Journal of the American Medical Informatics Association, 26(10), 2019, 1120–1128
doi: 10.1093/jamia/ocz035
Advance Access Publication Date: 17 April 2019
Review

Review

A review of measurement practice in studies of clinical decision support systems 1998–2017

Philip J. Scott,[1] Angela W. Brown,[1] Taiwo Adedeji,[1] Jeremy C. Wyatt,[2] Andrew Georgiou,[3] Eric L. Eisenstein,[4] and Charles P. Friedman[5]

[1]Centre for Healthcare Modelling and Informatics, University of Portsmouth, Portsmouth, UK, [2]Wessex Institute of Health Research, University of Southampton, Southampton, UK, [3]Australian Institute of Health Innovation, Macquarie University, Sydney, Australia, [4]Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina, USA, and [5]Schools of Medicine, Information and Public Health, University of Michigan, Ann Arbor, Michigan, USA

Corresponding Author: Philip J. Scott, PhD, Centre for Healthcare Modelling and Informatics, Room 1.37, Buckingham Building, University of Portsmouth, Portsmouth PO1 3HE, UK (Philip.scott@port.ac.uk)

Received 1 November 2018; Revised 20 February 2019; Editorial Decision 6 March 2019; Accepted 8 March 2019

ABSTRACT

Objective: To assess measurement practice in clinical decision support evaluation studies.
Materials and Methods: We identified empirical studies evaluating clinical decision support systems published from 1998 to 2017. We reviewed titles, abstracts, and full paper contents for evidence of attention to measurement validity, reliability, or reuse. We used Friedman and Wyatt's typology to categorize the studies.
Results: There were 391 studies that met the inclusion criteria. Study types in this cohort were primarily field user effect studies ($n = 210$) or problem impact studies ($n = 150$). Of those, 280 studies (72%) had no evidence of attention to measurement methodology, and 111 (28%) had some evidence with 33 (8%) offering validity evidence; 45 (12%) offering reliability evidence; and 61 (16%) reporting measurement artefact reuse.
Discussion: Only 5 studies offered validity assessment within the study. Valid measures were predominantly observed in problem impact studies with the majority of measures being clinical or patient reported outcomes with validity measured elsewhere.
Conclusion: Measurement methodology is frequently ignored in empirical studies of clinical decision support systems and particularly so in field user effect studies. Authors may in fact be attending to measurement considerations and not reporting this or employing methods of unknown validity and reliability in their studies. In the latter case, reported study results may be biased and effect sizes misleading. We argue that replication studies to strengthen the evidence base require greater attention to measurement practice in health informatics research.

Key words: health informatics, clinical decision support systems, measurement, validity, reliability

INTRODUCTION

Measurement is fundamental to empirical science and requires instruments that are valid and reliable: that provide reproducible results and measure what they claim to measure. For this reason, researchers use preexisting measurement instruments wherever possible and typically only develop their own instruments when there is no existing suitable instrument or when they are measuring a new construct unaddressed in published research. Before using a new instrument, investigators should carry out measurement studies that explore whether the methods are acceptably reliable and valid.[1] If these are absent, investigators must proceed carefully based only on assumptions about what their measurements mean. Just as poor

study design or inadequate sample size can jeopardize the integrity of a study, so too can measurements that are to a significant extent unreliable or invalid.[2]

It has been suggested that health informatics has a paucity of well-known and consistently used research constructs with established instruments for measuring them.[3] A robust library of reusable instruments creates an infrastructure for research that facilitates the work of study design, strengthens the internal and external validity of studies, and facilitates systematic reviews. Without this infrastructure, the health informatics evidence base will be weak and knowledge will not cumulate.[4–6] In other disciplines such as the behavioral sciences, there are bibliographic databases of measurement instruments,[7] and researchers are trained to use existing instruments with known validity and reliability whenever possible.[8] Previously validated instruments may require adaptation for changed circumstances, but, whether utilizing an existing instrument or developing a new one, explicit attention to measurement is important to the conduct and reporting of research.

Reliability and validity are core aspects of measurement.[9] Carmines and Zeller define *reliability* as "the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials" and *validity* as the extent to which an indicator "measures what it purports to measure."[10] An instrument can be reliable without being valid, but an instrument can only be valid if it is first found to be reliable.[2] Reliability assessment requires readministration of the instrument on successive occasions or a study of the internal consistency of independent observations within a measurement process. There are differing approaches to validity assessment but assessment of validity always requires use of external standards. *Face validity*, a relatively weak indicator, employs subjective assessments by experts of whether a measure appears to include all relevant facets of a construct and to measure what is intended. Assessment of *criterion–related validity* requires 1 or more external standards against which the measure should either be highly correlated or not correlated. Assessment of *construct validity* is the strongest approach but requires multiple additional constructs to be assessed revealing a pattern of correlations with the measure from which validity can be inferred.[1] However, validity is not simply a property of an instrument but arises from a combination of data collected when the instrument is used in the context and with the population for which it was intended.[11]

A previous review explored measurement practice in health informatics studies employing quantitative methods.[12] A significant majority of those studies addressed clinical decision support systems, examining 3 indicators of measurement practice: attention to the reliability of measures employed, the validity of those measures, and reuse of pre-existing instruments. In that review, of the 27 studies meeting the inclusion criteria, 3 reported reliability indices, and 8 suggested reuse of measurement methods, the majority of which were reused within the same research group. None of the studies explicitly considered the validity of the measurements employed.

