Challenges in Evaluating Decision Support Systems with Complex Outputs: Lessons from Design-a-Trial

Henry W. W. Potts*, Jeremy C. Wyatt*, Douglas G. Altman#

* School of Public Policy, University College London,
29/30 Tavistock Square, London WC1H 9QU, UK
{h.potts, jeremy.wyatt}@ucl.ac.uk

Centre for Statistics in Medicine, Institute of Health Sciences,
Old Road, Oxford OX3 7LF, UK
d.altman@icrf.icnet.uk

Abstract. Decision support system developers and users agree on the need for rigorous evaluations of system performance and impact. Fortunately, evaluating simple reminder systems is relatively easy, because there is usually a reasonable gold standard of decision quality and it is feasible to blind such studies. However, when the system generates complex or distinctive output (such as a critique, a paragraph interpreting lab tests, or a graphical report), it is much less obvious with which standard to compare the output, what support to offer control users in a comparative study and how to blind such studies. This paper discusses these issues and how one might resolve them, using as a case study Design-a-Trial, a DSS which helps doctors write a clinical trial protocol over 1000 words long.

1. Introduction

A clinical decision support system (DSS) is a computer tool which uses two or more items of data to generate patient- or encounter-specific advice (Wyatt & Spiegelhalter, 1991; Wyatt, 2000). There are a wide variety of DSSs, ranging from simple reminder/alert systems to those producing distinctive and complex output, such as the lengthy draft protocol document for a randomised clinical trial (RCT) generated by our own Design-a-Trial tool (http://www.design-a-trial.net).

At the first AIME conference, the need to evaluate even the accuracy of DSSs was not well established (Wyatt, 1987). This need is now recognised (Wyatt, 1997*a*) and there are even books discussing the evaluation of DSSs (Friedman & Wyatt, 1997; Brender, 1997). However, despite exhortations to carry out field trials (Tierney, Overhage & McDonald, 1994), evaluation of impact is still uncommon, although a systematic review by Hunt, Haynes, Hanna & Smith (1998) lists 65 RCTs of impact, showing these are feasible. However, most of the systems reviewed by Hunt *et al.* produced very simple output, such as one-word diagnoses or a phrase describing an alert event.

Inspired by the challenge of evaluating our Design-a-Trial system, we have identified a number of difficult and generic problems facing anyone trying to evaluate complex DSSs, particularly those with distinctive output. The next section describes Design-a-Trial and its output in more detail, but there are many other examples of clinical DSSs with complex or distinctive output including systems generating:

- prose reports of imaging studies (Taylor, Fox & Todd-Pokropek, 1999);
- reports detailing the interpretation of laboratory tests (*e.g.* Edwards, Compton, Malor, Srinivasan & Lazarus, 1993; Diamond, Mishka, Seal & Nguyen, 1994);
- critiques of a clinical management plan (*e.g.* Miller, 1986; van der Lei, Musen, van der Does, Man in't Veld & van Bemmel, 1991);
- graphical output such as a practice guideline (Fox, Johns, Rahmanzadeh & Thompson, 1997) or the design of a partial denture (Hammond, Davenport & Fitzpatrick, 1993);
- paragraphs summarising the clinical literature to aid patient care (Rennels, Shortliffe, Stockdale & Miller, 1989).

2. Design-a-Trial

Design-a-Trial (DaT) is a complex DSS to aid clinicians less experienced in RCTs to write a high-quality RCT protocol. Version 2 (Modgil, Hammond, Wyatt & Potts, *in press*), currently in development, uses a knowledge-based interface consisting of multiple adaptive data entry forms to guide the user through the process of designing the RCT. Using natural language generation routines, the software produces feedback in the form of critiques commenting on the statistical rigour and feasibility of the design, and assembles the data entered into a protocol of over 1000 words. DaT 2 uses a graphical user interface written in Visual Basic communicating with an inference engine written in Amzi! Prolog and a large knowledge base of relevant facts. The software runs in Windows 95 and is a development from DaT 1 (Wyatt, Altman, Heathfield & Pantin, 1994) which ran under OS/2.

Several difficulties have arisen in our plans to evaluate DaT which may generalise to the evaluation of other complex DSSs. In particular, we encountered special problems with the distinctive output of DaT, namely the trial protocol it generates. This raises issues about what to measure (including how we may determine the gold standard for a high quality answer) and how to ensure blinded assessment against a control. These issues are likely to be met by anyone evaluating a DSS with a distinctive output, so are expanded on in the subsections below. Other evaluation problems we encountered concern the choice of a control protocol. These are again generic problems for complex DSSs and are expanded on in section 3.2.

