Unpacking the Evaluation Proceeding of Clinical Decision Support Systems: A review of methodological approaches and categories

Ali Azadi [1] **0000-0002-7166-1565 , Francisco José García-Peñalvo** [1] **0000-0001-9987-5584**

[1] GRIAL Research Group, Computer Science Department, University of Salamanca, Salamanca, Spain

Abstract

Medical personnel must utilize Clinical Decision Support Systems (CDSS) to enhance clinical decision-making, minimize mistakes, and improve patient outcomes. Accurately evaluating the performance of CDSS is essential to avouch their effectiveness and efficiency. We have reviewed the literature to provide insights into evaluating CDSS, along with the criteria that need to be assessed, such as accuracy, usability, and efficiency. Researchers are instructed to pick an acceptable technique depending on their research aims and the situation in which they will analyze Clinical decision support systems after considering potential obstacles and constraints within the procedure. By conducting these types of research projects, we will be able to improve the quality of the decision-support systems and enhance their utility in clinical practice. This article provides valuable intuition for researchers, healthcare professionals, and decision-makers seeking to evaluate the performance of CDSS in healthcare settings.

Keywords: Clinical Decision Support Systems, Evaluation, Methodology

1. Introduction

Computer programs to boost medical decision-making have long been anticipated by physicians with both curiosity and accuracy [1]. It has been widely recognized that evaluating clinical decision support systems is both a vital component of the larger area of medical computing and a challenging and diverse topic unto itself [2] and has been considered a must [3], hence, to ensure that the Clinical Decision Support Systems (CDSS) have been effective at improving patient care and safety, raising the standard of care, lowering healthcare costs, and boosting healthcare providers' productivity, evaluation is crucial [4]. In other words, evaluation is not only exploited to assess the effectiveness of a program but can also be applied to measure the program's evolution over time [5].

Now that it is obvious why the CDSS evaluation must be done, to guarantee the results' reliability and validity and allow for the generalizability of findings to different contexts, it is imperative to employ the proper evaluation methods [6][7]. These methods, analyses of their comparative effectiveness, and their use have gradually evolved in various studies [8]. According to the predefined aims, the CDSS evaluation can be carried out in two main aspects: usability evaluation and accuracy evaluation, although some other categories have been recommended.

This paper mentions some of the principal methods to evaluate clinical decision support systems. These methods have been classified regarding the nature of the assessment and its purpose into two major categories: usability assessment and evaluating accuracy level. The fruitful methods included in each of the mentioned categories will be described. After getting familiar with the available methods, we will conclude. Since this type of classification of the CDSS assessment has not been performed so far, this study can contribute to future research projects.

28th Conference on Information Society and University Studies (IVUS'2023), May 12, 2023, Kaunas, Lithuania

EMAIL: ali.azadi@usal.es (A. Azadi);

CEUR Workshop Proceedings (CEUR-WS.org)

2. Evaluation of Clinical Decision Support systems

Nowadays, the authors may be focusing on measurement considerations rather than reporting this in their study or using methods of unknown assessment [9] opting for the optimum methodology for evaluating CDSS is a critical point in investigating these systems properly and perfectly and in different dimensions. The main aspects which have been addressed in the studies include usability, accuracy level, reliability, and validity.

2.1. Usability methods

Usability is the quality of a user's experience when interacting with products or systems, including websites, software, devices, or applications[3]. Some studies employ a single method and others combined methods to assess usability. In this article, both single and combined methods have been explained.

2.1.1. Single-one methods

The main (single) methods to evaluate the usability facets have been addressed as the following:

Think Aloud: A direct observational method of user testing where users are asked to think aloud while completing a task. Users are asked to say what they are looking at, thinking, doing, and feeling at any moment [10]. This technique is particularly useful for determining user expectations and identifying confusing aspects of the system [11].

Near live: During "Near Live" testing, participants can interact naturally with patient actors. Every participant demonstrates the same workflow regarding the order of events [12].

Heuristic evaluation: Nielsen and Molich [13] introduced a new method for evaluating user interfaces called heuristic evaluation. In this method, a small group of usability experts evaluates user interfaces against a set of guidelines, noting the severity and presence of each usability problem. Andrea et al. [14] applying this method, have reinforced the CARTIER-IA platform pertaining to incorporating medical data (both structured data and images) through actuating Artificial Intelligence algorithms.