OBJECTIVE

This work extends the previous study[12] by examining a significantly larger body of articles using the same indicators of attention to measurement with a specific focus on studies of clinical decision support systems used by medically qualified practitioners (specifically, physicians or surgeons). While not providing an exact comparison with the previous study, this review will help indicate whether attention to measurement practice in health informatics has changed over time. To provide a more detailed analysis, this paper describes the spectrum of study types in the published literature using Friedman and Wyatt's typology[1] and examines the extent to which explicit attention to measurement is associated with the study type. Related work[13,14] has reported development of an inventory of measurements applicable in health informatics but apparently without quality assessment of attention to measurement practice.

This aim of this study is to address 3 research questions (RQs):

- RQ1 – What fraction of a cohort of studies of clinical decision support systems (CDSS) used by medically qualified practitioners have indicators of measurement reliability, validity, or reuse?
- RQ2 – What is the distribution of study types within this cohort?
- RQ3 – To what extent is attention to measurement reliability, validity, or reuse related to study type?

MATERIALS AND METHODS

Search strategy

We identified a cohort of published studies and developed criteria to assess the 3 categories of attention to measurement. We applied the criteria to data extracted from each study to address RQ1. We categorized the specific study types to address RQ2. We examined the association between study type and evidence of attention to measurement to address RQ3, using the nonparametric Kruskal-Wallis test in IBM SPSS version 25.

We first conducted a search to identify CDSS system evaluation studies, using the PubMed database (given our focus on usage by medically qualified practitioners). We selected articles written in English that had abstracts, were classified as clinical trials, and published between January 1998 and December 2017. We limited our search in this way based upon the fact that studies classed as clinical trials would reasonably be assumed to be ones where mature measurement practice might be found. The MeSH terms used in the previous study[12] directed this search: medical records systems, computerized; decision support systems, clinical; hospital information systems; therapy, computer assisted; diagnosis, computer assisted. Due to the high volume of results, we further restricted some searches to MeSH major topics.[15] To complement the MeSH search strategy, we identified 3 seed papers[16–18] from earlier work and conducted a "snowball" search from their references.

Inclusion criteria

We manually filtered the search results based on title and abstract. We included studies that examined CDSSs used by a medically qualified practitioner, such as a physician or surgeon. Studies that stated "clinician" use were included if it could be reasonably assumed that clinician referred to a medically qualified practitioner. In this review, CDSSs are defined as computer systems that utilize patient data to provide timely patient-specific information or advice to support decision making.[19] Example systems are computerized alerts or reminders, computerized templates, order sets or clinical guidelines, diagnostic support, and other relevant information supplied to the physician to facilitate decision making.

We excluded studies about medical devices, decision aids used by patients, telemedicine studies (unless a CDSS was involved), study protocols, and systems used by health care professionals other than medically qualified practitioners. Studies where only a minor part of the intervention involved a CDSS were also excluded. Developmental IT system validation studies were also excluded.

Table 1. Classifications of generic study types by broad study questions and the stakeholders concerned,[1] with kind permission from Springer Science and Business Media. © Springer Science and Business Media, Inc. 2006.

	Study type	Aspect studied	Broad study question	Audience/stakeholders primarily interested in results
1	Needs assessment	Need for the resource	What is the problem?	Resource developers, funders of the resource
2	Design validation	Design and development process	Is the development method in accord with accepted processes?	Funders of the resource; professional and governmental certification agencies
3	Structure validation	Resource static structure	Is the resource appropriately designed to function as intended?	Professional indemnity insurers, resource developers, professional and governmental certification agencies
4	Usability test	Resource dynamic usability and function	Can intended users navigate the resource so it carries out intended functions?	Resource developers, users
5	Laboratory function study	Resource dynamic usability and function	Does the resource have the potential to be beneficial?	Resource developers, funders, users, academic community
6	Field function study	Resource dynamic usability and function	Does the resource have the potential to be beneficial in the real world	Resource developers, funders users
7	Laboratory user effect study	Resource effect and impact	Is the resource likely to change behavior?	Resource developers and funders, users
8	Field user effect study	Resource effect and impact	Does the resource change user actual user behavior in ways that are positive?	Resource users and their clients, resource purchasers and funders
9	Problem impact study	Resource effect and impact	Does the resource have a positive impact on the original problem?	The universe of stakeholders

Criteria for attention to measurement

We based our general approach on the methods used in the previous review, as we had the same aim to explore attention to reliability, validity, or instrument reuse (RQ1).[12] We defined *reliability indicators* as the explicit report of any measure of reliability associated with a method, measure, or instrument within the study or explicit reference to separately published reliability indices. We defined *validity indicators* as reported validation methods within the study or explicit reference to separately published validation methods. We excluded studies that solely employed clinical and laboratory measures that might reasonably be presumed to be valid, but where the paper did not otherwise demonstrate attention to measurement. We defined *reuse indicators* as the presence of any statement in which a study utilized a measurement instrument or method (in whole or in part) derived from previous work, whether published or not, and regardless of authorship.

