Technology
Institute for Health Information Systems
Eduard-Wallnöfer-Zentrum 1
6060 Hall in Tyrol
Austria
E-mail: elske.ammenwerth@umit.at