Intelligent Decision Support Systems for Medicine: Inherent Performance Evaluation

A.E. Smith[1], C.D. Nugent[1], S.I. McClean[1]

[1]Medical Informatics, Faculty of Informatics, University of Ulster, Jordanstown, Newtownabbey, Co. Antrim, Northern Ireland, U.K. BT37 0QB. Email ae.smith@ulst.ac.uk

Abstract - **Researchers in the artificial intelligence community, who design decision support systems for medicine, are aware of the need for response to real clinical issues, in a problem driven approach, rather than just an academic exercise. They recognise that their systems need to meet the specific goals of the domain requirements and also to have been thoroughly evaluated, for acceptability. Attempts at compliance, however, are hampered by lack of guidelines. Evaluation can be thought of as being subjectivist and objectivist. Subjectivist evaluation appears to be addressed in the literature and also some objectivist evaluation, but the core evaluation of performance accuracy appears to be the area that receives least attention in evaluation papers. It is hoped to rectify this, by concentrating on the methodology of formal quantitative evaluation and disseminating the information, allowing progression towards the production of guidelines for a sufficiency of performance evaluation. Not carrying out this core evaluation avoids answering – "Does the system do what it claims?" and "is it more accurate than current methods?" Such questioning is essential for giving evidence that a real, scientific process has been applied to meet the safety – critical requirements of medical systems.**

Keywords: **intelligent medical systems; evaluation; performance**

I. INTRODUCTION

The term "decision support system" (DSS) is a generic term used to cover many types of intelligent systems which can be applied in the medical field [1]. Their use is not yet widespread in the domain, but it is growing rapidly as a means of handling a surfeit of information and knowledge. Their acceptability to clinicians demands that they be thoroughly evaluated, but therein lies a weakness as there is little in the way of laid down criteria.

The word "evaluation" is used loosely and inconsistently in assessing DSSs designed for clinical application. In its global form it can encompass both subjectivist and objectivist measures through all the processes of development and implementation. Subjectivist evaluation [2,3,4] involves mostly qualitative measures of organisational and human interface issues. These evaluations are of primary importance to the impact of any system on clinical usage, and most researchers and developers do address these issues. These qualitative measures, however, can lead to accusations of being a flawed and incomplete evaluation process if carried out in isolation. Objectivist evaluation centres on the use of quantitative measurement techniques to assess a system's effectiveness. Such an approach utilises all the identifiable stages of the development from needs assessment through to cost-effectiveness analysis [5] in order to try to identify the "truth" at each stage.

Evaluation can also be regarded as an umbrella term with elements of verification (system functioning accurately), validation (domain knowledge accurately represented), and assessment (end user and clinical impact) embedded in it [6] but these definitions are not generally agreed or utilised. Sometimes, evaluation is interpreted solely as the output performance of the system without reference to any other aspect of the system [7].

II. DEFINITION OF PERFORMANCE

Researchers vary in their interpretation of performance, for example, some regard it as how the system operates in the clinical setting, others may regard it as its user friendliness, or many other functions at different stages in its development, as well as overall decision accuracy. We intend performance here to mean all the measures which are carried out to examine how well the direct output from the system meets the "Gold Standard" (the correctly measured and agreed result, as recognised in the relevant domain), or if none can be identified, then against another methodology. These measures include accuracy, precision and assessment of errors. Performance measures as defined here can be seen (Fig. 1) as being at the core of objectivist evaluation, with all

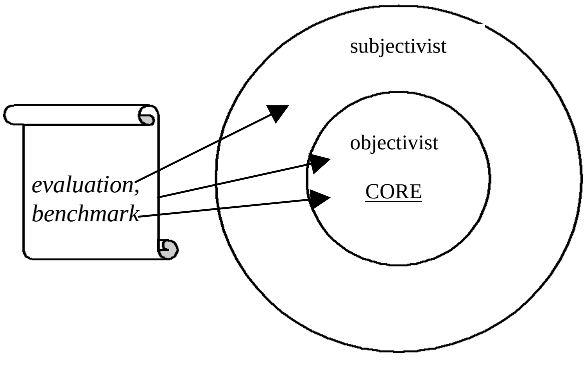

Fig. 1. Whole evaluation process for decision support systems.

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Intelligent Decision Support Systems for Medicine: Inherent Performance Evaluation	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Medical Informatics Faculty of Informatics University of Ulster Jordanstown, Newtownabbey, Co Antrim, Northern Ireland	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor’s Acronym(s)**
	Sponsor/Monitor’s Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 4	

the subjectivist approaches as a shell around this. These performance measures are essentially statistical in nature.

