Standards for reporting randomized controlled trials in medical informatics: a systematic review of CONSORT adherence in RCTs on clinical decision support

K M Augestad,1,2 G Berntsen,1 K Lassen,2,4 J G Bellika,1,3 R Wootton,1 R O Lindsetmo,2,4 Study Group of Research Quality in Medical Informatics and Decision Support (SQUID)

ABSTRACT
Introduction The Consolidated Standards for Reporting Trials (CONSORT) were published to standardize reporting and improve the quality of clinical trials. The objective of this study is to assess CONSORT adherence in randomized clinical trials (RCT) of disease specific clinical decision support (CDS).
Methods A systematic search was conducted of the Medline, EMBASE, and Cochrane databases. RCTs on CDS were assessed against CONSORT guidelines and the Jadad score.
Result 32 of 3784 papers identified in the primary search were included in the final review. 181 702 patients and 7315 physicians participated in the selected trials. Most trials were performed in primary care (22), including 897 general practitioner offices. RCTs assessing CDS for asthma (4), diabetes (4), and hyperlipidemia (3) were the most common. Thirteen CDS systems (40%) were implemented in electronic medical records, and 14 (43%) provided automatic alerts. CONSORT and Jadad scores were generally low; the mean CONSORT score was 30.75 (95% CI 27.0 to 34.5), median score 32, range 21–38. Fourteen trials (43%) did not clearly define the study objective, and 11 studies (34%) did not include a sample size calculation. Outcome measures were adequately identified and defined in 23 (71%) trials; adverse events or side effects were not reported in 20 trials (62%). Thirteen trials (40%) were of superior quality according to the Jadad score (≥3 points). Six trials (18%) reported on long-term implementation of CDS.
Conclusion The overall quality of reporting RCTs was low. There is a need to develop standards for reporting RCTs in medical informatics.

INTRODUCTION
Randomized controlled trials (RCTs) are considered the gold standard for investigating the results of clinical research because they inherently correct for unknown confounders and minimize investigator bias.1–3 The results of these trials can have profound and immediate effects on patient care. When RCTs are reported, it is recommended that the Consolidated Standards of Reporting Trials (CONSORT)4 are followed. CONSORT was first published in 1996 and has been revised several times since.5 The CONSORT statement is widely supported and has been translated into several languages to facilitate awareness and dissemination. An extension of the CONSORT statement was published in 2008, focusing on randomized trials in non-pharmacologic treatment.6 CONSORT consists of a checklist of information to include when reporting on an RCT; however, inadequate reporting remains common among clinicians.7–12 Higher quality reports are likely to improve RCT interpretation, minimize biased conclusions, and facilitate decision making in light of treatment effectiveness.7 Furthermore, there is evidence that studies of lower methodological quality tend to report larger treatment effects than high quality studies.13–15

Research on clinical decision support (CDS) tools has rapidly evolved in the last decade. CDS provides clinicians with patient specific assessment or guidelines to aid clinical decision making16 and improve quality of care and patient outcome.17 18 CDS has been shown to improve prescribing practices,19 reduce serious medication errors,20 21 enhance delivery of preventive care services,22 and improve guidelines adherence,23 and likely results in lasting improvements in clinical practice.24 However, clinical research on CDS tools faces various methodological problems25–28 and is challenging to implement in the field of health informatics.29 Guidelines for reporting studies in health informatics have been published,26 but there is no universal consensus.

Numerous RCTs examining (disease specific) CDS tools aimed at improving patient treatment have been performed. It is unclear whether these studies provided CONSORT statements when the trials were reported. Although several studies have evaluated the quality of RCTs in medical journals,3 7 8 to date none have been directed at medical informatics literature published in dedicated journals. The objective of this paper is to perform a systematic review of RCTs to assess the quality of clinical CDS research focusing on disease specific interventions. We aimed to score the identified RCTs according to the CONSORT checklist and Jadad score.3 Finally, we discuss the implications of these results in the context of evidence-based medicine.

MATERIALS AND METHODS
The review followed the PRISMA statements (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)30 and the CONSORT statement.31 The CONSORT statement has been translated into several languages and is widely supported.32–34

This paper is freely available online under the BMJ Journals unlocked scheme, see http://jamia.bmj.com/site/about/unlocked.shtml
and Meta-Analyses) and was divided into two work phases: (a) identification of RCT trials assessing disease specific CDS and (b) data extraction and assessment of RCT quality.

The Study Group of Research Quality in Medical Informatics and Decision Support (SQUID) is a multidisciplinary study group. Members have expertise in hospital medicine (KL, ROL, KMA), RCTs in medicine (surgery) (KL), RCTs in telemedicine (RW), trials of medical informatics (JGB, KMA), and epidemiological research (GB). The group’s objective is to assess and improve the quality of clinical informatics research with special focus on randomized controlled trials aimed at enhancing physician performance.