3. Problems in Evaluating Complex Decision Support Systems

3.1. Issues Arising in All Studies

Table 1 outlines a number of challenges arising in all evaluation studies. We will discuss these problems within the context of our plans to evaluate DaT. One could evaluate DaT in many different ways. An ergonomic analysis will be useful to see whether the software is quick and easy to use. One could assess user satisfaction, but the target audience, those inexperienced with RCTs, may be the least suited to judging whether DaT and its advice is of use. Users may have poor introspection: for example, they may not recognise good advice and may dislike being corrected by the computer system. A user assessment of the DSS is an important part of the evaluation of the software, but so is a more objective assessment of the DSS's output.

Table 1. Issues arising in all studies

Issue	Options	Implications
What measures to make	Quality of decision/output	See below
	Time or effort taken to complete a decision/case	Measures resources used rather than quality
	User satisfaction	Users may have poor introspection
What is a high quality solution	Reference to/use of standard guidelines; consistency of solutions	Increasingly important measure (*e.g.* for risk management); standards may not exist
	Exact match with DSS knowledge base	Circularity; lack of generality; tests animation of rules, not their quality
	Match with a quality checklist chosen by DSS developers	Circularity; which checklist to choose; bias
	Get external expert(s) to choose a checklist	Which checklist to choose; duplication of effort
	Use expert judges (without a specified checklist)	Criteria unclear
	Indirect, external measure of quality	Logistic challenges; influenced by other factors

DaT produces complex and distinctive output, namely a draft protocol of over 1000 words describing a planned RCT. That output is the result of guided input from the user and text generated by the software. It is our intent that this protocol should be of a high quality in two respects: both that the design of the underlying trial is statistically and

methodologically valid; and that the protocol is an apt, concise and precise description of the planned trial. A descriptive study of DaT could just assess protocols produced by DaT; a comparative trial of DaT would have a control condition in which users produce a protocol by other means. Either way, assessing the quality of protocols is clearly central to assessing DaT.

Assessing the quality of a protocol appears straightforward. One of the chief uses of a protocol is for submission to bodies who must make a decision about the trial described: whether to fund it, whether to allow it to go forwards on ethical grounds and so on. Bodies making these decisions routinely assess protocols in some manner. However, in considering how to formulate the assessment of protocols in a explicit manner, we became aware of some inherent problems.

In general, evaluators should check to see if there is already a scale for measuring the outcome of interest (Friedman & Wyatt, 1997). If a gold standard exists, then that is the clear way forward and criteria for gold standards exist (Wyatt, 1997*b*). However, in the absence of a gold standard the evaluator needs to prioritise the study questions. What issues, at that stage of evaluation, are most important?

A complex output normally calls for a composite measure of quality. While we are not aware of any formal scoring systems in the literature for assessing RCT protocols, there has been considerable work on assessing published reports of RCTs, offering an embarrassment of riches when it comes to different systems (Moher, Jadad, Nichol, Penman, Tugwell & Walsh, 1995; Jüni, Altman & Egger, 2001). The issues in assessing published reports of RCTs are very similar to those for RCT protocols. Both have dual notions of quality: the quality of the underlying trial and the quality of the description of that trial. It seems a simple task to adapt one of the many scoring systems for an RCT report to one for RCT protocols.

However, Jüni, Witschi, Bloch & Egger (1999) have shown that the use of different scoring systems for RCT reports leads to different conclusions when used in meta-analyses (see also Jüni *et al.*, 2001). The different systems put differing weights on different aspects of the report and design quality. For example, some concentrate more on issues of internal validity, some more on external validity. There is no inherent right answer to these relative weights; no obvious gold standard. This problem must generalise to a potential scoring system for RCT protocols. Moreover, we suggest that it will also generalise to most composite measures of quality in a manner similar to Arrow's Theorem on the theoretical limits of group decision making (Arrow, 1963; Bose, Heathfield & Andrew, 1992).

If the conclusions drawn using a composite measure of quality depend on the weighting of issues within that measure, then that raises a number of problems. How can one ensure that the choice or design of such a measure is not biased in favour of the DSS being evaluated? Even if the measure can be chosen in an objective fashion, there is still the potential for the DSS designers to concentrate their efforts on those issues that the chosen scoring system weights highest.

