journal homepage: www.intl.elsevierhealth.com/journals/ijmi

The quality of evidence in health informatics: How did the quality of healthcare IT evaluation publications develop from 1982 to 2005?

N.F. de Keizer[a,*], *E. Ammenwerth*[b]

[a] *Department of Medical Informatics, J1b-114, Academic Medical Centre, Universiteit van Amsterdam, P.O. Box 22700, 1100 DE Amsterdam, The Netherlands*
[b] *Institute for Health Information Systems, UMIT-University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Austria*

ARTICLE INFO

Article history:
Received 10 July 2006
Received in revised form
28 November 2006
Accepted 28 November 2006

Keywords:
Information management
Evaluation studies
Quality of publications
Review

ABSTRACT

Objective: To obtain an overview of study designs and study methods used in research evaluating IT in health care, to present a list of quality criteria by which all kinds of reported evaluation studies on IT systems in health care can be assessed, and to assess the quality of reported evaluation studies on IT in health care and its development over time (1982–2005).
Methods: A generic 10-item list of quality indicators was developed based on existing literature on quality of medical and medical informatics publications. It is applicable to all kind of IT evaluation papers and not restricted to randomized controlled trials. One hundred and twenty explanatory papers evaluating the effects of an IT system in health care published between 1982 and 2005 were randomly selected from PubMed, the study designs and study methods were extracted, and the quality indicators were used to assess the quality of each paper by two independent raters.
Results: The inter-rater variability of scoring the 10 quality indicators as assessed by a pre-test with nine papers was good ($K = 0.87$). There was a trend towards more multi-centre studies and authors coming more frequently from various departments. About 70% of the studies used a design other than a randomized controlled trial (RCT). Forty percent of the studies combined at least two different data acquisition methods. The quality of IT evaluation papers, as defined by the quality indicators, was only slightly improving in time (Spearman correlation coefficient $[r_s] = 0.19$). The quality of RCTs publications was significantly higher than the quality of non-RCT studies ($p < 0.001$).
Conclusion: The continuous and dominant number of non-RCT studies reflects the various approaches applicable to evaluate IT systems in health care. Despite the increasing discussion on evidence-based health informatics, the quality of published evaluation studies on IT interventions in health care is still insufficient in some aspects. Journal editors and referees should take care that reports of evaluation on IT systems contain all aspects needed for a sufficient understanding and reproducibility of a paper. Publication guidelines should be developed to support more complete and better publications of IT evaluation papers.

[*] *Corresponding author*. Tel.: +31 20 566 5205; fax: +31 20 691 9840.
E-mail address: n.f.keizer@amc.uva.nl (N.F. de Keizer).

1386-5056/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2006.11.009

1. Introduction

The implementation of information technology (IT) is often advocated as being able to make health care more effective and efficient (e.g. [1]). Despite an accumulation of successful implementations, a not negligible amount of studies showed negative effects of IT implementation (e.g. [2,3]). Therefore, evaluation of the effects of IT interventions on health care is essential. 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 [4]. Evidence-based health informatics [5], meaning that all decisions with regard to IT systems should be based on available scientific evidence, need systematically conducted and published evaluation studies. Consequently, evaluation studies are increasingly considered as an integral part of planning, development, introduction and operation of IT in health care [6–8], and a rising number of evaluation studies is being performed [9]. However, evaluation studies can only form the basis of evidence-based health informatics when they are adequately described in publications so that readers can reasonably evaluate the study in the context of existing information.

Systematic reviews try to give an overview on the available evidence in a given field, e.g. on telemedical systems [10,11], on decision support systems [12], or on patient-owned medical records [13]. However, several review authors have found the quality of reported evaluation studies to be insufficient which makes it hard to draw conclusions about the general effects of an IT system.

Typical problems that are noted by authors of systematic reviews are, e.g. insufficient description or lack of the control situation [11], use of inadequate designs or inadequate methods [11,13–15], limited power of studies, or inconsistent results of studies [12]. The problem of inadequate and incomplete reporting of IT evaluation studies has also been discussed at previous European workshops, leading to the initiative to develop standards for reporting on IT evaluation studies in health care [4].[1]

Also in other fields such as medicine research points to the fact that medical publications vary considerably in their quality of reporting [16–19].

To improve reporting of evaluation results in medicine, guidelines have been developed. One well-known example is the CONSORT group that developed a statement consisting of 21 items that should be included in a study report [19]. This CONSORT statement has been adopted by BMJ, JAMA, Lancet and some other journals [17,20]. However, CONSORT is only restricted to the reporting of randomized controlled trials (RCT) and does not help with other kinds of study design such as quasi-experimental or qualitative designs. Now, in health informatics, there are many situations where an RCT is either not helpful or not feasible. For example, to explore a setting or to identify influencing user acceptance factors, other approaches such as qualitative case studies, user surveys or action research are better suited [21,22]. Sometimes, even when an RCT would be the best approach, it may be not feasible due to a low number of participants or the difficulty of undertaking true randomisation. In that case the researcher may have to choose other quasi-experimental designs such as controlled before-after trials or time-series analysis [23].