We defined CDS as ‘any electronic or non-electronic system designed to aid directly in clinical decision making, in which characteristics of individual patients are used to generate patient specific assessments or recommendations that are subsequently presented to clinicians for consideration.’ We defined disease specific CDS as ‘a clinical decision support aimed at a specific disease, describing symptoms, diagnosis, treatment, and follow-up.’

Search strategy
This systematic review is based on a PubMed, EMBASE, and Cochrane Controlled Trials Register search using EndNote X5 (EndNote, San Francisco, California, USA) for relevant publications published through November 2010. We piloted search strategies and modified them to ensure they identified known eligible articles. We combined keywords and/or subject headings to identify CDS (clinical decision support system, computer-assisted decision making, computer-assisted diagnosis, hospital information systems) in the area of RCTs (ie, randomized controlled trials). We searched publications accessible from the web pages of the International Journal of Medical Informatics, Journal of the American Medical Informatics Association, and BMC Medical Informatics and Decision Making. We systematically searched the reference lists of included studies. Reviews addressing CDS were investigated and papers fulfilling the inclusion criteria were included. The searches were individually tailored for each database or journal. Experienced clinicians reviewed all search hits and decided whether a CDS was aimed at a specific disease and fulfilled inclusion criteria. The titles, index terms, and abstracts of the identified references were studied and each paper was rated as ‘potentially relevant’ or ‘not relevant.’ Disagreements regarding inclusion were resolved by discussion. Only trials performed the last 10 years were included.

Inclusion criteria were:
- Randomized controlled trial
- CDS describing specific diseases and treatment guidelines
- CDS aimed at physicians.

Exclusion criteria were:
- Papers published before the year 2000
- Not published in English
- Proceedings, symposium, and protocol papers.

The search strategy yielded 3784 papers. We retrieved and reviewed the full text of 364 papers; 32 papers were included in the final review (figure 1).

Assessing RCT quality
Scoring according to CONSORT
A checklist of 22 items from the revised 2001 CONSORT guidelines was analyzed. The score for each item ranged from 0 to 2 (0=no description, 1=adequate description). The maximum score a paper could obtain was 44 points.

Scoring according to Jadad
The Jadad scale is a 5-point scale for evaluating the quality of randomized trials in which three points or more indicates superior quality. The Jadad scale is commonly used to evaluate RCT quality. The scale contains two questions each for randomization and masking, and one question evaluating reporting of withdrawals and dropouts.

Scoring according to the sequential phases of a complex intervention
An RCT evaluating a CDS tool is defined as a complex intervention, that is an intervention consisting of various interconnecting parts. Cambell et al suggested four sequential phases for developing RCTs for complex interventions: theory, modeling, exploratory trial, definitive randomized controlled trial, and long-term implementation. Included trials were scored according to these sequential phases, that is one point was given for each phase.

Scoring according to CDS features critical for success
Kawamoto et al identified certain CDS factors associated with clinical improvement. These factors are: automatic provision of CDS, CDS at the time and location of decision making, provision of a recommendation rather than just an assessment, computer based assessment, and automatic provision of decision as part of clinician workflow. The identified CDS tools were scored according to these factors, giving one point for each feature. All appraised papers were discussed by the two reviewers and, if necessary, by a third independent reviewer to verify the
appraisal process and resolve disagreement; when consensus could not be reached, the third reviewer assessed the items and provided the tiebreaker score.

**Statistics**

Trial characteristics and CONSORT adherence were analyzed and interpreted with the trial as unit of analysis. Descriptive statistics were analyzed using percentages, standard deviation, confidence intervals, $2 \times 2$ contingency tables, $\chi^2$ test, and Fisher’s exact test when appropriate. We used proportions for categorical variables and mean for continuous variables. For reasons of comparison, trials were divided into groups according to whether or not their outcome was positive. A positive outcome was defined as either a primary or secondary outcome with $p < 0.05$. All tests were two-sided and a probability (p) value of $<0.05$ was considered statistically significant. Microsoft Excel and SPSS PASW Statistics v 18.0 were used for the statistical analyses.

**RESULTS**

**Clinical features**

Of 3784 potentially relevant articles screened, 52 papers met all our inclusion criteria (table 1).

Fourteen (43%) of the trials were performed in the US, seven (21%) in the Netherlands, and four (12%) in the UK. Four of the trials were published in medical informatics journals, and the rest in medical journals. The trials included 181 702 patients and 7315 physicians. The majority (22 trials) were performed in primary care, including 897 general practitioner (GP) offices. Of the 11 trials performed at hospital level, two were performed in an outpatient department, three in internal medicine departments, one in a surgical department, one in an intensive care unit, two in emergency departments, one in a trauma unit, and one in various different departments. Asthma ($n=4$), diabetes ($n=4$), and hyperlipidemia ($n=3$) were the most common diseases addressed (table 1).