Objectivist evaluation can be thought of as having the following elements. (a) The ability to handle the characteristic features of medical data. (b) The correct representation of domain knowledge within the database. (c) Perspicuity of the processes involved. (d) Proven to be effective in the clinical environment. (e) The ability to deal with dynamic refinement of the knowledge based on feedback and as the environment changes. (f) Demonstrate generaliseability by being transferable to other, similar environments. All this is underpinned by the core or primary requirement of the inherent performance component (Fig. 2).

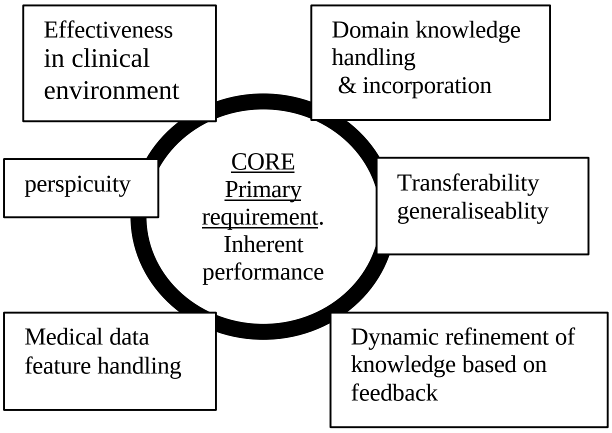
Effectiveness in clinical environment
Domain knowledge handling & incorporation
perspicuity
CORE Primary requirement. Inherent performance
Transferability generaliseablity
Medical data feature handling
Dynamic refinement of knowledge based on feedback

Fig. 2. Formal objectivist evaluation of decision support systems, with the core of inherent performance.

IV. WHY IS THERE INSUFFICIENT PERFORMANCE EVALUATION?

Many reports of objectivist approaches leave out this core measurement, that is, the system's inherent performance, or carry out inappropriate tests. A partial reason for this may be that researchers and developers in (a) academic institutions or (b) commercial organisations have different priorities from those in the target domain. Issues of publication and education dominate in the first case and marketing of their products dominate in the second case. This is often incompatible with the requirements of the end-users, who want reliable, user-friendly systems which are practicable in their operating environment. Another partial reason is that researchers see such critical evaluation as "hampering their creativity". Researchers and developers of DSSs cannot be expected to be expert statisticians and not all of them have access to this expertise. Statistical and classification texts tend to be written by academics who cover the theory very well but are presented in such formal and complex mathematical language that it is difficult for non-mathematicians to understand.

V. THE ADVANTAGES OF PERFORMANCE EVALUATION

For academics and developers, the main reason for carrying out formal objectivist evaluation, including performance measures, is that a level of confidence can be obtained as to whether their system (or model) A is better than system (or model) B. Also, numerical measurements can be considered to be less open to misinterpretation than verbal descriptions. Specifically, for the medical domain the advantages are that these evaluations can be shown as:

(1) Overcoming the lack of transparency in the processes, giving evidence that a real scientific approach has been applied, at least to the outputs.
(2) Increasing clinical acceptability, by cutting through allegations of empty hype.
(3) More likely to offset any product liability, or medico-legal claims.
(4) Being seen to have met the requirements of the CE kite mark [8,9]

This CE mark is a symbol for the EU Medical Devices Directive and means that the efficacy of a medical device now has to be demonstrated, documented and clinically evaluated, and the risk to benefit ratio assessed. This should include benchmarking by comparing the system with the nearest "substantially equivalent" approach. For many intelligent methods this means statistical methodology, for example, neural networks have been shown to be comparable with multiple logistic regression (MLR) techniques [10].

The reference for systems or the gold standard can be in the form of agreed measurements by experts, but comparison with only one expert is insufficient, and a Delphi approach is required. Any other validated and agreed objective reference, compatible with the study, can be used if appropriate. It is occasionally not possible to identify a gold standard so then direct comparisons with another approach are necessary, for example, sometimes laboratory tests give continuous output without an identified threshold indicating normal or abnormal. Self-testing methods for the system are not sufficient as this can give rise to unexpected results, for example, testing the same system in another location can lead to unexpected problems when a different population is examined. Also, different data gathering techniques could give dissimilar performance measures and indicate lack of transferability [11].