Summarizing, the quality of evaluation papers in health informatics is often seen as being insufficient, but this has not been systematically assessed yet. A more systematic analysis of the quality of IT evaluation studies (not only RCT studies, but all IT evaluation studies) would help to better understand frequent problems of evaluation reports and would form the basis to develop comprehensive guidelines for IT evaluation papers. As a consequence this would support better and more complete evaluation papers that form the basis for evidence-based health informatics. As CONSORT is restricted to the quality of reporting RCTs and as many other study designs than RCTs are used in evaluating IT interventions in health care, an adapted set of quality indicators is necessary.

This study has therefore two objectives. First, to get an overview of the major (quantitative and qualitative) study designs and study methods used in IT evaluation studies in health care and their development over time. Second, to assess the quality of reported IT evaluation studies in health care and its development over time based on a list of quality indicators by which all kinds of reported IT evaluation studies (not only RCT) can be assessed.

2. Methods

2.1. Selection of evaluation studies

This study has been based on the systematic literature search in PubMed extensively described in [9]. For this query, we used a combination of MeSH Headings and title words to search for all types of evaluation studies on 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 care to process patient-related data, information or knowledge. We restricted the papers on health information systems to those that primarily concern the evaluation of the system. We defined an *evaluation study* as the systematic, empirical assessment of a component of a health information system. We did not include editorials, letters and papers that only describe a study design, or that just contain system descriptions. The complete PubMed query is available from the authors. The search included papers from 1982 to 2005. All 1.258 references and abstracts are available at http://evaldb.umit.at.

We divided the period 1982–2005 into eight periods of 3 years. For each 3-year period we randomly selected 15 full papers using the random sample selector of SPSS statistical software 12.0.2. We restricted our selection on summative studies that evaluated effects of IT on quality of care processes (e.g. time needed for care, appropriateness of care, quality of communication and cooperation) or outcome of care (e.g. effect on mortality, morbidity, patient satisfaction or costs) from the 1.258 abstracts found in the database. The 120 (8 × 15) papers selected for this study can also be

[1] http://iig.umit.at/efmi/stare-hi/Stare-HIv0.12.doc.

found at http://evaldb.umit.at (choose "Search", then select "Only papers of Quality Review 1982–2006" from the list in the "Search" field).

2.2. Classification of study designs and methods

For each of the 120 selected papers the study design was described by the following aspects:

- *Single-unit study*, *single-centre study*, or *multi-centre study*. Typically, single-centre and multi-centre studies are distinguished in the literature, with multi-centre studies defined as a "trial conducted at several geographical sites" [24]. In IT evaluation studies, however, studies are often conducted at pilot units and not at one or more hospitals. To reflect this, we used the following classification:
  - Single-unit study: The study was conducted only in one distinct unit within a larger health care organization (e.g. one nursing ward, one intensive care unit).
  - Single-centre study: The study was conducted in one organizational centre (e.g. a hospital, a general practitioner practice, an outpatient clinic).
  - Multi-centre study: The study was conducted in several organizational centres that are geographically distributed (e.g. several general practitioners, several hospitals).

  When it was explicitly stated that an individual GP, instead of a group GPs within one practice, was involved in an evaluation study this study was classified as single-unit. Note that telemedical applications were classified as single-centre in the case were there was only one major centre offering the telemedical service (e.g. several GPs using the telemedical service of one hospital).
- *Prospective or retrospective data collection*. This describes the time of data collection in relation to the study period and gives an indication for the quality of the data:
  - Prospective trial: The data were collected after the study began.
  - Retrospective trial: The data were collected before the study began.
- *Experimental or quasi-experimental design*. According to Harris [23], quasi-experiments are studies that aim to evaluate interventions but that do not use randomization. This includes, e.g. before-after-studies, time-series analysis or non-randomly selected control groups (e.g. by matching). We therefore distinguished:
  - Uncontrolled trial: No explicit control group is given (e.g. post-test analysis of effects of system based on user survey, or analysis of organizational impact based on qualitative methods).
  - Non-random controlled trial (quasi-experiments): An intervention group is compared to a non-random control group (e.g. pre-test versus post-test trial, time-series analysis, matched-pair design).
  - Randomized controlled trial (RCT, experiment): Participants are randomized to intervention and control group.

In addition, to get a feeling for the size of an evaluation project and the interdisciplinarity of the researchers, we counted the number of authors and the number of distinct departments they came from.

We described the data acquisition methods used in a study based on the following list. This list is based on a comparable list used by van der Loo [25]:

- Questionnaires.
- Interviews.
- Observations.
- Documentation analysis, chart review.
- Automated logging of usage data.
- Work sampling, time measurements.