Cognitive walkthrough: The cognitive walkthrough is a usability inspection method that evaluates the design of a user interface for its ease of exploratory learning based on a cognitive model of learning and use [15].

Pluralistic usability walkthrough: The multidimensional usability walkthrough adapted the traditional usability walkthrough to include representative users, product developers, product team members, and usability experts in the process[16]. This is defined by her five characteristics:

- Inclusion of representative users, product developers, and human factors professionals.
- The application's screens are presented in the same order as they appear to the user.
- All participants are asked to assume the role of the user.
- Participants write down what actions they, as users, would take for each screen before the group discusses the screens.
- When discussing each screen, the representative users speak first.

Formal usability inspections: A formal usability test reviews a user's potential task performance by an interface designer and their peers. As with multidimensional usability walkthroughs, this involves stepping through the user's tasks [8]. However, because the reviewers are made up of human factors experts, the review can be quicker, more thorough, and more technical than a multidimensional walkthrough. The aim is to identify the maximum number of errors in the interface as efficiently as possible.

Quick and dirty usability testing: John Brooke [17] has claimed in his study that each tool or system's usability must be evaluated in terms of the context in which it is used and its suitability for that context. He explained that, generally, it is impossible to characterize a system's usability without first identifying its intended users, the tasks they will carry out with it, and the features of the physical, organizational, and social environments in which they will use it. He has explained that SUS is a Likert scale. It is usually assumed that a Likert scale is just a type of forced-choice question, where the responder is asked a statement and is then asked to rate how much they agree or disagree with it on a scale of 5 (or 7). He has suggested a questionnaire including 10 questions to measure the system usability scale. In this method, the SUS score will be calculated as follows: first, sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1,3,5,7 and 9, the score contribution is the scale position minus 1. For items 2,4,6,8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SU. SUS scores have a range of 0 to 100.

2.1.2. Multimodal Methods

Based on the research necessities, several methods have been formed to assess the system´s usability in various dimensions and in a stricter manner. In this section we have addressed two multimodal ones:

Development and design approaches (mixed methods): In this paper conducted by Horsky et al.[18] has been addressed with a set of useful suggestions and references to resources that may be utilized to guide the development of Clinical Decision Support Systems, to achieve the best possible human-computer interaction properties. They have claimed that the optimal design approaches for CDSS developers comprise iterative development, user-centered design, collaborative design teams, usability inspection, clinician interviews, log analysis, and cognitive walkthrough, as the principal components.

Since in this study, several aspects have been pointed out to evaluate the CDSS, it is considered a mixed method.

Integrating think-aloud and near-live: In [19] has been innovated a usability method through integrating two other methods, including “near-live” and “think-aloud,” which have been discussed above. In this study, the two phases of evaluation have been explained before establishing the deployment of the integrated clinical prediction rules and clinical decision support. Phase I involved usability testing associated with “think-aloud” protocol analysis to evaluate human–computer interaction as the healthcare providers performed specific tasks for invoking the CDSS [20], [21]. Phase II involved a “near-live” clinical simulation in assessing how stakeholders interact with the CDSS while interviewing a simulated patient [22]. They have demonstrated that both types of testing offer various insightful perspectives essential for the successful development and integration of CDSS in Electronic Health Records.

2.2. Accuracy level methods

Another perspective that can be investigated in CDSS is the accuracy and reliability level. By assisting with tasks like diagnosis, decision-making, and ordering tests and treatments, accurate CDSS may cut down on wasteful spending and boost the standard of care. They serve as a (basic) support system, but experts continue to have the final say in all decisions [23]–[25].

Statistical analysis for happened errors: In [26] Chantal et al. proposed a statistical method to compare the error cases that have taken place to calculate a risk score manually and by CDSS. In this research, a retrospective analysis was exploited to determine the degree of correlation for the score criteria: hypertension, diabetes, thromboembolic disorders (cerebrovascular accident, transient ischemic attack, long embolus, and deep venous thrombosis), heart failure, symptomatic arteriosclerosis in the legs and symptomatic coronary disease. In this study, A Bland-Altman plot and regression analysis has been used to visualize the agreement between two different interventions (automated CDSS, which is called aCDSS, and manual CDSS, which is called mCDSS). This study demonstrated that calculations performed by an aCDSS might be more accurate and time efficient than a manual calculation.