**General trial features**

Twenty-six trials (81%) did not provide an RCT registration number (ie, http://Clinicaltrials.gov and others), while only seven trials (21%) offered web access to the full trial protocol. One trial did not state funding sources (table 2). In nine trials (28%), more than half of the authors were medical doctors; in 10 trials, information on the background and education of the author(s) was not provided. Twenty-two (68%) trials chose a cluster-randomized design, which was the most common design among trials in primary care (21 of 22). Of the nine trials performed in a hospital setting, four had a cluster-randomized design and in these cases the department was chosen as the clustering unit. Two trials provided information on changes to the trial protocol, and one trial addressed CONSORT guidelines.

**CDS features**

Less than half of the CDS tools were implemented in an electronic medical record, and 14 (43%) of the CDS tools provided automatic alerts (table 2). Twenty-four (75%) of the developed CDS tools provided decision support at the time and location of the decision need. Eighteen (56%) of the CDS tools did not disrupt the natural workflow of the physician. None of these CDS features had a significant influence upon the primary endpoint or overall conclusions.

**Addressing sequential phases of a complex intervention**

None of the trials defined the intervention as complex or discussed the definition of a complex intervention.77 78 90 Four trials defined all phases of a complex intervention and these phases were described in detail (table 2).

**Trials reporting on long-term CDS implementation**

Six trials reported on the long-term implementation of the CDS tool used in the RCT (table 1).

Four of these trials addressed all phases of a complex intervention and had a statistically higher CONSORT score compared to trials not reporting long-term implementation (OR 1.64, $p=0.04$). Three of these trials were performed at a hospital level, with the largest trial including 87 000 patients.

**Inter-rater reliability and CONSORT score**

The intraclass correlation coefficient used to establish inter-rater reliability was 0.69 for the 22-item CONSORT scale. The mean CONSORT score was 30.75 (95% CI 27.0 to 34.5), median score 32, range 21–38.

**CONSORT: title, abstract, and background**

Five trials did not identify a randomized design in their title. All trials had a structured abstract and gave a solid background and rationale for the trial (table 3).

**CONSORT: materials and methods**

One trial addressed the CONSORT guidelines in their Material and Methods section. Twenty-seven trials (84%) clearly defined their participants, eligibility, and ethics approval. Fourteen trials (45%) did not clearly define the study objective or hypothesis. Twenty-three trials (72%) had an adequate definition of outcome measures. Fourteen studies (37%) did not perform or had an inadequate sample size calculation (table 3).

**CONSORT: randomization**

Most trials described mechanisms to generate random allocation (59%) and the method of implementing the random sequence (47%). In contrast, only five trials (15%) gave adequate information regarding blinding (whether or not blinding was necessary and if necessary, how it was performed) (table 3).

**CONSORT: results**

Most trials (87%) provided a detailed description of statistical methods (table 3). Five trials had no figure showing participant flow and four trials did not include a table showing demographics. Nine trials did not address exclusions during the trial, and 10 trials did not define the date of trial initiation and termination. Only two trials performed an interim analysis, and only one trial addressed the ‘harms or unintended effects’ of the intervention.

**CONSORT: discussion**

The interpretation of results was justified in 28 trials (87%). Four trials did not discuss limitations and six trials did not address generalizability or provide recommendations for the future (table 3).

**Jadad score**

Thirteen trials (40%) were classified as superior quality trials ($\geq 3$ points). Nineteen (59%) described the study as randomized, and the sequence of randomization was explained and was appropriate. Twenty-seven (85%) did not describe blinding. Ten (32%) did not describe dropout (table 4).