2.3. Assessment of paper quality

The list of quality indicators to assess the quality of IT evaluation papers was developed after thorough review of earlier work on quality indicators. CONSORT [19], Bath et al. [16], Mihan and Windeler [26] and others were analysed to check whether indicators used to report on RCTs can be transferred to a checklist appropriate for other quasi-experimental study designs such as controlled before–after trials and for uncontrolled trials. We found that many indicators used here can be generalized: for example, blinding of participants is a typical technique that aims at increasing credibility of results. Instead of asking for blinding in our study, we tried to assess the credibility of the results. Another example of the CONSORT criteria is that the intervention should be described. We tried to assess whether the IT system, considering it as an intervention, is described in sufficient detail. Defining the statistical analysis it is important to judge the adequate use of methods in quantitative analyses. Since our study also includes qualitative analyses we tried to assess whether the methods used, both quantitative and qualitative, were adequate to answer the study question. We then reviewed papers analysing the quality of controlled trials (both RCT and non-RCT) such as Mair and Whitten [15], Downs and Black [27], Johnston et al. [28], Roine et al. [29], Verhagen et al. [30], Hall et al. [31]. We also reviewed quality indicators used in the qualitative research domain as discussed by Greenhalgh and Taylor [32] and Borreani et al. [33]. Finally, we reviewed guidelines for authors and reviewers as used by major health informatics journals and by the IMIA Yearbook of Medical Informatics [34].

It was important to find quality indicators that were applicable not only to randomized controlled quantitative trials, but to all kinds of studies (e.g. qualitative case studies, longitudinal descriptive studies, uncontrolled clinical trials, etc.) as described above.

The resulting list of 10 quality indicators is presented in Table 1. Note that the quality indicators primarily aim at measuring the quality (and completeness) of reporting the evaluation study instead of the quality of the study itself. The quality of a study itself often cannot completely be assessed only based on its publication. Nevertheless some criteria assess (in some way) the quality of the study itself, e.g. "Q7. Methods seem adequate to answer study question". We believe this quality indicator is important to include since many study designs (not only RCTs) and methods are used in evaluation studies on IT interventions in health care. A description of the choice of the study design and methods used as well as their motivation is therefore included in the set of quality indicators.

Table 1 – Ten items that should be included in reports of evaluation studies on IT interventions in health care

	Paper heading	Description
1	Introduction	*Motivation, problems and study questions are clearly described*: Does the study have a scientific basis with relevant literature references and a clear study objective?
2	Introduction/Methods	*The evaluated information technology (the intervention), is described in sufficient detail*: Is the information technology under study sufficiently described including hardware, software, position of the information system in the total information infrastructure, functionality that is available and functionality that is really used, number and types of regular users, usage patterns, age/maturity of the technology, integration into workflow? Is timing and procedure of the intervention (e.g. a new IT system, a new IT function) described in sufficient details?
3	Methods	*Type, number, and sampling of involved study population are clearly described*: What is the unit of analysis (e.g. patients, doctors, departments)? Is the number and type of participants clearly indicated (e.g. junior physicians, only outpatient units)? How are the subjects chosen and recruited? Are inclusion and exclusion criteria clear?
4	Methods	*Setting and population seem justified to answer study question*: Is the unit of analysis appropriately chosen to answer the study question? Is the setting in which this study took place sufficient representative and appropriate to answer the study question?
5	Methods	*Methods for collecting data are sufficiently clear*: Is clearly described how data is collected (e.g. interview, focus groups, document analysis, extraction from patient record, time measurement, etc.) and whether it was collected retrospectively or prospectively? Were validated measures used or new methods? Who collected the data, was this person independent? Is reliability and validity of the use tools addressed?
6	Methods	*Methods for analysing data are sufficiently clear*: Are methods for analysing qualitative data clearly described (who did the analysis, what is the role of this person, were accepted methods used and referred to)? Were exact statistical tests used to analyse quantitative data? Are confidence intervals, *p* values, measures of variation given where appropriate? If only a subset of all data is presented, is it clear how this was selected? Is it clear how incomplete data were dealt with?
7	Methods	*Methods seem adequate to answer study question*: Was an appropriate study design used to answer study question? Do the chosen methods provide sufficient valid data to answer the study questions? Is triangulation of methods and data used?
8	Methods/Results	*Any comparison that is done between groups is fair*: Are the groups comparable with regard to baseline data? Are any differences here discussed? Is verified that any later differences in groups are only due to the IT intervention and not to external other factors such as staff changes?
9	Results	*All results are credible and seem valid*: Do the results give an answer to the initial research questions? Have the results been presented in a credible way? Is all data presented and interpreted? Are objective and subjective findings distinguished clearly?
10	Discussion (Conclusion)	*All conclusions seem justified by the results*: Have the results been well summarized? Have the results been well interpreted and reflected in the context of existing literature? Have the weaknesses of the study been mentioned? Is it made clear whether the results are transferable to others settings?