Positive and Negative predictive value: In this method, some of the monitoring and critical values have been defined to calculate accuracy, sensitivity, and specificity and to review other screening performance characteristics, including positive and negative predictive values (PPV and NPV). PPV and NPV are true positive and negative results of a diagnostic test, respectively [27]. In other words, in a certain diagnosis process by the test, predictive values explain how probable it is for the diagnosis to be correct. Safari et al. [28] have managed a study and defined some expressions to clarify the accuracy measurement as the following:

- True positive (TP)= the number of cases correctly identified as patient
- False positive (FP) = the number of cases incorrectly identified as patient
- True negative (TN) = the number of cases correctly identified as healthy
- False negative (FN) =the number of cases incorrectly identified as healthy

Altman and Bland [29] proved that positive predictive value is the proportion of cases giving positive test results which are already patients. They have expressed that it is the ratio of patients truly diagnosed as positive to all those with positive test results (including healthy subjects who were incorrectly diagnosed as patients). These criteria will be able to predict how likely it is for a person to truly be patient in case of a positive test result and has been formulated in this study as below:

$$\textit{Positive predictive value (PPV)}= \frac{TP}{TP+FP} \qquad \textit{Negative predictive value (NPV)}= \frac{TN}{TN+FN}$$

Robert Trevethan [30], in another article, has determined the mentioned criteria but with to some extent different expressions. He has utilized the expressions “sensitivity” and “specificity” and explained them:

Sensitivity*:* The sensitivity of a screening test will be described in various manners, often such as sensitivity being the ability of a screening test to detect a true positive, being based on the true positive rate, reflecting a test’s ability to correctly identify all persons who have a circumstance, or, if 100%, identifying all persons with a condition of interest by those people testing positive on the test.

Specificity*:* The specificity of a test is defined in a variety of manners, usually such as specificity being the ability of a screening test to detect a true negative, being based on the true negative rate, correctly identifying the persons who do not have a circumstance or, if 100%, identifying all patients who do not have the condition.

Robert Trevethan has concluded that Sensitivity and specificity must be emphasized as having different origins, and purposes, from PPVs and NPVs. All four metrics should be considered substantial when explaining and evaluating a screening test’s adequacy and usefulness.

Comparison with Golden Standard: This method defines a golden standard for a specific test to compare other automated medical functionalities with the ideal one. In other words, an expert-prepared "gold standard" for making diagnoses was verified in an earlier investigation [31]. The mentioned manner has been exploited in the research project by Helena et al [32].

Indeed, this study was Cross-sectional descriptive with a quantitative approach. In this investigation, The Wilcoxon nonparametric test was employed to compare two paired samples and is considered the number of differences (between the gold standard and the data extracted).

These kinds of methods need to be ascertained as a golden standard by the experts obsessively and precisely otherwise, the evaluation proceeds and its result will not be reliable.

3. Discussion and Conclusion

It can be claimed that assessing Clinical Decision Support Systems is a crucial undertaking that necessitates careful consideration of the most effective assessment techniques. Usability and accuracy assessment, the two main components of CDSS evaluation included in this study, are essential for assuring the successful application of CDSSs in healthcare settings, and experts in this field must carefully consider and choose assessment techniques that are appropriate for the CDSS being assessed. In other words, if CDSSs are assessed successfully, will lead to widespread acceptance and improved patient outcomes. These kinds of studies, by collecting various evaluation methods, categorizing, and comparing their exclusive attributes, will help the assessors to opt for the most optimized option according to the circumstances and limitations of the study.

4. Acknowledgments

I would like to express my sincere thanks to my supervisor, Professor Francisco José García-Peñalvo, for his invaluable guidance and support throughout the preparation of this study. His expertise and insights were instrumental in shaping the direction and focus of the research.

5. References

[1] E. H. Shortliffe, “Computer programs to support clinical decision making.,” *JAMA*, vol. 258, no. 1, pp. 61–66, Jul. 1987.