Identifying and appraising the variables measured

We evaluated the measurement indicators in each study considering both primary and secondary outcomes if they were explicitly stated as such. Where this was not stated and it was unclear from the text, we made an assessment of what measures to include from the study objectives, data analysis, and results sections of the article. To ensure consistent data extraction and to calibrate our assessment, we examined the reliability of our appraisal of measurement indicators using Cohen's kappa.[20]

Assessment of measurement indicators

We searched the manuscripts for measurement indicators by determining if they contained any keywords relating to validity and reliability, namely: validity (construct, criterion, concurrent, predictive, content, face, divergent, discriminant, convergent); reliability (inter-rater/abstractor/coder, kappa, Cronbach's alpha, Spearman-Brown, test-retest reliability, and agreement); and synonyms, such as accuracy and precision. Apart from a few papers that we had to obtain as hard copies through inter-library loans, we executed this as an electronic search of the full text.

As a formative exercise to calibrate our assessment of measurement indicators, we calculated Cohen's kappa[20] from 50 randomly selected studies independently reviewed by a second rater. Following this calibration exercise, the bulk of the assessments were made independently by the 2 research assistants. The assessments were reviewed by the lead author, but no further inter-rater reliability calculation was made.

Study type categorization

After classifying studies according to evidence of measurement practice, we categorized them using Friedman and Wyatt's typology[1] to assess whether measurement indicators were associated with specific study types. This framework consists of 9 study types distinguished on the basis of the aspect of the information system ("resource") studied, the study question, and the audience most interested in the results (Table 1).

RESULTS

Literature review

Figure 1 summarizes the literature review process and results. The initial corpus comprised 8780 articles. Title review and removal of duplicates reduced the number of articles to 926. The first snowball search based on seed papers resulted in 683 studies. The second snowball search, based on 36 systematic reviews, retrieved a total of 812 papers. The 3 search strategies —the MeSH-driven search and 2 snowball searches—thus resulting in a total of 2421 papers.

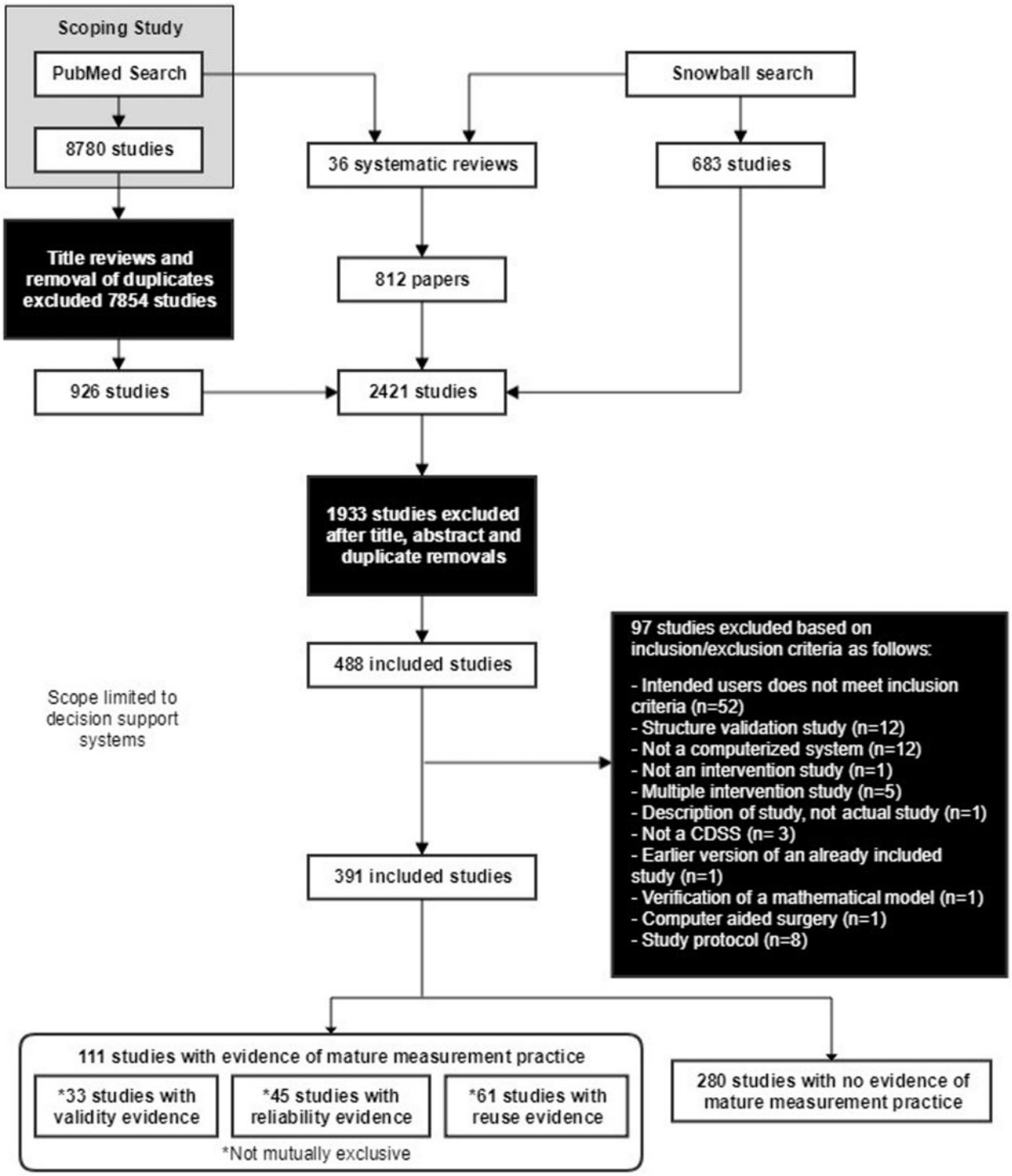

Figure 1. Literature review process and results.