For each indicator, 0–2 points (0 = not fulfilled, 1 = partly fulfilled, 2 = fully fulfilled) were assigned, therefore a maximum of 20 points could be obtained as an overall quality score for one study publication. Mean scores between different groups of papers, e.g. to compare quality scores between different study designs, were compared by Student's *t*-test. Spearman rho correlation was calculated to test the development of quality in time.

To calculate inter-rater reliability for the quality scores and to test whether the indicators were clear to both reviewers, the first nine randomly selected papers were independently analysed in a pre-test using Kappa statistics. After calculating the Kappa on this pre-set and discussing unclear items left, one of the raters scored the other 111 papers and the other rater commend on these scores. Discussions continued until consensus was reached.

SPSS 12.0.2 was used to perform all statistical analyses.

3. Results

3.1. Study design of IT evaluation papers

The study designs of all 120 papers were analysed. As shown by Fig. 1, the percentage of single-unit and single-centre studies strongly decreased during the last years. In the period 2002–2005 almost half of all selected papers reported on multicentre studies. Although for each 3-year period 15 papers were

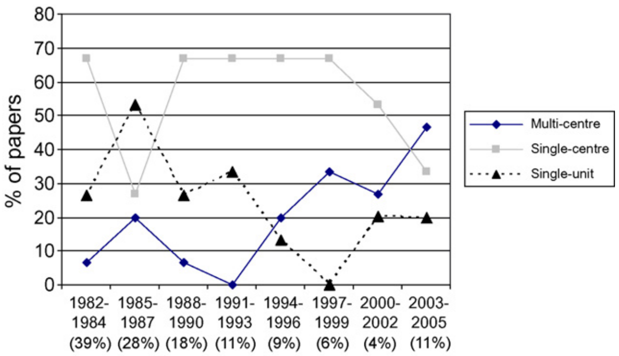

Fig. 1 – Percentage of IT evaluation studies performed in single-unit, single-centred or multi-centred trials 1982–2005 ($n = 15$ per 3-year period.) The percentage selected studies from each time period (15/total number of evaluation studies in 3-year period) is presented bracketed on the X-axis.

selected, the original database contains different numbers of evaluation studies per 3-year period. The percentage selected studies for each 3-year period (15/total number of evaluation studies in 3-year period) is presented bracketed on the *X*-axis.

Overall, 79% of the studies used prospectively collected data; this percentage remained stable over the last 24 years. Of all evaluation studies, 16% ($n = 19$) did not use any control group, 55% ($n = 66$) used non-random controls and 29% ($n = 35$) used random controls (Fig. 2).

As shown in Fig. 3 45% ($n = 54$) of the papers were written by 1–3 authors, 42% ($n = 50$) have 4–6 authors, and 13% ($n = 16$) have more than 6 authors. Thirty percent of the papers have been written by authors of the same department. During the years the ratio of papers written by authors from various departments has strongly increased (mean number of departments involved in 1982–1984 was 2.2 versus 3.6 departments in 2002–2005).

A majority ($n = 79$, 66%) of the studies used document analysis/chart review as the main data acquisition method. Forty percent of the studies ($n = 48$) combined at least two different data acquisition methods, 10 studies (8%) combined three data sources, and only one study (0.8%) combined more than three different data sources for answering their study questions. Especially in the last 3 years the use of multiple data acquisition methods increased (mean number of data acquisition methods between 1982 and 2002 was rather constant around 1.4 versus 1.7 in 2002–2005). Table 2 shows the details of the chosen methods.

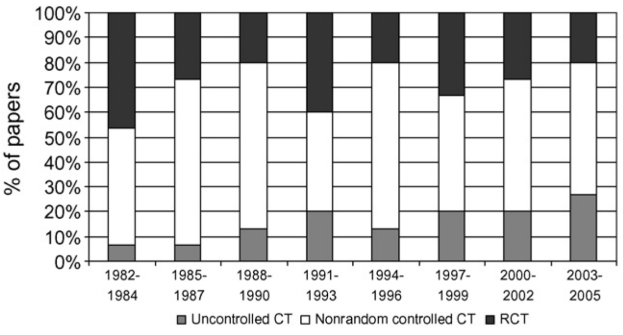

Fig. 2 – Distribution of different study designs used in IT evaluation studies 1982–2005 ($n = 15$ per 3-year period).

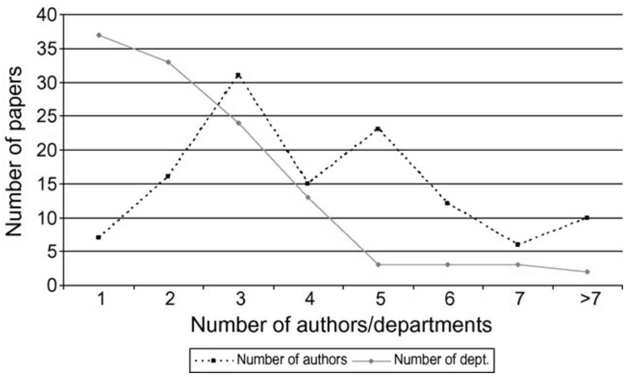

Fig. 3 – Number of authors and number of distinct departments the authors come from of IT evaluation studies 1982–2005 ($n = 120$).