[2] P. L. Miller and D. F. Sittig, “The evaluation of clinical decision support systems: What is necessary versus what is interesting,” *Inform Health Soc Care*, vol. 15, no. 3, pp. 185–190, 1990, doi: 10.3109/14639239009025266.

[3] J. M. Toribio-Guzmán, A. García-Holgado, F. Soto Pérez, F. J. García-Peñalvo, and M. Franco Martín, “Usability Evaluation of a Private Social Network on Mental Health for Relatives,” *J Med Syst*, vol. 41, no. 9, 2017, doi: 10.1007/s10916-017-0780-x.

[4] T. M. Rawson *et al.*, “A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately?,” *Clinical Microbiology and Infection*, vol. 23, no. 8, pp. 524–532, 2017, doi: 10.1016/j.cmi.2017.02.028.

[5] R. E. Glasgow, T. M. Vogt, and S. M. Boles, "Evaluating the public health impact of health promotion interventions: the RE-AIM framework.," *Am J Public Health*, vol. 89, no. 9, pp. 1322–1327, Sep. 1999, doi: 10.2105/AJPH.89.9.1322.

[6] C.-P. Lin, T. H. Payne, W. P. Nichol, P. J. Hoey, C. L. Anderson, and J. H. Gennari, "Evaluating clinical decision support systems: monitoring CPOE order check override rates in the Department of Veterans Affairs' Computerized Patient Record System.," *J Am Med Inform Assoc*, vol. 15, no. 5, pp. 620–626, 2008, doi: 10.1197/jamia.M2453.

[7] T. Schleyer, H. Spallek, and P. Hernández, "A Qualitative Investigation of the Content of Dental Paper-based and Computer-based Patient Record Formats," *Journal of the American Medical Informatics Association*, vol. 14, no. 4, pp. 515–526, 2007, doi: 10.1197/jamia.M2335.

[8] T. Hollingsed and D. G. Novick, "Usability inspection methods after 15 years of research and practice," *SIGDOC'07: Proceedings of the 25th ACM International Conference on Design of Communication*, no. October 2007, pp. 249–255, 2007, doi: 10.1145/1297144.1297200.

[9] P. J. Scott *et al.*, "A review of measurement practice in studies of clinical decision support systems 1998-2017," *Journal of the American Medical Informatics Association*, vol. 26, no. 10, pp. 1120–1128, 2019, doi: 10.1093/jamia/ocz035.

[10] M. J. Van Den Haak, M. D. T. De Jong, and P. J. Schellens, "Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue," *Behaviour and Information Technology*, vol. 22, no. 5, pp. 339–351, 2003, doi: 10.1080/0044929031000.

[11] S. Richardson *et al.*, "'Think aloud' and 'Near live' usability testing of two complex clinical decision support tools," *Int J Med Inform*, vol. 106, no. November 2016, pp. 1–8, 2017, doi: 10.1016/j.ijmedinf.2017.06.003.

[12] S. Richardson *et al.*, "'Think aloud' and 'Near live' usability testing of two complex clinical decision support tools," *Int J Med Inform*, vol. 106, pp. 1–8, 2017, doi: 10.1016/j.ijmedinf.2017.06.003.

[13] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," *Conference on Human Factors in Computing Systems - Proceedings*, no. April, pp. 249–256, 1990, doi: 10.1145/97243.97281.

[14] A. Vázquez-Ingelmo *et al.*, "Usability Study of CARTIER-IA: A Platform for Medical Data and Imaging Management," 2021, pp. 374–384. doi 10.1007/978-3-030-77889-7_26.

[15] J. Nielsen, "Usability inspection methods," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 1994-April, pp. 413–414, 1994, doi: 10.1145/259963.260531.

[16] S. Riihiaho, "The Pluralistic Usability Walk-Through Method," *Ergonomics in Design: The Quarterly of Human Factors Applications*, vol. 10, 2002, doi: 10.1177/106480460201000306.

[17] J. Brooke, "SUS : A quick and dirty usability scale SUS - A quick and dirty usability scale," no. November 1995, 2020.