The studies were then limited to decision support systems used by medically qualified practitioners, which excluded 1933 papers and left 488 in the corpus. Deduplication and further review of abstracts reduced the corpus to 391 studies. The large number of studies excluded for not meeting the intended user criteria was due to abstracts that failed to identify users of the system.

Reliability of the assessment of measurement indicators

The result of the formative inter-rater reliability assessment ($\kappa = 0.34$) is conventionally interpreted as "fair agreement," but showed room for improvement given that $\kappa = 0.41$–0.60 is considered "moderate agreement."[20] We then reviewed how we were

applying the criteria and explored the reasons for differing assessments. Following this calibration process, we reached agreement for all 50 studies in the sample set and the rest of the appraisals were made independently by the 2 research assistants (AB, TA).

Research question 1: indicators of measurement reliability, validity or reuse

We found measurement indicators in 111/391 studies (28%) listed in the supplementary file. It was also found that 45/391 (12%) had reliability indicators, 33/391 (8%) had validity indicators, and 61/391 (16%) had reuse indicators. These categorizations were not mutually exclusive, as shown in Figure 2. In 280/391 studies (72%) no evidence of measurement indicators was found.

Reliability

We found reliability indicators in 45 studies (12%). Evidence primarily comprised reported measurement of chance-corrected inter-rater agreement/reliability (Cohen's kappa) for the abstraction of data from medical records to facilitate measurement (eg, identifying the documentation of certain items, whether a particular test or adverse event had occurred, or for categorization purposes). A small number of studies measured inter-rater agreement with a percentage or employed other measures of reliability such as test-retest, intraclass correlation coefficient, Cronbach's alpha, or claimed reliability with no measurement given as shown in Table 2. The total number of instances is 50, as some studies reported more than 1 reliability measure.

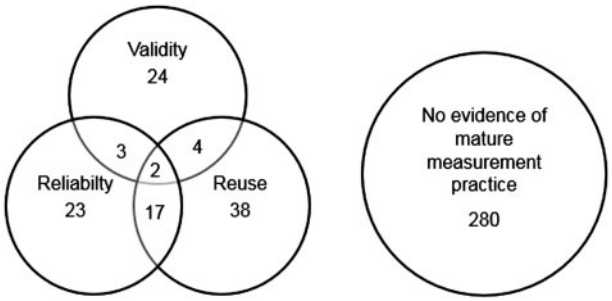

Figure 2. Measurement indicators in all included studies.

Table 2. Reliability indicators

Studies with Reliability Indicators	
Indicator	Instances
Inter-rater Cohen's kappa	28
Inter-rater percentage	8
Test-retest	1
Intraclass correlation coefficient	2
Cronbach's alpha	5
Claimed (no measurement specified)	6
TOTAL	**50**

Validity

From the 33 studies (8%) with validity indicators, we identified 68 distinct measurements. Of these, 63 had validity measured elsewhere, and 5 had validity measured within the study.

Most of the measures that had validity evidence that were patient health outcomes or process of care measures; only 6 were not:

- a continuous diagnostic quality score[a],[21] where validity had been assessed in a previous study;
- a composite quality score calculated for diagnostic and management plans[b],[22] which carried out a thorough validity and reliability assessment within the study;
- a known usability measure: the Standard Usability Score[c] was used by 2 studies[23,24];
- a survey measuring house staff attitudes toward CPOE[d] which had been face validated[25];
- the semantic differential power perception survey[e] which had also been shown to be valid in a previous study.[26]

Table 3 shows the categorization by measurement domain, with the alphabetic superscripts referencing the 6 measures listed previously.

Reuse

We found reuse of 68 measurement artefacts from 61 studies (some studies reused more than 1 artefact, others reused the same artefact). The majority of reused artefacts were modified instruments ($n = 13$). Of the reused instruments, 4 had evidence of reliability. A number of established methods for identifying and classifying adverse drug events were identified, most of which were internally reused by the same research group. A number of studies showed evidence of reuse of other artefacts as shown in Table 4.

Table 3. Valid measures by domain

Primary Category	User Measures			Patient Health			Process of Care		IT System	
Secondary Category	Physician Knowledge, attitudes, or beliefs.	Physician decisions or diagnostic/ therapeutic accuracy	Physician satisfaction or perceptions	Laboratory	Clinical Measure or Outcome	Patient Reported Outcome	Patient Safety	Patient Reported Experience	Usability/ usefulness	Total
Validity Measured Elsewhere		1[a]	1[e]		15	30	4	10	2[c]	**63**
Validity Measured in Study	2[d]	1[b]				1			1	**5**
Total	**2**	**2**	**1**	**0**	**15**	**31**	**4**	**10**	**3**	**68**

[a-e]Measures that were not categorized as patient health outcomes or process of care measures.