3.2. *Quality of IT evaluation papers*

The inter-rater variability of scoring the 10 quality indicators in the pre-test of nine papers was good: the averaged Kappa (the average of the 10 Kappa's per quality indicator) was 0.87. The right column of Table 3 presents the Kappa-values per indicator. In the other 111 papers there was no disagreement on quality indicators "Q1. Motivation, problems and study questions are clearly described" and "Q2. The evaluated information technology (the intervention), is described in sufficient detail". In less than five papers the quality indicators "Q3. Type, number, and sampling of involved study population are clearly described", "Q4. Setting and population seem justified to answer study question", "Q5. Methods for collecting data are sufficiently clear", "Q6. Methods for analysing data are sufficiently clear", "Q7. Methods seem adequate to answer study question", "Q9. All results are credible and seem valid" and "Q10. All conclusions seem justified by the results" were discussed. We mostly (12 times) discussed disagreements on the

Table 2 – Data acquisition methods of evaluation studies 1982–2005 ($n = 120$)

Methods used	Number of studies using the indicated methods (multiple nominations possible)
Documentation analysis, chart review	79 (65.8%)
Questionnaires	38 (31.7%)
Work sampling, time measurements	16 (13.3%)
Interviews	14 (11.7%)
Observations	13 (10.8%)
Automated logging of usage data	4 (3.3%)
Other/unclear	16 (13.3%)

Table 3 – Mean, standard deviation, incidence of scoring values (0–2), and Kappa value on the pre-test for each quality indicator of IT evaluation papers 1982–2005 ($n = 120$)

Quality criteria	Mean	Standard deviation	Scoring value			Kappa
			0	1	2	
1. Motivation, problems and study questions are clearly described	1.78	0.46	2	23	95	0.89
2. The evaluated information technology (the intervention), is described in sufficient detail	1.39	0.75	19	35	66	0.53
3. Type, number, and sampling of involved study population are clearly described	1.43	0.63	9	50	61	0.88
4. Setting and population seem justified to answer study question	1.92	0.28	0	10	110	1
5. Methods for collecting data are sufficiently clear	1.43	0.65	10	48	62	0.76
6. Methods for analysing data are sufficiently clear	1.35	0.76	21	36	63	0.88
7. Methods seem adequate to answer study question	1.71	0.53	4	27	89	0.88
8. Any comparison that is done between groups is fair[a]	1.38	0.75	15	27	50	1
9. All results are credible and seem valid	1.81	0.40	0	23	97	0.88
10. All conclusions seem justified by the results	1.83	0.43	2	17	101	1

[a] Not every study compares groups, therefore the total scores do not count up to 120.

indicator “Q8. Any comparison that is done between groups is fair” as it was not clear to one rater that this indicator was about the fairness of the comparison and not about the existence of a comparison.

Table 3 shows the mean scores for each quality indicator, while Fig. 4 shows the mean total quality score of the evaluation studies (as sum of the 10 quality indicators) in time. The quality of the evaluation studies was rather stable over time, there seems to be only a slight improvement in quality score in time, Spearman correlation coefficient $[r_S] = 0.192$; 95% confidence interval 0.013–0.359.

Of the 10 quality indicators, the indicators “Q2. The evaluated IT intervention is described in sufficient detail”, “Q3. Type, number, and sampling of involved study population are clearly described”, “Q5. Methods for collecting data are sufficiently clear”, “Q6. Methods for analysing data are sufficiently clear” and “Q8. Fair comparison between groups” had the lowest scores. The mean total quality score for RCTs (17.8) was significantly higher than for non-RCTs (14.8) (t-value = −5.09, d.f. = 118, $p < 0.001$). RCTs scored higher then non-RCTs on all 10 quality indicators. For all indicators but “Q1. Motivation, problems and study questions are clearly described” and “Q2. The evaluated information technology (the intervention) is described in sufficient detail” this difference was statistically significant ($p < 0.05$). No significant differences could be detected between single-unit, single-centre and multi-centre studies or between prospective and retrospective studies.