**DISCUSSION**

**Summary of findings**

This is the first review assessing the quality of RCTs of disease specific CDS as a primary intervention. We have analyzed their
| Author            | Author MD/total | Setting | Disease                          | Country | Sample       | Primary outcome (PO) | Effect of PO (CI, p value) | Overall conclusion                  | Journal     |
|-------------------|-----------------|---------|----------------------------------|---------|--------------|----------------------|----------------------------|------------------------------------|-------------|
| Apkon et al (2005) | 5/11            | Hospit  | Multiple diseases                | USA     | 1902 pts, 4639 consul | % of 24 healthcare quality process measures | NE                        | NE                                 | Clinical    |
| Bell et al (2010)  | 6/9             | GP      | Asthma                           | USA     | 12 offices, 19450 pts | Updated asthma care plan | 14% (0.03)               | CDS improved adherence to guidelines | Clinical    |
| Bertoni et al (2009) | 4/9             | GP      | Hyperlipidemia                   | USA     | 61 offices, 5057 pts | Proportion of patients screened with appropriate decision making | 10% (2.8% to 16.8%, 0.01) | Improved adherence to guidelines | Clinical    |
| Bosworth et al (2008) | 3/10            | GP      | Blood pressure                   | USA     | 32 GPs, 588 pts | % <140/90 mm Hg, 2-year follow up | NE                        | NE                                 | Clinical    |
| Cleveringa et al (2008) | 3/4             | GP      | Diabetes                         | Netherl.| 55 offices, 3391 pts | A1C levels           | NE                        | NE on A1C levels, but reduced cardiovascular risk | Clinical    |
| Dexter et al (2001) | 4/6             | Hospit  | Pneumococcal influenza vaccination, thrombo-embolism | USA     | 202 phys, 10865 pts | Rates of preventive therapies ordered | 35% (0.001)               | Increased preventive treatment | Clinical    |
| Emery et al (2007)  | NA/10           | GP      | Genetic risk assessment          | Australia | 45 offices | % of referrals consistent with guidelines | OR 5.2 (1.7 to 15.8, 0.006) | Increased number and quality of referrals | Clinical    |
| Feldstein et al (2006) | 3/8             | GP      | Osteoporosis                      | USA     | 159 GPs, 311 pts | % BMD measurement | OR 1.27 (0.79 to 1.75, 0.001) | Improved process of care and some clinical markers | Clinical    |
| Flottorp et al (2002) | NA/5            | GP      | UVI sore throat                   | Norway   | >20000 consul | Rate of AB use | 45% (0.001)               | Increased BMD measurement | Clinical    |
| Frieling et al (2002) | NA/8            | GP      | Diabetes                         | Netherl.| 185 GPs, 2800 consul | Rate of foot and eye examination | OR 1.68 (1.19 to 2.39) | Little overall effect on changing practice | Clinical    |
| Gilutz et al (2009) | 13/15           | GP      | Hyperlipidemia                    | Israel   | 112 offices, 448 pts | Appropriate lipoprotein monitoring | 6% (0.001)               | Facilitated adherence to guidelines | Clinical    |
| Hicks et al (2008)  | 5/7             | GP      | Blood pressure                    | USA     | 14 GP offices, 2027 pts | Blood pressure | 0%                        | No reduction in BP but better prescribing routines | Clinical    |
| Holtbrook et al (2009) | 4/9             | GP      | Diabetes                         | Canada   | 46 GPs, 511 pts | Process composite score (PCS) | OR 1.80 (1.04 to 3.05) | Improved process of care and some clinical markers | Clinical    |
| Kuijber et al (2006) | NA/8            | GP      | Asthma                           | Netherl.| 32 offices, 40 GPs, 10863 pts | Contact frequency | 10% (0.034)               | Improved adherence to guideline | Clinical    |
| McGovan et al (2008) | NA/4            | GP      | Multiple diseases                 | Canada   | 88 GPs | Impact assessment scores | 48% (no p value) | CDS positive impact on decision making | Clinical    |
| McKinley et al (2000) | 4/10            | Hospit  | Acute respiratory distress syndrome | USA     | 67 pts | Survival and ICU stay | No negative impact on survival | Effectively standardized ventilator management | Clinical    |
| Maclean et al (2009) | 1/4             | GP      | Diabetes                         | USA     | 132 GPs, 64 GP offices, 7412 pts | Glycemic control mean A1C and A1C <7% | NE                        | CDS feasible for patients and providers but no physiologic effect | Clinical    |
| Mulvaney et al (2008) | 1/7             | Hospit  | Several diseases                  | USA     | 226 phys | Action index, CDS induced different or new treatment | OR 8.19 (1.04 to 64.0) | Treatment issues were significantly impacted by CDS | Informa     |
| Poels et al (2007)  | NA/8            | GP      | Asthma                           | Netherl.| 78 GPs | GP diagnosis was compared with an expert panel judgment | OR 1.08 (0.70 to 1.88) | No effect on patients and providers' achievement | Clinical    |
| Rollman et al (2003) | 3/6             | GP      | Depression                        | USA     | 200 pts | Hamilton rating scale for depression | NE                        | Little effect on patients and providers' achievement | Clinical    |
| Rosenboom et al (2005) | 5/9             | Hospit  | Several diseases                  | USA     | 202 phys | Rates of opportunity, access, and utilization of CDS | NE                        | Little effect on patients and providers' achievement | Clinical    |
| Roukema et al (2008) | 3/4             | Hospit  | High fever in children            | Netherl.| 164 pts | Total ED time | NE                        | Increased CDS utilization | Informa     |
| Roy et al (2009)    | 11/13           | Hospit  | Pulmonary embolism                | France   | 20 hospit, 1015 pts | % sequence of tests that yielded a post-test probability <5% | 19% (2.9% to 35%, 0.023) | Improved diagnostic decision making | Clinical    |
| Thomas et al (2009) | NA/7            | GP      | Urology                          | UK       | 66 offices, 959 pts | Guideline compliance | Increase 0.5 points (0.2 to 0.8, 0.001) | Improved process of out-patient referral | Clinical    |
outcome, CONSORT adherence and Jadad score. Methodologically, research quality varies and adherence to CONSORT guidelines is low for certain checklist items. Thirteen trials (40%) were classified as superior quality trials according to their Jadad score (≥3 points). According to our analysis, there is considerable room for improving methodology in areas such as the description of specific research objectives, randomization methods, sample size calculations, reporting of adverse events, and a general focus on CONSORT. Similarly, the Jadad score was low on several checklist items. Surprisingly few studies defined their CDS intervention as a complex intervention; only four studies described all phases of a complex intervention including long-term implementation.