Such problems are perhaps soluble. Decisions about the assessment tool can be made separately from the design team. Expert judges could be set the task of choosing or designing the composite measure while blind to the details of the DSS, while the developers produce their software while blind to the details of the assessment exercise. However, this seems a rather perverse way of duplicating labour. The developers of the DSS and the judges are both concerned with quality: the former with producing it, the latter with assessing it. If the expert judges have useful insights into quality, sharing those insights would lead to a better DSS.

There is, we suggest, another generic problem here: circular definitions of quality. A measure of quality defines what quality means, but the DSS in its operation may also define what quality means. We could construct a checklist for assessing RCT protocol quality, but that checklist and the DaT rule base then exist as parallel definitions of quality. As both involve arbitrary decisions, as will be the case with most complex DSSs and composite scoring systems, does it makes sense to compare one (the DSS) against another (assessment tool)? The latter is no more a gold standard than the former.

We considered whether a valid alternative approach to using a composite measure of quality would be to ask expert judges to rate quality directly, without reference to an explicit scoring system. This, however, is only a superficial way of avoiding the problems raised. Judges will still use implicit quality criteria and, without a specified scoring system, we do not know what criteria they have used.

One possible way of untying this Gordian knot (Brender, 1997) is to use a pragmatic measure, one that is readily measured and an indirect result of quality. This could be seen as a test of predictive validity. In the context of RCT protocols, this might be the acceptance rate by some decision-making body, for example making funding decisions or an ethics committee. If dealing with a prescribing system, it may be unclear what constitutes good prescribing, but it is clear if patients show longer survival or higher quality of life. Although such external measures are attractive, they do pose logistic difficulties and again fail to make the quality criteria explicit.

Many of these problems arise because of the lack of a gold standard, which in turn is partly to do with the problems of composite quality scores. The best way forward may be to use multiple dimensions of quality: a ‘component approach’ (Jüni *et al.*, 2001). However, this will make it harder to draw clear conclusions from an evaluation exercise.

3.2. Issues specific to comparative studies

Comparative studies can be more powerful than simply assessing the DSS in isolation (Tierney *et al.*, 1994). Often, the most rigorous design is one in which users are randomised to the DSS or a control and we will concentrate on this here, although one might make other types of comparisons, *e.g.* to historical controls (provided sensible precautions are taken). Table 2 outlines challenges that arise in comparative studies.

Table 2. Issues specific to comparative studies

Issue	Options	Implications
What assistance to give to users in control group	No assistance	DSS as a black box
	Print-out of DSS rules	Tests value of software animation of the rules; awkward for control users; rules may not be readily available in such a format
	A textbook	Provides incentive for control participants
	A checklist	Controls for checklist effect
	Attenuated version of DSS, minus key functions	Choice of functions to omit; practicalities of developing alternate version
	Domain expert	May be unrealistic
How to blind assessors (for distinctive output)	No blinding	Assessor may be biased
	Use attenuated version of DSS as control	Limits what one is testing; may be impractical
	Abstract both control and DSS solutions into common third format	Abstractor may be biased; may lose added value of DSS (text assembly *etc.*)
	Abstract control solution into DSS	Abstractor may be biased; may be impractical

It is unclear what the control in any comparison of DaT should be. At present the inexperienced clinician will generally rely on what knowledge they do have, help from colleagues, advice from experts (often a limited resource), textbooks and other publications. There is some software of help at present, like power calculators or Methodologist's Toolchest (Idea Works, Inc.), but nothing with a comparable functionality to DaT. This nebulous mix of current support is difficult to formalise in a trial.

One can compare the DSS against nothing other than users' own experience. This is a pragmatic approach, but effectively assumes a worst-case scenario for the control, increasing the chance of finding a significant effect with the DSS. A fairer control would be to provide the user with some other tool, perhaps something of about the same price, like a textbook. However, any such evaluation is as much an evaluation of the particular control chosen as of the DSS in question. One could provide access to a human expert as a control, although that is perhaps an unfair control when resources in research and medical environments are rarely so generous. If this option was taken, then logging the length and type of communication required to achieve a result of similar quality to that with use of the DSS would be revealing.

An alternative is to compare the DSS against a hardcopy print out of the DSS's own rule base. This presumes such is obtainable; for example, neural net systems do not have explicit rules. This would test the computer animation of the rules, but not the value of the actual rules. However, a print out of the rules may be seen as an artificial way of

presenting rules, not something normally done and thus hampering control subjects, again leading to a potential bias in favour of the DSS.