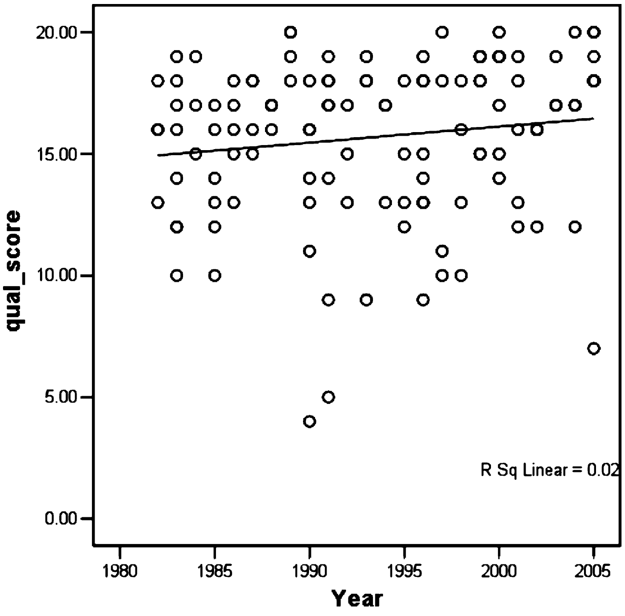

Fig. 4 – Total quality score of IT evaluation studies (possible range 0–20) 1982–2005 ($n = 120$). Spearman correlation coefficient $[r_S] = 0.192$.

4. Discussion

4.1. *Discussion of results*

We observed an increasing trend in using multi-center evaluation studies, a rising number of authors from various departments and a recently developed trend towards the usage of different data acquisition methods in an evaluation study. This all can be interpreted as an indication for maturation and better quality of evaluation research in medical informatics, as both multi-center studies, authors from multiple disciplines and triangulation of data sources point to more extensive and larger studies that may be better able to generate reproducible and valid findings. The slight increase of the overall quality of evaluation studies on IT in health care as measured with our newly developed quality score confirms this finding.

The developed quality score includes 10 indicators aiming to measure the quality of reporting the evaluation study since evidence based health informatics requires high-quality publications. Five out of the 10 quality indicators had scores lower than 1.5 while the other five had scores above 1.7. The indicator “Q8. Fair comparison between groups” is of special relevance for all controlled trials. Here we observed that these groups are not compared on baseline characteristics which makes it hard to check whether both groups were comparable. The indicator "Q2. The evaluated IT intervention is described in sufficient detail" is a very important one for the reader to translate the findings of the study to their own IT situation or to decide on future developments in their information architecture. Here we found several publications with an insufficient description, e.g. of the functionality, technology, usage patterns and workflow integration of the evaluation IT system. The same counts for the indicator "Q3. Type, number, and sampling of involved study population are clearly described" and hence hampers the reader to translate the study setting to their own situation. The indicators "Q5. Methods for collecting data are sufficiently clear" and "Q6. Methods for analyzing data are sufficiently clear" are important to assess the reliability of the reported results. Here we also found a large number of studies that, e.g. reported on results of user surveys, documentation analysis or time measurement, but without giving sufficient information to reproduce and assess their findings.

In contrast to others [18,25] that observed a decrease in quality we found a slight increase in quality of evaluation studies of IT interventions in health care between 1982 and 2005. However, we want to emphasize that the correlation coefficient is very small and it is generally known that significant correlation coefficient does not necessarily indicate a strong and important relationship [35]. Overall, the quality of published IT evaluation studies seems insufficient especially with regard to a clear description of the IT system and to the methods used to capture and analyse data. This is supported by remarks from authors of systematic reviews [11–13].

An important difference between our study and former ones is that we did not restrict our review to RCTs. In fact, we found that only 29% of all studies used an RCT design. This means that 71% of all summative evaluation studies on IT interventions in health care are either non-randomly controlled or uncontrolled clinical trials. To our knowledge, most of comparable work on the quality of medical and medical informatics research papers has focussed on the quality of randomized controlled trials, borrowing quality indicators for clinical trials from the medical sciences. For example, Godman [36] described a checklist for clinical trials, containing 18 quality attributes such as: specify alpha and beta level; make sample calculation; use blinding; describe treatment and control; or provide test statistic. In a comparable way, Johnston [28] analysed 28 controlled trials on clinical decision support systems. While these studies both used established quality indicators for quantitative clinical trials, they are not appropriate for other types of study designs in health informatics. Therefore we generalized the quality attributes to make them applicable to other evaluation papers.

4.2. *Discussion of methods*

Our 10-item list of quality indicators for publications evaluating IT in health care primarily aims at assessing the quality of reporting studies instead of the quality of the study itself. In this sense it is comparable to the goal of CONSORT for RCTs. However, since our list focuses on all kind of study designs, not limited to RCTs, it also includes the indicator "Q7. Methods seem adequate to answer study question" which assess the quality of the study itself. Therefore, one might argue that there is some mixture of assessing the quality of study and the quality of reporting. Our set of quality indicators has three important weaknesses that need to be discussed. First, the 10 criteria were unweighted for scoring purposes. This means that some criteria which might be more important than others are not weighted to reflect their importance. However, any weighting would be arbitrary and subjective and therefore equally subject to criticism. Hence for simplicity we have left criteria unweighted. Second, several criteria seem to be particular subjective ("Q9. Results are credible and seem valid"), which is due to the fact that we had to develop a comprehensive checklist. Finally, the overall report quality score and its parts were not tested on its properties (e.g. test–retest reliability, internal consistency and construct validity). Although in our pre-test the inter-rater variability of our quality score was good, indicating that our quality score was useful for objectively assessing reports of studies on evaluating IT interventions in health care we did not evaluate more formally the reliability, consistency and validity characteristics of our quality score. As far as we know, these three weaknesses of using our list of quality indicators also apply to well-accepted quality measures such as CONSORT [16]. Further research should prove the general usefulness, objectiveness and reliability of our quality score.