Research challenges of complex interventions

A complex intervention was defined by Campbell et al77 as an intervention that is ‘built up from a number of components, which may act both independently and interdependently.’ Similarly, Campbell defined an intervention with a decision support system as a complex intervention.77 In 2000, the Medical Research Council in the UK proposed a framework for the development and evaluation of RCTs for complex interventions (theory, modeling, exploratory trial, definitive RCT, long-term implementation);77 which was further improved in 2007.81 The methodological challenges of complex interventions have been thoroughly discussed in the field of medical informatics,25 29 as well as in the area of health service research.79 82 There have been arguments against over-standardization of complex interventions. Complex and large health organizations are characterized by flux, contextual variation, and adaptive learning rather than stability, and a standardized approach will not fit such organizations.80 However, our review shows that most trials do not address the term ‘complex intervention’ and as many as 23 trials (71%) did not perform an exploratory trial before the definitive RCT. This problem is well discussed by Friedman, who introduces the ‘tower of achievements’.

According to Friedman, integration across research phases is of utmost importance to success in the field.

Quality of RCTs in medical informatics versus clinical trials

Our survey shows generally low CONSORT adherence and only 13 trials were defined as superior quality trials according to their Jadad score. However, the research quality of RCTs has been of varying quality in medical research as well. In a review from 200686 assessing 69 RCTs of surgery, only 37% of trials were classified as of superior quality. CONSORT scores were generally low but significantly higher in trials with higher author numbers, multi-centre trials, and trials with a declared funding source.25 It has been concluded that there is a need to improve awareness of the CONSORT statement among authors, reviewers, and editors.8 The similar concerns were recently reported in several medical journals, which concluded that there was low adherence to key methodological items.83–85 These conclusions from the medical literature are in accordance with our review findings.

Strength and limitations of our study

This study has several important strengths. First, our literature search was thorough and we screened more than 3700 articles. Second, this is the first review to evaluate the general trial quality and CONSORT adherence of RCTs evaluating CDS tools as a clinical intervention. Research on CDS tools is methodologically challenging.28 Thus, a focus on research methods in medical informatics is important, and adherence to CONSORT
**Table 2** Characteristics of RCTs of clinical decision support and impact on outcome and implementation

| General trial features                  | Outcome +, n=24 (%) | Outcome -, n=8 (%) | Total, n=32 (%) |
|----------------------------------------|---------------------|--------------------|-----------------|
| Protocol access                         | 7 (29)              | 0                  | 7 (21)          |
| Identification of RCT number           | 6 (25)              | 0                  | 6 (18)          |
| Funding sources identified              | 23 (95)             | 8 (100)            | 31 (96)         |
| >50% MD authors                        | 8 (33)              | 1 (12)             | 9 (28)          |
| Primary care                            | 17 (70)             | 5 (62)             | 22 (68)         |
| Clustered design                       | 18 (75)             | 4 (50)             | 22 (68)         |
| Patients                                | 158 (240)           | 23 (462)           | 181 (702)       |
| Participating MDs                       | 2270                | 846                | 3116            |

**CDS feature**

| CDS time/location of decision          | 17 (70)             | 7 (87)             | 24 (75)         |
| Automatic alert                         | 11 (45)             | 3 (37)             | 14 (43)         |
| Implemented in EMR                     | 10 (41)             | 3 (37)             | 13 (40)         |
| No disruption to workflow               | 15 (62)             | 3 (37)             | 18 (56)         |
| All features present                    | 8 (33)              | 0                  | 8 (25)          |

**Phases of complex interventions**

| Theory                                  | 24 (100)            | 8 (100)            | 32 (100)        |
| Modeling                                | 8 (33)              | 5 (62)             | 13 (40)         |
| Exploratory trial                       | 7 (29)              | 2 (25)             | 9 (28)          |
| Definitive RCT                          | 24 (100)            | 8 (100)            | 32 (100)        |
| Long-term implementation                | 4 (16)              | 2 (25)             | 6 (18)          |
| All phases present                      | 3 (12)              | 1 (12)             | 4 (12)          |

**RCT assessment tool**

| CONSORT score mean (SD)                  | 29.9 (4.7)          | 33.1 (3.9)         | 30.7 (4.7)      |
| Jadad score mean (SD)                    | 2.0 (1.5)           | 2.75 (1.6)         | 2.2 (1.5)       |

Outcome + is defined as either a positive primary or positive secondary endpoint (p < 0.05).