Moving on to the need to blind comparative studies, while earlier DSSs may have given simple binary answers, DSSs are increasingly producing sophisticated output. Such output may be distinctive. With the user's input, DaT produces a lengthy draft RCT protocol in natural language (English). This appears as if it could have been written by a human working unaided, but the format and use of certain set phrases by DaT means that protocols written with DaT are distinctive if one sees more than a few. That distinctiveness, something shared by other DSSs, presents particular problems when it comes to blinding any comparative test. Blinding will be compromised if judges can identify whether they are dealing with output from the DSS group or the control group.

How can we preserve blinding? We could try to make the DSS output less distinctive or we could try to make the control output look more like the DSS output. These approaches, however, are liable to introduce bias into the evaluation. If we change what we are judging to make it look different, are we making a fair assessment? Those changes will probably have to be made by some person, who we call the abstractor, but how can we ensure they are unbiased? One approach may be to ensure both the control and DSS solutions conform to some third, pre-existing format.

It may be possible to alter the DSS under evaluation in some manner so that it has reduced functionality. This attenuated DSS can then be compared to the full DSS. If both versions of the software have a similar output, this avoids problems arising from a distinctive output. Of course, such a comparison can only assess those functions omitted from the attenuated version of the software and some software may not be adaptable in this fashion. Our original version of DaT, DaT 1, with separate data input, protocol generation and critiquing functions, would have allowed this approach, with critiquing being omitted from the attenuated version. However, DaT 2 uses adaptive input forms and pre-emption as well as critiquing in a more integrated fashion (Modgil *et al.*, *in press*) and cannot be readily attenuated. Some complex software may allow a related approach: a disaggregation of multiple functions, which could then be tested against each other.

While blinding is generally recommended, it is not essential. If one cannot avoid potential bias on the part of the judges in this manner, one may be able to minimise the problem in other ways (through a more objective outcome measure). If one cannot avoid the biases at all, one can still try to assess their magnitude: for example, judges' attitudes to the software can be recorded and compared with their ratings. Judges can also be tested to see whether they can guess which protocol came from which condition. In the context of DaT, it is reassuring that studies on peer reviewing papers, a similar task to rating RCT protocols, suggests that masking author identity has little effect on reviewers (Van Rooyen, Godlee, Evans, Smith & Black, 1998; Justice, Cho, Winker, Berlin, Rennie & the PEER Investigators, 1998).

4. Conclusions

It is important to evaluate the output and impact of DSSs. Doing so is often no more difficult than the tasks involved in developing a DSS. However, we believe that complex DSSs, particularly those with distinctive outputs, do present some special problems when it comes to evaluation. We have underestimated these challenges in our own project and hope this paper serves to forewarn other researchers.

Complex outputs may not have a gold standard. Assessment of such outputs may rely on composite scoring systems, but there are potential biases in designing or choosing such systems. Moreover, there is a potential circular argument if both the operation of the DSS and the scoring system are trying to define high quality. Distinctive outputs present problems with blinding, while complex functions make the choice of a control difficult. There are still other issues upon which we have not touched, like the choice of participants in a trial.

However, while we will have to accept some compromises, our own work on evaluating DaT is moving ahead. Applying some ingenuity and the insights discussed above, we hope it is not too hard to perform rigorous evaluations of the increasing number of complex DSSs. Moreover, we believe that many of these insights also apply to evaluations of other types of decision support, such as text checklists, or ready-formed but distinctive material, such as learning materials and Web sites.

Acknowledgements: Peter Hammond and Sanjay Modgil (Eastman Dental Institute); Charles Pantin; EPSRC; BUPA Foundation; and an anonymous reviewer.

References

Arrow, K. J. (1963). Social Choice and Individual Values, 2nd ed. Wiley, New York.

Bose, D. K., Heathfield, H. A., Andrew, M. (1992). Collective Decision Problems in Medicine – A Basic Approach Looking for Cross-Fertilization in Clinical Surgery. Theoretical Surg. **7**, 186-193.

Brender, J. (ed) (1997). Methodology for Assessment of Medical IT Based Systems. IOS Press, Amsterdam.

Diamond, L. W., Mishka, V. G., Seal, A. H., Nguyen, D. T. (1994). Multiparameter Interpretative Reporting in Diagnostic Laboratory Haematology. Int. J. Bio-Medical Comp. **37**, 211-224.

Edwards, G., Compton, P., Malor, R., Srinivasan, A., Lazarus, L. (1993). PEIRS: A Pathologist-Maintained Expert System for the Interpretation of Chemical Pathology Reports. Pathology **25**, 27-34.