We analysed a random sample of just 15 papers from each 3-year period. We cannot be sure that this is representative of the overall number of evaluation studies published, especially for the last years where a lot more studies have been performed (see percentages on X-axis of Fig. 1). However, we decided to have a fixed number in each period to make the assessment feasible, and to give each time period an equal weight to be able to detect developments over time.

5. Conclusion

We conclude that the quality of evaluation studies on IT systems in health care still has to be improved in some aspects. Despite the increasing discussion about evidence-based health informatics [5], we could not see a strong increase in the quality of IT publications. This poses a severe problem as evaluation publications form the basis of any further analysis about the effects and quality of IT in health care. The continuous and dominant number of non-RCT studies reflects that various approaches are applicable to evaluate IT systems in health care. Therefore a general list of quality indicators to assess reports on evaluation studies on IT in health care independent on study design is essential. Journal editors should update instructions for authors to better cover the reporting of evaluation of IT systems to ensure that issues which affect the understanding of a paper and how the study was undertaken are adequately described. Referees should then be asked to judge papers in this context. The list of 10 quality indicators described in this paper (Table 1) could serve as a basis for improving the quality of reporting on evaluation studies of IT interventions in health care and for developing publication guidelines for authors.

Summary points

What was known before the study:

- Quality indicators for describing evaluation studies in medicine are mostly restricted to randomized controlled trials.
- Earlier studies showed that the quality of evaluation studies in medicine or on IT interventions in health care decreased over time.
- Evidence based health informatics requires high-quality publications on health informatics evaluation studies.

What this study has added to our knowledge:

- There is an increasing trend in using multi-centre designs in evaluation studies on IT interventions in health care.
- The quality of publications on evaluation studies has only slightly been increased in the last 24 years.
- Reports on randomized controlled trials show a significant higher quality score than non-RCTs.
- The quality of publications on IT evaluation studies in health care still has to be improved, especially the description of the IT intervention under evaluation and the description of the methods for data collection and data analysis.

Acknowledgement

Preliminary versions of this work have been presented on MIE2005 and AMIA2005.

REFERENCES

[1] Institute of Medicine, Crossing the Quality Chasm: A New Health System for the 21st Century, National Academy Press, Washington, 2001.

[2] R. Koppel, J. Metlay, A. Cohen, B. Abaluck, A. Localio, K. SE, et al., Role of computerized physician order entry systems in facilitating medication errors, JAMA 293 (10) (2005) 1197–2003.

[3] Y.Y. Han, J.A. Carcillo, S.T. Venkataraman, Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system, Pediatrics 116 (6) (2006) 1506–1512.

[4] E. Ammenwerth, J. Brender, P. Nykänen, H.U. Prokosch, M. Rigby, J. Talmon, 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. 73 (6) (2004) 479–491.

[5] M. Rigby, Evaluation: 16 powerful reasons why not to do it—and 6 over-riding imperatives. In: V. Patel, R. Rogers, R. Haux (Eds.), Proceedings of the 10th World Congress on Medical Informatics (Medinfo 2001), 84th ed., IOS Press, Amsterdam, 2001, pp. 1198–1202.

[6] C. Friedman, J.C. Wyatt, Evaluation Methods in Medical Informatics, Springer, New York, 1997.

[7] H. Heathfield, V. Peel, P. Hudson, S. Kay, L. Mackay, T. Marley, et al. Evaluating large scale health information systems: from practice towards theory. In: D. Masys (Ed.), AMIA Annual Fall Symposium, Hanley & Belfus, Philadelphia, 1997, pp. 116–120.

[8] J.R. Moehr, Evaluation: salvation or nemesis of medical informatics? Comput. Biol. Med. 32 (3) (2002) 113–125.

[9] E. Ammenwerth, N. de Keizer, An inventory of evaluation studies of information technology in health care: trends in evaluation research 1982–2002, Meth. Inf. Med. 44 (2005) 44–56.

[10] N. Aoki, K. Dunn, K. Johnson-Throop, J. Turley, Outcomes and methods in telemedicine evaluation, Telemed. J. E: Health 9 (4) (2003) 393–401.