Outcome + is defined as either a positive primary or positive secondary endpoint (p < 0.05).

There were no significant differences between Outcome + and Outcome −.

CDS, clinical decision support; EMR, electronic medical record; MDs, medical doctors, including hospital physicians and general practitioners; RCT, randomized controlled trial.

---

**Table 3** The CONSORT checklist: scoring of 32 RCT trials of disease specific clinical decision support systems

| Item* | Description | No description, n (%) | Inadequate, n (%) | Adequate, n (%) |
|-------|-------------|-----------------------|-------------------|-----------------|
| 1     | Allocation (eg, ‘random allocation,’ ‘randomly assigned,’ or ‘randomized’) | 0                     | 5 (15.6)       | 27 (84.4)       |
| 2     | Justification | 0                     | 3 (9.4)         | 29 (90.6)       |
| 3     | Eligibility criteria for participants and location of data collection | 2 (6.3)             | 3 (9.4)        | 27 (84.4)       |
| 4     | Details and timing of interventions | 8 (25.0)           | 2 (6.3)        | 22 (68.8)       |
| 5     | Specific objectives and hypotheses | 3 (9.4)            | 11 (34.4)      | 18 (56.3)       |
| 6     | Identification and definition of outcome measures | 4 (12.5)          | 5 (15.6)       | 23 (71.9)       |
| 7     | Prestudy sample size calculation | 11 (34.4)          | 3 (9.4)        | 18 (56.3)       |
| 8     | Method of generation of the random sequence | 5 (15.6)          | 8 (25.0)       | 19 (59.4)       |
| 9     | Method of implementation of the random sequence | 10 (31.3)         | 7 (21.9)       | 15 (46.9)       |
| 10    | Details of personnel involved in recruitment, allocation, and outcome measurement | 14 (43.8)         | 7 (21.9)       | 11 (34.4)       |
| 11    | Whether subjects, treatment providers, or assessors/analysts were blinded | 24 (75.0)          | 3 (9.4)        | 5 (15.6)        |
| 12    | Statistical methods | 0                     | 4 (12.5)       | 28 (87.5)       |
| 13    | Flow of participants through each stage | 5 (15.6)          | 2 (6.3)        | 25 (78.1)       |
| 14    | Dates defining the periods of recruitment and follow-up | 9 (28.1)          | 1 (3.1)        | 22 (68.8)       |
| 15    | Baseline demographic and clinical characteristics of each group | 4 (12.5)          | 0              | 28 (87.5)       |
| 16    | Number of participants in each group analysis; whether the analysis was by ‘intention to treat’ | 1 (3.1)           | 16 (50.0)      | 15 (46.9)       |
| 17    | Complete reporting of results with CIs | 2 (6.3)            | 12 (37.5)      | 18 (56.3)       |
| 18    | Multiple testing and corrections | 0                     | 0              | 0               |
| 19    | All important adverse events or side effects | 20 (62.5)         | 11 (34.4)      | 1 (3.1)         |
| 20    | Interpretation of the results, including trial limitations and weaknesses | 1 (3.1)           | 3 (9.4)        | 28 (87.5)       |
| 21    | Generalizability (external validity) of the trial findings | 1 (3.1)           | 5 (15.6)       | 26 (81.3)       |
| 22    | General interpretation of the results in the context of current evidence | 0                     | 1 (3.1)       | 31 (96.9)       |

The mean CONSORT score for the 32 included trials was 30.75 (95% CI 27.0 to 34.5), median score 32, range 21–38. The intraclass correlation coefficient used to establish inter-rater reliability was 0.69. All appraised papers were discussed by the two reviewers and, if necessary, by a third independent reviewer to verify the appraisal process and resolve disagreement; when consensus could not be reached, the third reviewer assessed the items and provided the tiebreaker score.

*Score for each item: 0=no description, 1=inadequate description, 2=adequate description; maximum score=44.

RCT, randomized controlled trial.

---

has never been assessed. Third, we are currently recruiting patients into an RCT addressing the use of disease specific CDS tools and thus have experienced the inherent methodological challenges. In addition to technological problems, these trials also face the challenges of a complex intervention. These research questions have been addressed in this review.