VI. THE POSSIBILITIES FOR REALISABLE PERFORMANCE EVALUATION

A few websites exist with the aim of giving evaluation guidance, such as the Statlog Project [12], where the aim was comparison studies of different machine learning, neural, or statistical classification algorithms. The comparisons are extensive and the evaluation of these include the development of a software tool, "Evaluation Assistant", which tests by applying N cross validation, leave-one-out, bootstrap and train & test methods. The tool is a self-testing approach, however, which only examines the one system rather than making valid comparisons, especially output measurements against other methodologies. Other websites are limited in that they do not cover a full range of possibilities or are not applicable to the medical domain [7].

A sufficiency of evaluation is needed and so some sort of structured framework is required for guidelines. Fig. 3 is the start of such an approach for the comparison of two models, or one model with a gold standard. The necessary comparisons with the gold standard, or other methodology, are almost exclusively statistical.

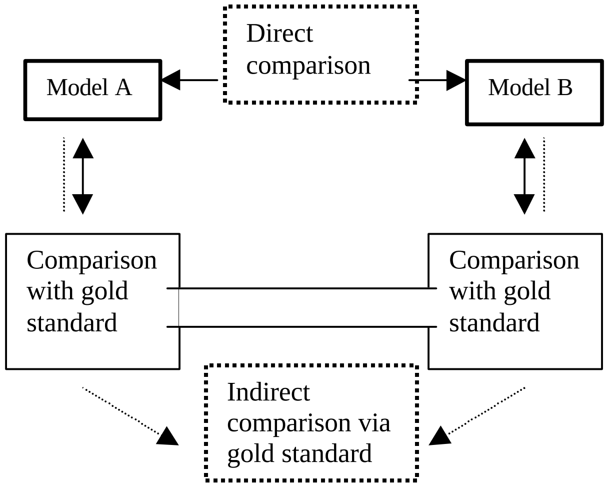
Direct comparison
Model A
Model B
Comparison with gold standard
Comparison with gold standard
Indirect comparison via gold standard

Fig. 3. A framework for evaluating the performance of a system.

An example of this would be the paper by Anand *et al.* [13] of a prognostic system for survival of colorectal cancer patients, where the estimates from a neural network (Model A) were compared with another methodology, Cox's regression (Model B), by pair-wise statistics (Wilcoxon's signed rank sum test). Their performances against a gold standard (a record of actual survivals of the patients) were measured in the form of errors and these were also compared by *t*-tests of the means, thus fulfilling the sufficiency of evaluation of performance, consistent with the goals and the data.

Most outputs of systems are in the form of either categorical data or continuous data. Although most outputs can be reduced to a percentage of overall classification accuracy, this causes loss of information. If the output is categorical, then the data are essentially non-parametric requiring tests such as odds ratios, relative risks, χ^2 etc. If binary in nature, then all Receiver Operating Characteristic (ROC) analyses are based on these, plus simple Bayes theorem (conditional probability) approaches. Unpaired and paired data can be handled in this way with appropriate tests applied. If the output is continuous, then the data can be paired or unpaired, requiring tests of a parametric or non-parametric nature, such as paired or unpaired *t*-test, Wilcoxon's or Mann Whitney test, etc.

Our intention is to give details of the specific tests appropriate to the type of data output, along with examples and constraints. These measures essentially treat the system processes as a "black box" and this is considered necessary, as the processes need to be hidden at this particular stage of evaluation [14].

Randomised Controlled Trials (RCTs) are used routinely in medical research, where they perform well in the original concept of direct comparison of treatment regimes, but are now being regarded as generally unsuitable for the evaluation of intelligent systems, by those in the intelligent software field [15]. They take little account of the barriers to the introduction of new technologies and are limited in their range of coverage.

VIII. DISCUSSION

Many evaluation papers for systems designed for medical application have been published, e.g. [16-20]. Studies of this nature have mostly concentrated on organisational issues and are valid, for the researcher and developer, for eventual application. Of these, however, very few give specific information on the exact nature of the system's inherent performance evaluation. This seems to be largely left to those who are interested in mathematical classification issues and are published in a form not easily understood by the non-experts in the field.

There is a general recognition that new technologies should be medical problem driven with output that is both appropriate and understandable to the end-user. Clinicians are sceptical of "black boxes" and require systems to be understandable as in rule-based approaches, or thoroughly evaluated if not transparent, before they will accept them. They are also wary of new inroads into their domain and so systems should be compared with statistical and other current acceptable approaches utilised in the domain

We would recognise that the imposition of too judgmental a form of assessment criteria, while the system or model is being finalised, would mean that many systems and ideas would not reach all the intended recipients who might benefit from it, or develop it further. If it is intended that the system or model be applied in the domain, however, then adequate formal evaluation is necessary.