Fox, J., Johns, N., Rahmanzadeh, A., Thompson, R.: ProForma (1997). A General Technology for Clinical Decision Support Systems. Comp. Meth. Prog. Biomed. **54**, 59-67.

Friedman, C., Wyatt, J. (1997). Evaluation Methods in Medical Informatics. Springer Verlag, New York (reprinted 1998).

Hammond, P., Davenport, J. C., Fitzpatrick, F. J. (1993). Logic-Based Integrity Constraints and the Design of Dental Prostheses. Artificial Int. in Med. **5**, 431-446.

Hunt, D. L., Haynes, R. B., Hanna, S. E., Smith, K. (1998). Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review. J. Amer. Med. Assoc. **280**, 1339-1346.

Jüni, P., Witschi, A., Bloch, R., Egger, M. (1999). The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. J. Amer. Med. Assoc. **282**, 1045-1060.

Jüni, P., Altman, D. G., Egger, M. (2001). Assessing the Quality of Randomised Controlled Trials. In: Egger, M., Davey Smith, G., Altman, D. G. (ed.s): Systematic Reviews in Health Care, 2nd edition. BMJ Books, London, chapter 5.

Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., Rennie, D. and the PEER Investigators (1998). Does Masking Author Identity Improve Peer Review Quality: A Randomized Controlled Trial. J. Amer. Med. Assoc. **280**, 240-242.

Miller, P. (1986). Expert Critiquing Systems. Springer Verlag, New York.

Modgil, S., Hammond, P., Wyatt, J.C., Potts, H. (*in press*). The Design-a-Trial Project: Developing a Knowledge-based Tool for Authoring Clinical Trial Protocols. In: 1st European Workshop on Computer-based Support for Clinical Guidelines and Protocols (EWGLP-2000). IOS Press, Amsterdam.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., Walsh, S. (1995). Assessing the Quality of Randomized Controlled Trials. Control. Clin. Trials **16**, 62-73.

Rennels, G., Shortliffe, E., Stockdale, F., Miller, P. (1989). A Computational Model of Reasoning from the Clinical Literature. AI Magazine **10**, 49-56.

Taylor, P., Fox, J., Todd-Pokropek, A. (1999). The Development and Evaluation of CADMIUM: A Prototype System to Assist in the Interpretation of Mammograms. Med. Image Analysis **3**, 321-337.

Tierney, W. M., Overhage, M. L., McDonald, C. J. (1994). A Plea for Controlled Trials in Medical Informatics. J. Amer. Med. Informatics Assoc. **1**, 353-355.

Van der Lei, J., Musen, M. A., van der Does, E., Man in't Veld, A. J., van Bemmel, J. (1991). Comparison of Computer-Aided and Human Review of GP's Management of Hypertension. Lancet **338**, 1504-1508.

Van Rooyen, S., Godlee, F., Evans, S., Smith, R., Black, N. (1998). Effect of Blinding and Unmasking on the Quality of Peer Review: A Randomized Trial. J. Amer. Med. Assoc. **280**, 234-237.

Wyatt, J. (1987). The Evaluation of Clinical Decision-support Systems: A Discussion of the Methodology Used in the ACORN Project. In: Fox, J., Fieschi, J., Engelbrecht, R. (eds.): Proc. 1st European Conference on Artificial Intelligence in Medicine, Marseilles. Spring Verlag, Heidelberg, 15-24.

Wyatt, J. (1997*a*). Quantitative Evaluation of Clinical Software, Exemplified by Decision Support Systems. Int. J. Med. Informatics **47**, 165-173.

Wyatt, J. (1997*b*). What are the Criteria for Selecting a 'Gold Standard' Measure? J. Health Serv. Res. Policy **2**, 60.

Wyatt, J. C. (2000). Knowledge for the Clinician 9. Decision Support Systems. J. Roy. Soc. Med. **93**, 629-633.

Wyatt, J., Spiegelhalter, D. (1991). Field Trials of Medical Decision-Aids: Potential Problems and Solutions. In: Clayton, P. (ed.): Proc. 15th Symposium on Computer Applications in Medical Care, Washington 1991. McGraw Hill Inc, New York, 3-7.

Wyatt, J., Altman, D., Heathfield, H., Pantin, C. (1994). Development of Design-a-Trial, a Knowledge-based Critiquing System for Authors of Clinical Trial Protocols. Comp. Meth. Prog. Biomed. **43**, 283-291.