One limitation of our study might be that only RCTs assessing CDS systems aimed at physicians were included. However, when planning this review, the research group wanted to identify CDS trials to improve patient treatment as these trials should ideally adhere to research conventions in general medical society. In this context the research group felt it natural to exclude CDS not addressing physicians.

Another limitation might be the reporting of the various phases in a complex intervention. Our review shows that only six trials (18%) report on long-term implementation. However, all studies were RCTs and thus were in the stage prior to implementation. It may be that implementation did occur after the RCT was published but was not part of the publication. It might also be that some providers implemented their long-term intervention, but as the RCT did not support this, they were reluctant to report on it. Similarly, it is possible that theoretical and preliminary work might have been carried out but was not fully described in an RCT paper.

Finally, it is unclear whether or not ‘complex intervention’ is a term widely accepted in medical informatics circles. We identified the term ‘complex intervention’ in one JAMIA article from 2008, with the other mentions of this concept all being in BMJ. Since JAMIA readership is largely within the US, it is unclear whether it is mandatory for CDS and their evaluation to be declared as complex interventions and thus follow the required phases.
Challenges of RCTs in medical informatics

Recently, Liu$^{22}$ discussed the pros and cons of RCTs in medical informatics. We agree with their view that RCTs are not the only method for evaluation. Medical informatics interventions are usually performed in a complex organizational environment. In this context, there is a need for different research methods, and often a mixture of qualitative and quantitative methods, depending on the research subject. However, when an RCT is deemed the proper design, standards of reporting must be followed. In addition, RCTs in medical informatics face several methodological challenges, some of which have been clarified in this review.

Choice of outcome measures

In principal, outcomes can either be patient orientated, process orientated, or system orientated. The choice of outcome measures should be clearly related to the research question. Our review shows a large mixture of primary outcomes, which makes meta-analyses of effects impossible. Thus, a clear conclusion regarding the effects of CDS (in the form of a meta-analyses) cannot be reached.

Sample size calculations

The planning of an RCT should begin with sample size calculation. This assessment is closely related to the choice of primary outcome, as different primary outcomes can result in different sample size estimates. The sample estimate is crucial to determine the resources and time needed to conduct a properly designed RCT with enough power to reject or accept the null hypothesis. Kiehan et al.$^{11}$ address concerns about the poor standards of reporting sample size calculations. They conclude that many of these trials are flawed from the start due to inadequate power to assess any real difference between interventions. In this review, approximately 50% of the trials had an inadequate estimate of sample size, a surprisingly low number.

Randomization

Should randomization be performed at an individual or an organizational level? In this review, 68% preferred a clustered design, clustered at the level of hospitals, departments, or GP offices. There are obvious advantages to a cluster design in complex health organizations, as problems of blinding and random sequence implementation will be avoided. In addition, clustering randomization is usually less demanding of resources, as randomization can be performed before the actual trial period with fewer personnel involved.

Conclusion

The research methodology in the identified trials is of low quality, suggesting a need for increased focus on the methods of conducting and reporting RCT trials. Study designs that adhere to CONSORT are not always appropriate in medical informatics research.$^{26}$ However, RCTs evaluating CDS tools in a clinical setting should adjust to the accepted consensus. Thus, CONSORT guidelines for conducting RCT trials should be addressed and subsequently implemented in the trial. CONSORT guidelines for non-pharmacological treatment$^{3}$ provide a solid basis for reporting RCTs evaluating CDS systems, but an adjustment for medical informatics is needed. The societies for medical informatics should aim for a consensus statement to improve the quality of reporting RCTs, trials of informatics applications, and CDS.

Funding

The Norwegian Health Authorities Research Fund supported this survey (grant number 40614).

Competing interests

None

Ethics approval

No patients were involved in the survey and thus approval by the hospital IRB board was not required.

Contributors

Study concept and design: KMA and GB. Acquisition of data: GB and KMA. Analysis and interpretation of data: KMA, GB, RW, and KL. Drafting of the manuscript: KMA. Critical revision of the manuscript for important intellectual content: KMA, GB, KL, JGB, RW, and ROL. Statistical analysis: KMA and GB. Administrative, technical, and material support: KMA and GB. Study supervision: KMA, RW, GJB, and ROL. RW had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Provenance and Peer review

Not commissioned; externally peer reviewed.