[18] S. Palit, A. Datta, J. Lyu, and P. Chen, “D Ecision S Upport and S Ystems I Nteroperability,” *Landscape*, no. September, 2007.

[19] A. C. Li *et al.*, “Integrating usability testing and think-aloud protocol analysis with ‘near-live’ clinical simulations in evaluating clinical decision support,” *Int J Med Inform*, vol. 81, no. 11, pp. 761–772, 2012, doi: 10.1016/j.ijmedinf.2012.02.009.

[20] A. W. Kushniruk and V. L. Patel, “Cognitive and usability engineering methods for the evaluation of clinical information systems.,” *J Biomed Inform*, vol. 37, no. 1, pp. 56–76, Feb. 2004, doi: 10.1016/j.jbi.2004.01.003.

[21] J. Daniels, S. Fels, A. Kushniruk, J. Lim, and J. M. Ansermino, “A framework for evaluating usability of clinical monitoring technology.,” *J Clin Monit Comput*, vol. 21, no. 5, pp. 323–330, Oct. 2007, doi: 10.1007/s10877-007-9091-y.

[22] J. J. Saleem, E. S. Patterson, L. Militello, M. L. Render, G. Orshansky, and S. M. Asch, “Exploring barriers and facilitators to the use of computerized clinical reminders.,” *J Am Med Inform Assoc*, vol. 12, no. 4, pp. 438–447, 2005, doi: 10.1197/jamia.M1777.

[23] E. S. Berner and T. J. La Lande, “Overview of Clinical Decision Support Systems. In: Berner E.S. (eds) Clinical Decision Support Systems. Health Informatics,” *Springer*, vol. 3, pp. 1–18, 2007.

[24] A. Wulff *et al.*, “CADDIE2{\textemdash}evaluation of a clinical decision-support system for early detection of systemic inflammatory response syndrome in paediatric intensive care: study protocol for a diagnostic study,” *BMJ Open*, vol. 9, no. 6, 2019, doi: 10.1136/bmjopen-2019-028953.

[25] A. Wulff, B. Haarbrandt, E. Tute, M. Marschollek, P. Beerbaum, and T. Jack, “An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR,” *Artif Intell Med*, vol. 89, pp. 10–23, 2018, doi: https://doi.org/10.1016/j.artmed.2018.04.012.

[26] C. van Giersbergen, H. H. M. Korsten, A. J. R. De Bie Dekker, E. H. J. Mestrom, and R. A. Bouwman, “Quality Improvement in the Preoperative Evaluation: Accuracy of an Automated Clinical Decision Support System to Calculate CHA2DS2-VASc Scores,” *Medicina (Lithuania)*, vol. 58, no. 9, 2022, doi: 10.3390/medicina58091269.

[27] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.,” *Emerg (Tehran)*, vol. 3, no. 2, pp. 48–49, 2015.

[28] S. Safari, A. Baratloo, M. Elfil, and A. Negida, “Evidence Based Emergency Medicine Part 2: Positive and negative predictive values of diagnostic tests.,” *Emerg (Tehran)*, vol. 3, no. 3, pp. 87–8, 2015, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26495390%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4608333

[29] D. G. Altman and J. M. Bland, “Statistics Notes: Diagnostic tests 2: predictive values,” *BMJ*, vol. 309, no. 6947, p. 102, 1994, doi: 10.1136/bmj.309.6947.102.

[30] R. Trevethan, “Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice,” *Front Public Health*, vol. 5, no. November, pp. 1–7, 2017, doi: 10.3389/fpubh.2017.00307.

[31] F. G. de Oliveira Azevedo Matos and D. de Almeida Lopes Monteiro da Cruz, “Development of an instrument to evaluate diagnosis accuracy,” *Revista da Escola de Enfermagem*, vol. 43, no. SPECIALISSUE.1, pp. 1087–1095, 2009, doi: 10.1590/S0080-62342009000500013.

[32] H. H. C. Peres, R. Jensen, and T. Y. De Campos Martins, “Assessment of diagnostic accuracy in nursing: Paper versus decision support system,” *ACTA Paulista de Enfermagem*, vol. 29, no. 2, pp. 218–224, 2016, doi: 10.1590/1982-0194201600030.