Realistic and achievable guidelines for benchmarking are lacking in the field of medical decision support. It is hoped that we can contribute to the debate that will proceed towards rectifying this situation. We intend to give specific advice of a practicable, easy look-up nature to allow achievement of adequate performance evaluation by the DSS designers themselves. This may enable the plethora of papers that have been written describing systems, to be turned into reality and applied in the medical domain, rather than just remaining on the shelf.

ACKNOWLEDGMENT

This research is part of a Fellowship, funded by the Medical Research Council, London, UK.

REFERENCES

[1] http://www.mieur.nl/mihandbook/r_3_3/handbook/home.htm

[2] B. Kaplan, “Addressing organisational issues into the evaluation of medical systems,” *JAMIA*, vol. 4, pp. 94–110, 1997.

[3] J. Dowie, “The evaluation of decision aids: the role of the decision owner,” *Med. Inform*., vol. 15, 219-228, 1990.

[4] H.A. Heathfield and J.C. Wyatt, “Philosophies for the design and development of clinical decision-support systems,” *Meth Inform Med.*, vol. 32, pp. 1-8, 1993.

[5] J.C. Wyatt, “Evaluation of information systems,” in “*Handbook of medical informatics*,” J.H. van Bemmel and M.A. Musen, Eds. Heidelberg: Springer-Verlag, 1997, pp. 463-469.

[6] R. Engelrecht, A. Rector and W. Moser, “Verification and validation,” in: ”*Assessment and evaluation of information technologies*,” E.M.S.J. van Gennip and J.L. Talmo, Eds. Amsterdam: IOS Press, 1995, pp. 51-66.

[7] http://www.eksil-www.cs.umass

[8] “Medical device directive 93/42/EC,” 1993, *Official Journal of the European Communities*, L139, 1.

[9] “In vitro diagnostic medical devices directive 98/79/EC, 1998,” *Official Journal of the European Communities*, L331, 1.

[10] J. Lette, B.W. Colletti, M. Cerino, D. McNamara, M.-C. Eybalin, A. Levasseur, S. Nattel. "Artificial intelligence versus logistic regression statistical modelling to predict cardiac complications after noncardiac surgery," *Clin Cardiol.*, vol 17, pp. 609-614, 1994.

[11] J.M. Garibaldi, J.A. Westgate, and E.C. Ifeachor, "The evaluation of an expert system for the analysis of umbilical cord blood," *Art Int Med.*, vol. 17, pp. 109-130, 1999.

[12] http://www.ncc.up.pt/liacc/ml/statlog/

[13] S.S. Anand, A.E. Smith, P.W. Hamilton, J.S. Anand, J.S. Hughes, and P. Bartel, "An evaluation of intelligent prognostic systems for colorectal cancer," *Art Int Med.*, vol. 15, pp. 105-119, 1999.

[14] A. Rossi-Mori, D.M. Pisanelli, and F. Ricci, "Evaluation stages and design steps for knowledge-based systems in medicine," *Med Inform.*, vol. 15(3), pp. 191- 204, 1990.

[15] H. Heathfield, D. Pitty, and R. Hanka, "Evaluating information technology in healthcare: barriers and challenges," *B.M.J.*, vol. 316 (7149**)**, pp. 1959-1961. 1998.

[16] J. Hornberger, and M.K. Goldstein, "Clinical decision support systems: Evaluating the Evaluation," *Med Decis Making*, vol. 20, pp. 130-131, 2000.

[17] P.L. Miller, "Issues in the evaluation of artificial intelligence systems in medicine," *IEEE*, vol, 195, pp. 281-286, 1985.

[18] P.L. Miller, and D.F. Sittig, "The evaluation of clinical decision support systems: what is necessary versus what is interesting," *Med Inform*, vol. 15(3), pp. 185-190, 1990.

[19] C. Nohr, "The evaluation of expert diagnostic systems. How to assess outcomes and quality parameters," *Art Int Med* vol. 6, pp. 123-135, 1994.

[20] R. O’Moore, and R. Englebrecht, "The evaluation of medical decision support and expert systems: reflections on the literature," in "*Lecture notes in medical informatics*", New York: Springer-Verlag, 1991, pp. 263-73.