Research question 2 – categorization of study types

Using the typology, 6 types were identified in the cohort[1]: studies of usability, laboratory user effect, laboratory function, field function, field user effect, and problem impact. Figure 3 shows the study type distribution for the 391 included studies. Studies identified were predominantly field user effect and problem impact studies.

Research question 3: relationship of study type with measurement indicators

Figure 4 shows the distribution of study types by measurement indicators. The percentage is the proportion of studies with that indicator (or the absence of indicators).

Most studies were problem impact studies or field user effect studies. The Kruskal-Wallis test showed a significant association between study type and the presence of validity indicators ($P = .007$) and a significant association with the *absence* of mature measurement indicators ($P = .005$) but no significant association with reliability or reuse indicators. We interpret this to suggest a bi-modal distribution: while the majority of studies have no measurement indicators, there is a significant minority (mostly problem impact studies) that do address validity. Of course, there is an inherent bias in our sample: the predominant study types in the cohort reflect our decision to exclude developmental system validation studies.

Table 4. Reuse indicators

Studies with Reuse Indicators	
Reuse Artefact	No of Studies
Modified or un-validated instrument	23
Methodology (all or part)	17
Measurement	6
Categorization	8
Guideline/protocol	3
Criteria	8
Definition	3
TOTAL	**68**

DISCUSSION

We set out to answer 3 research questions on measurement practice in health informatics studies, focusing on CDSS evaluation. We found that 28% (111/391) of the eligible studies had some evidence of at least 1 of the 3 defined measurement indicators. Assessment of reliability was identified in 12% (45/391) of studies. However, the majority of these measurements did not directly assess the reliability of an instrument or measure, but demonstrated the reliability of data abstraction from medical records to facilitate measurement.

Validity evidence was identified in 8% (33/391) of studies, comprising 68 individual measures. However, the majority of valid measures (93%, 63/68) had no direct evidence of validity assessment indicated in the study. Only 5 studies (7%, 5/68) had evidence of direct measurement of validity.

Reuse of measurement artefacts was identified in 16% (61/391) of studies. The majority of these either modified previously valid instruments or reused instruments and methods where there was no indication of validity. Of these studies 38% (23/61) referenced additional measurement data such as reliability or validity. In the majority of studies in our cohort where an instrument had been modified, no evidence of the validity or reliability of the 'new' instrument was provided.

A direct comparison with the previous study of measurement practice in health informatics[12] cannot be made due to the different inclusion criteria and categorizations. However, this review echoes the earlier conclusion that measurement practice is immature in the field of health informatics. Our study included 18 of the 27 studies in the earlier review. Of the 9 not included, 5 were outside the defined date range and 4 did not meet our inclusion criteria. Identification and categorization differed in 8 studies due to the modified criteria for the measurement indicators.

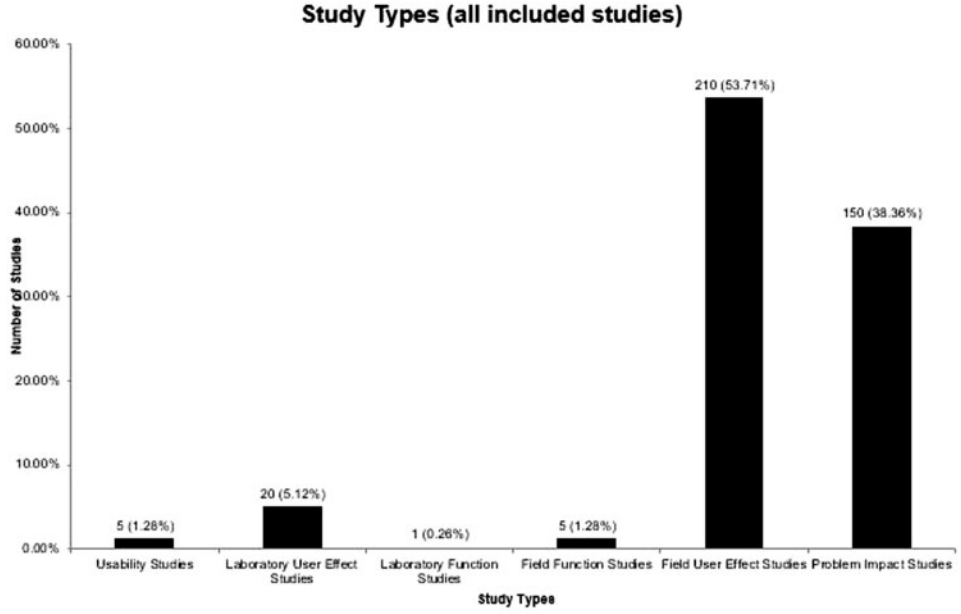

Figure 3. Distribution of study types (all included studies $n = 391$).