[11] P. Whitten, F. Mair, A. Haycox, C. May, T. Williams, S. Hellmich, Systematic review of cost effectiveness studies of telemedicine interventions, BMJ 324 (2002) 1434–1437.

[12] A. Garg, N. Adhikari, H. McDonald, M. Rosas-Arellano, P. Devereaux, J. Beyene, et al., Effects of computerised clinical decision support systems on practitioner performance and patient outcomes. A systematic review, JAMA 293 (2005) 1223–1238.

[13] S. Ross, C. Lin, The effects of promoting patient access to medical records: a review, J. Am. Med. Inf. Assoc. 10 (2003) 129–138.

[14] R. Currell, P. Wainwright, C. Urquhart, Nursing record systems: effects on nursing practice and health care outcomes (Cochrane review), in: The Cochrane Library, Issue 1, Update Software, Oxford, 2000.

[15] F. Mair, P. Whitten, Systematic review of studies of patient satisfaction with telemedicine, Br. Med. J. 320 (7248) (2000) 1517–1520.

[16] F.J. Bath, V.E. Owen, P.M. Bath, Quality of full and final publications reporting acute stroke trials. A systematic review, Stroke (29) (1998) 2203–2210.

[17] C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, et al., Improving the quality of reporting of randomized controlled trials—the CONSORT statement, JAMA 276 (8) (1996) 637–639.

[18] R.H. Fletcher, S.W. Fletcher, Clinical research in general medical journals: a 30-year perspective, N. Engl. J. Med. 301 (1979) 180–183.

[19] D. Moher, The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials, Ann. Intern. Med. 134 (8) (2001) 657–662.

[20] D. Altman, Better reporting of randomised controlled trials: the CONSORT statement, BMJ (313) (1996) 570–571.

[21] H. Heathfield, I. Buchan, Current evaluations of information technology in health care are often inadequate, BMJ 313 (7063) (1996) 1008.

[22] B. Kaplan, Evaluating informatics applications–some alternative approaches: theory, social interactionism, and call for methodological pluralism, Int. J. Med. Inf. 64 (2001) 39–56.

[23] A.D. Harris, J.C. McGregor, E.N. Perencevich, J.P. Furuno, J. Zhu, D.E. Peterson, et al., The use and interpretation of quasi-experimental studies in medical informatics, J. Am. Med. Inf. Assoc. 13 (1) (2006) 16–23.

[24] Cochrane Collaboration, Glossary of Terms in The Cochrane Collaboration, Version 4.2.5, update May 2005. Available from: URL: www.cochrane.org.

[25] R. van der Loo, Overview of published assessment and evaluation studies, in: E.M.S.J. van Gennip, J.S. Talmon (Eds.), Assessment and Evaluation of Information Technologies in Medicine, IOS Press, Amsterdam, 1995, pp. 261–282.

[26] L. Mihan, J. Windeler, Methodological quality of controlled studies in the "Medizinische Klinik" journal. Analysis of

contributions appearing between 1979 and 1996 (in German), Med. Klin. 94 (1) (1999) 1–8.

[27] S.H. Downs, N. Black, The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions, J. Epidemiol. Community Health 52 (Jun 6) (1998) 377–384.

[28] M. Johnston, K. Langton, R. Haynes, A. Mathieu, Effects of computer-based clinical decision support systems on clinician performance and patient outcome–a critical appraisal of research, Annu. Intern. Med. 120 (1994) 135–142.

[29] R. Roine, A. Ohinmaa, D. Hailey, Assessing telemedicine: a systematic review of the literature, CMAJ 165 (6) (2001) 765–771.

[30] A. Verhagen, H. de Vet, R. de Bie, A. Kessels, M. Boers, L. Bouter, et al., The Delphi list; a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus, J. Clin. Epidemiol. (51) (1998) 1235–1241.

[31] G. Hall, R. Smith, G. Drummond, J. Norman, A. Spence, J. Lilleyman, et al., How to Write A Paper, BMJ Publishing Group, 1994.

[32] T. Greenhalgh, R. Taylor, How to read a paper: papers that go beyond numbers (qualitative research), BMJ 315 (1997) 740–743.

[33] Borreani, Miccinesi, Lina Brunelli, An increasing number of qualitative research papers in oncology and palliative care: does it mean a thorough development of the methodology of research? Health Qual. Life Outcomes 2 (1) (2004) 1–23.

[34] E. Ammenwerth, A. Wolff, P. Knaup, H. Ulmer, S. Skonetzki, J. van Bemmel, et al., Developing and evaluating criteria to help reviewers of biomedical informatics manuscripts, J. Am. Med. Inf. Assoc. 10 (5) (2003) 512–514.

[35] D. Altman, Relation between two continuous variables, Pract. Statist. Med. Res. (1991) 277–324.

[36] C. Godman, Literature Searching and Evidence Interpretation for Assessing Health Care Practices, Norstedts Tryckeri AB, Stockholm, 1993.