REFERENCES

1. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001;285:1992–5.
2. Kassell NF, Dumont AS. Randomized controlled trials in surgery: comic opera no more? J Thorac Cardiovasc Surg 2006;132:243–4.
3. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1–12.
4. Consort 2010. Lancet 2010;375:1186.
5. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285:1987–91.
6. Boutron I, Moher D, Altman DG, et al. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008;148:295–309.
7. Kiehna EN, Starke RM, Pouratian N, et al. Standards for reporting randomized controlled trials in neurosurgery. J Neurosurg 2010;114:280–5.
8. Baldasaranian SP, Wiener M, Alshameen Z, et al. Standards of reporting of randomized controlled trials in general surgery: can we do better? Ann Surg 2006;244:663–7.
9. Horton R. Surgical research or comic opera: questions, but few answers. Lancet 1996;347:984–5.
10. Neugebauer EA, Morino M, Habermehl B. Surgical research or comic opera? Let’s give answers! Surg Endosc 2010;24:1411–12.
11. Sade RM. “Surgical research or comic opera” redux. Ann Thorac Surg 2008;82:1173–4.
12. Sellier R, Whitlee I. The lessons learnt from ISAT: surgical research is rescued from comic opera. Br J Neurosurg 2004;18:405–6.
13. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1996;352:609–13.

Table 4 The Jadad instrument: scoring of 32 RCT trials of disease specific clinical decision support systems

| Item* | Description | % max score (n) | % 0 points (n) |
|-------|-------------|----------------|---------------|
| 1     | Was the study described as randomized? | 59 (19) | 41 (13) |
|       | Additional point if the method for generating the sequence of randomization was described and it was appropriate | 59 (19) | 41 (13) |
|       | Deduct 1 point if the method for generating the sequence of randomization was described and it was inappropriate | 59 (19) | 41 (13) |
| 2     | Was the study described as double blind? | 15 (5) | 85 (27) |
|       | Additional point if the method of double-blinding was described and it was appropriate | 15 (5) | 85 (27) |
|       | Deduct 1 point if the method of double-blinding was described and it was inappropriate | 15 (5) | 85 (27) |
| 3     | Was there a description of withdrawals and dropouts? | 68 (22) | 32 (10) |

*Yes=1, for a total of 5 possible points; ≥3 points indicates a superior quality trial. RCT, randomized controlled trial.
72. van Wyk JT, van Wijk MA, Sturkenboom MC, et al. Electronic alerts versus on-demand decision support to improve dyslipidemia treatment: a cluster randomized controlled trial. Circulation 2008;117:371–8.

73. Watson E, Clements A, Yudkin P, et al. Evaluation of the impact of two educational interventions on GP management of familial breast/ovarian cancer cases: a cluster randomised controlled trial. Br J Gen Pract 2001;51:817–21.

74. Weir CJ, Lees KR, MacWalter RS, et al. Cluster-randomized, controlled trial of computer-based decision support for selecting long-term anti-thrombotic therapy after acute ischaemic stroke. QJM 2003;96:143–53.

75. Whelan T, Levine M, Willan A, et al. Effect of a decision aid on knowledge and treatment decision making for breast cancer surgery: a randomized trial. JAMA 2004;292:435–41.

76. Wilson BJ, Torrence N, Mollison J, et al. Cluster randomized trial of a multifaceted primary care decision-support intervention for inherited breast cancer risk. Fam Pract 2006;23:537–44.

77. Campbell M, Fitzpatrick R, Haines A, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000;321:694–6.

78. Ukoumunne OC, Guilford MC, Chinn S, et al. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess 1999;3:iii–92.

79. Kernick D. Assessing complex interventions. Guidelines perpetuate inappropriate methods. BMJ 2008;337:a2473.

80. Ukoumunne OC, Guilford MC, Chinn S, et al. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess 1999;3:iii–92.

81. Campbell NC, Murray E, Darbyshire J, et al. Designing and evaluating complex interventions to improve health care. BMJ 2007;334:455–9.

82. Bond L, Craig P, Egan M, et al. Evaluating complex interventions. Health improvement programmes: really too complex to evaluate? BMJ 2010;340:c1332.

83. Mackenzie M, O’Donnell C, Halliday E, et al. Do health improvement programmes fit with MRC guidance on evaluating complex interventions? BMJ 2010;340:c185.

84. Shiell A, Have P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. BMJ 2008;336:1281–3.

85. Thomson R. Evidence based implementation of complex interventions. BMJ 2009;339:b3124.

86. West PA. Evaluating complex interventions. What is and what could be. BMJ 2010;340:c1340.

87. Friedman CP. Where’s the science in medical informatics? J Am Med Inform Assoc 1995;2:65–7.

88. Naunheim MR, Walcott BP, Nahed BV, et al. The Quality of Randomized Controlled Trial Reporting in Spine Literature. Spine (Phila Pa 1976) 2011;36:1226–30.

89. Reveiz L, Cortes-Jofre M, Asernio Lobos C, et al. Influence of trial registration on reporting quality of randomized trials: study from highest ranked journals. J Clin Epidemiol 2010;63:1216–22.

90. Mills EJ, Wu P, Gagnier J, et al. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. Contemp Clin Trials 2005;26:480–7.