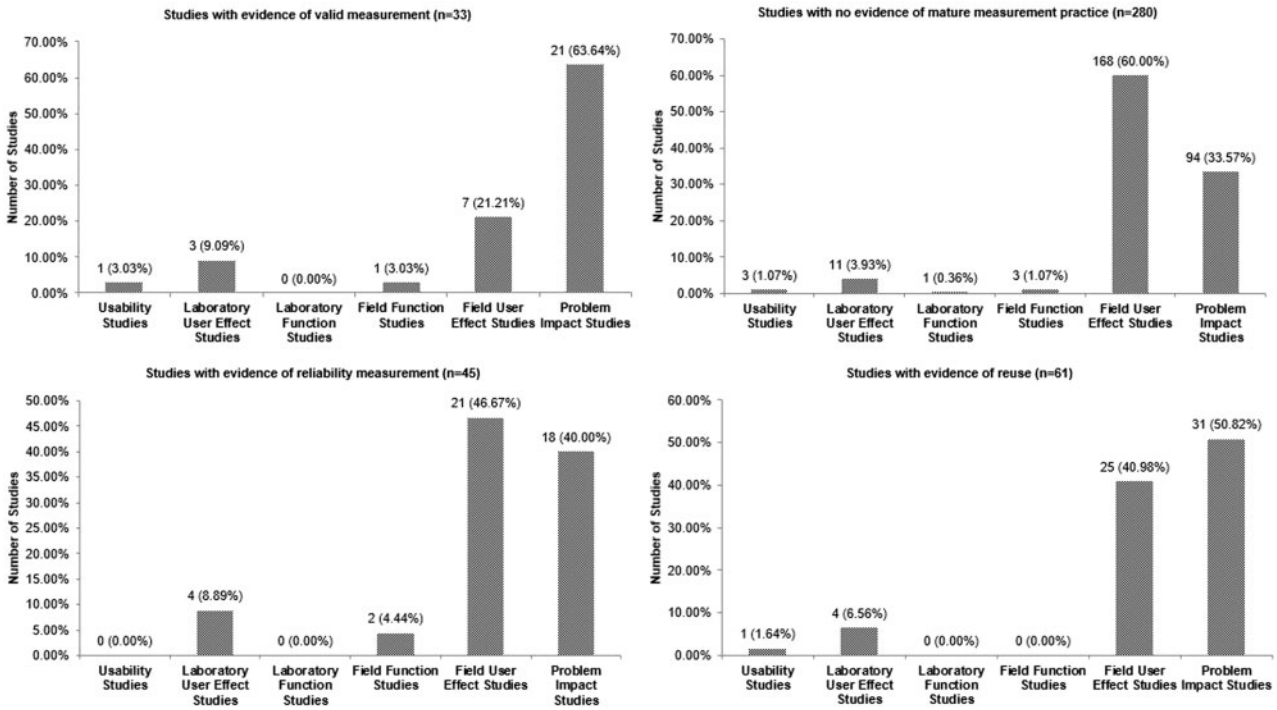

Figure 4. Distribution of study types by measurement indicators.

RQ2 addressed the prevalence of different study types in our study cohort. The study type analysis revealed that 54% (210/391) of the studies were field user effect studies and 38% (150/391) were problem impact studies. RQ3 addressed the relationship between study type and evidence of any of the 3 measurement indicators. There was a significant association between study type and validity indicators, but also between study type and absence of measurement indicators. Only 3% (7/210) of field user effect studies had validity indicators and only 14% (21/33) of problem impact studies.

In the 280 studies with no evidence of measurement practice, the most prevalent measures were behavioral, such as compliance measured using objective counts (eg, the number of compliant prescriptions). Even though these measures might be assumed to be perfectly reliable since they are "counts," the subjective construct of "compliance" raises the distinct possibility that multiple assessors of compliance might not agree. Further investigation of this type of measure needs to be undertaken to assess how and whether sources of error are being quantified and if more attention to good measurement practice is required in studies that employ these measures. The absence of good measurement practice does cast doubt on the extent to which the measured outcome is a true reflection of reality. The previous review[12] also stated that the measurement aspects of studies should be separate from the demonstration aspects in order for researchers to benefit from utilizing each other's measurement tools. Studies may not report reliability and validity measurements if researchers regard an instrument as well-known and authoritative (eg, established clinical scales). This is acceptable if the instrument is being used under the same conditions for which reliability and validity have previously been assessed; however, this should be clearly stated in the article. It is also necessary to account for attenuation, which will make measured effect sizes smaller than actual effect sizes due to measurement error.

It is immensely challenging to evaluate a unique health informatics system situated in an already complex environment that involves numerous variables.[27] However, unless our field begins to develop a range of valid well-understood measures, the evidence base will remain weak and incomplete. This methodological weakness is not unique to health informatics but appears to be common in other areas of health care evaluation.[28] Significant activities have been undertaken to work toward the goal of evidence-based health informatics,[29] however there is still progress to be made.

One of the difficulties we found was the varied and sometimes unclear reporting styles and language used when trying to describe measurement methods, identify evidence, and categorize studies. The European Federation for Medical Informatics guideline for Good Evaluation Practice in Health Informatics[30] and the associated Statement of Reporting of Evaluation Studies in Health Informatics[31,32] provide clear guidance on how to plan, perform, and publish a methodologically sound evaluation study, which includes explicit recommendations about attention to measurement issues. These resources can be combined with other standards such as the Consolidated Standards of Reporting Trials.[33] Comprehensive textbooks on health informatics evaluation and handbooks of methods exist to assist researchers to select the most appropriate methodology for the study being undertaken and explain the issues of measurement practice in detail.[1,34,35] Databases of measures exist for health care, such as the National Quality Measures Clearinghouse.[36] The Agency for Healthcare Research and Quality has published a small collection of health informatics evaluation measures,[37] and initiatives such as the Core Outcome Measures in Effectiveness Trials aim to establish agreed standardized sets of outcome measures.[6] These works have assisted in moving toward evidence-based health informatics.

Further work

An extension to this review could be the development of a database of measures for health informatics researchers, which would cover not only patient outcome or process of care measures but user, financial, system, and other aspects. Some work in this area has begun with a project to identify and evaluate measures for patient-facing technologies.[38] A further consideration is to identify and describe the theoretical foundations of validated measures.[13]

Limitations

This review is potentially limited by the use of only 1 database (PubMed). Given the defined scope, we did not search nursing bibliographic databases as studies were only included if used by a medically qualified practitioner. A previous systematic review of clinical decision support interventions that searched a number of databases found that all the studies included in the final study sample were also indexed and available in MEDLINE. Therefore, we also believe this limitation to be negligible.[39] However, we acknowledge that the selection of PubMed using the methodology employed here may not be completely reproducible over time.

A further limitation is that single researchers independently carried out the initial study selection, evidence assessment, and categorization process; however, formative inter-rater reliability assessment was carried out to mitigate this. We did not compare the distribution of study types in the inter-rater sample with the full set of papers, so there is a risk that the inter-rater reliability assessment was biased by an unrepresentative subset of studies.

This review has only looked at CDSS interventions; it is possible that other health informatics studies may demonstrate attention to measurement practice not identified here. The purpose of this review was purely to identify evidence—not to assess the quality of the evidence. It was also not our intention to assess the quality of the studies overall or to question the methodologies employed. Studies often do not clearly state who an intervention is used by, which can be problematic for non-medical researchers, and, in some cases, evidence is not clear and could be misinterpreted. We acknowledge that researchers may have carried out reliability or validity measurement but not reported this in their article.

CONCLUSION

We do not question that holistic evaluation requires mixed methods and a range of epistemological perspectives,[40,41] and that qualitative studies play an important role in addressing the why and how of health informatics interventions. However, we maintain that, as a basic scientific principle, any evaluation that reports quantitative results should give due consideration to sound measurement. This should be taken into account when designing the study, so that the evaluation is scoped and resourced as necessary to deliver robust results. Given suitable reuse, not every evaluation will need its own measurement study; but the limitations imposed by using any untested measures should always be acknowledged.

We argue that this review of outcomes in CDSS evaluation studies shows that attention to measurement practice remains weak. This review has also highlighted the prevalence of field user effect studies utilizing behavioral measures with little discussion of validity. We echo the recent call from Coiera and colleagues[6] to take seriously the scientific challenge facing our discipline: evidence-based health informatics requires replication studies to strengthen or question previous findings. This requires a toolset of validated and reliable measurement instruments.

We call on leaders in the health informatics field, researchers and funders, educators, professional bodies, and journal editors and referees to promote the practice of undertaking and reporting measurement studies in health informatics evaluation.

AUTHOR CONTRIBUTIONS

PS conceived and directed the review. All authors contributed to the study design, methods, discussion, conclusions and commented upon iterations of the complete text. AB and TA did the detailed searches, filtering and study categorisation. PS, AB and TA did the inter-rater reliability calibration. All authors agreed the final text.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Friedman CP, Wyatt JC. Evaluation of biomedical and health information resources. In: Shortliffe EH, Cimino JJ, editors. *Biomedical informatics*. London: Springer-Verlag; 2014: 355–87.
2. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm* 2008; 65 (23): 2276–84.
3. Friedman C, Wyatt J, Owens D. Evaluation and technology assessment. In: E. H. Shortliffe, J. J. Cimino, editors. *Biomedical Informatics*. New York: Springer; 2006: 403–43.
4. Clamp S, Keen J. Electronic health records: is the evidence base any use? *Med Inform Internet Med* 2007; 32 (1): 5–10.
5. Scott P, Prytherch D, Briggs J. Health *Informatics: Where's* the Evidence? Commissioned by the UK Faculty of Health Informatics, University of Portsmouth; 2010.
6. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? *J Am Med Inform Assoc* 2018: 25 (8): 963–8.
7. EBSCO. Health and Psychosocial Instruments (HaPI); 2018. https://www.ebsco.com/products/research-databases/health-and-psychosocial-instruments-hapi. Accessed June 1, 2018.
8. Langston W. *Research Methods Laboratory Manual for Psychology*. Belmont, CA: Cengage Learning; 2010.
9. Hammersley M. Some notes on the terms 'validity' and 'reliability'. *Br Educ Res J* 1987; 13 (1): 73–82.
10. Carmines EG, Zeller RA. *Reliability and Validity Assessment*. Vol. 17. Thousand Oaks, CA: Sage; 1979.
11. DiIorio CK. *Measurement in Health Behavior: Methods for Research and Evaluation*. Vol. 1. San Francisco, CA: John Wiley & Sons; 2006.
12. Friedman CP, Abbas UL. Is medical informatics a mature science? A review of measurement practice in outcome studies of clinical systems. *Int J Med Inform* 2003; 69 (2–3): 261–72.
13. Colicchio TK, Del Fiol G, Scammon DL, *et al*. Development and classification of a robust inventory of near real-time outcome measurements for assessing information technology interventions in health care. *J Biomed Inform* 2017; 73: 62–75.
14. Colicchio TK, Facelli JC, Del Fiol G, *et al*. Health information technology adoption: Understanding research protocols and outcome measurements for IT interventions in health care. *J Biomed Inform* 2016; 63: 33–44.
15. US National Library of Medicine. Medical Subject Headings (MeSH) in MEDLINE/PubMed: A Tutorial: Major Topics; 2012. http://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/majortopics/. Accessed January 9, 2014.
16. Black AD, Car J, Pagliari C, *et al*. The impact of eHealth on the quality and safety of health care: a systematic overview. *PLoS Med* 2011; 8 (1): e1000387.

17. Cork RD, Detmer WM, Friedman CP. Development and initial validation of an instrument to measure physicians' use of, knowledge about, and attitudes toward computers. *J Am Med Inform Assoc* 1998; 5 (2): 164–176.
18. Liu JL, Wyatt JC. The case for randomized controlled trials to assess the impact of clinical information systems. *J Am Med Inform Assoc* 2011; 18 (2): 173–180.
19. Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: potential problems and solutions. *Proc Ann Symp Comp Appl Med Care* 1991: 3–7.
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33 (1): 159–74.
21. Friedman CP, Elstein AS, Wolf FM, *et al*. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *JAMA* 1999; 282 (19): 1851–6.
22. Ramnarayan P, Kapoor RR, Coren M, *et al*. Measuring the impact of diagnostic decision support on the quality of clinical decision making: development of a reliable and valid composite score. *J Am Med Inform Assoc* 2003; 10 (6): 563–572.
23. Martins SB, Shahar Y, Galperin M, *et al*. Evaluation of KNAVE-II: a tool for intelligent query and exploration of patient data. *Stud Health Technol Inform* 2004; 107 (Pt 1): 648–52.
24. Martins SB, Shahar Y, Goren-Bar D, *et al*. Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data. *Artif Intell Med* 2008; 43 (1): 17–34.
25. Rosenbloom ST, Talbert D, Aronsky D. Clinicians' perceptions of clinical decision support integrated into computerized provider order entry. *Int J Med Inform* 2004; 73 (5): 433–441.
26. Bartos C, Butler B, Penrod L, Fridsma D, Crowley R. Negative CPOE attitudes correlate with diminished power in the workplace. *AMIA Ann Symp Proc* 2008; 2008: 36–40.
27. Koppel R. Is healthcare information technology based on evidence? *Yearb Med Inform* 2013; 8 (1): 7–12.
28. Lopetegui M, Bai S, Yen P, Lai A, Embi P, Payne P. Inter-observer reliability assessments in time motion studies: the foundation for meaningful clinical workflow analysis. *AMIA Annu Symp Proc* 2013; 2013: 889–96.
29. Rigby M, Ammenwerth E, Beuscart-Zephir M-C, *et al*. Evidence based health informatics: 10 years of efforts to promote the principle. *Yearb Med Inform* 2013: 34–46.
30. Nykänen P, Brender J, Talmon J, *et al*. Guideline for good evaluation practice in health informatics (GEP-HI). *Int J Med Inform* 2011; 80 (12): 815–827.
31. Brender J, Talmon J, de Keizer N, *et al*. Statement on Reporting of Evaluation Studies in Health Informatics: explanation and elaboration. *Appl Clin Inform* 2013; 4 (3): 331–358.
32. Talmon J, Ammenwerth E, Brender J, *et al*. STARE-HI—statement on reporting of evaluation studies in health informatics. *Int J Med Inform* 2009; 78 (1): 1–9.
33. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med* 2010; 8 (1): 18.
34. McNair JB. *Handbook of Evaluation Methods for Health Informatics*. Burlington, MA: Academic Press; 2006.
35. Ammenwerth E, Rigby M, eds. *Evidence-Based Health Informatics: Promoting Safety and Efficiency through Scientific Methods and Ethical Policy*. Vol. 222. Amsterdam: IOS Press; 2016.
36. Agency for Healthcare Research and Quality (AHRQ). *National Quality Measures Clearinghouse*. http://www.qualitymeasures.ahrq.gov/index.aspx. Accessed September 2014.
37. Agency for Healthcare Research and Quality, A. *Health IT Evaluation Measures: Quick Reference Guides*; 2014. http://healthit.ahrq.gov/health-it-tools-and-resources/health-it-evaluation-measures-quick-reference-guides. Accessed September 17, 2014.
38. Wakefield BJ. The eHealth measures compendium. *HSRD FORUM* 2014: 4.
39. Fillmore CL, Bray BE, Kawamoto K. Systematic review of clinical decision support interventions with potential for inpatient cost reduction. *BMC Med Inform Dec Making* 2013; 13 (1): 135.
40. Scott P, Briggs JS. A pragmatist argument for mixed methodology in medical informatics. *J Mixed Methods Res* 2009; 3 (3): 223–41.
41. Klecun E, Lichtner V, Cornford T, *et al*. Evaluation as a multi-ontological endeavour: a case from the English National Program for IT in healthcare. *JAIS* 2014; 15 (3): 147–76.