Heather A. Heathfield

Computing Department, Manchester Metropolitan University, Manchester, UK

Synopsis

Decision Support Systems

During the last few years there has been a proliferation of evaluation studies performed on clinical decision support systems. This apparent commitment to evaluation is a welcome indication that medical informaticists are serious in their intent to produce scientifically rigorous work. However, while the results from the majority of these evaluation studies are valuable and have greatly improved our knowledge of clinical decision-support systems, one might question whether evaluation (in the form that is currently prevalent) has become an end in itself and is contributing progressively less new knowledge or insights on the topic.

This preoccupation with a particular artefact has been noted before and it is important that we do not lose sight of the primary purpose of evaluation: to further the goal of using decision support systems in routine clinical practice [1].

Evaluation seeks to answer a plethora of questions concerning the accuracy, usefulness, acceptability and impact of clinical decision support systems. Obviously as a first step towards demonstrating that decision support systems can be of benefit in healthcare, we need to convince clinicians of their diagnostic accuracy. Two papers included in this section concentrate specifically on evaluating diagnostic accuracy.

The paper by Berner et al. compares the diagnostic capabilities of four systems in the domain of internal medicine: Dxplain, Iliad, Meditel and QMR [2]. A set of 105 clinical case summaries involving actual patient data was created by a set of 10 experts. These cases were selected on the basis of being diagnostically challenging, giving equal coverage to the major organ systems, and having an appropriate gold standard for the diagnosis designated as correct in each case (i.e., a definitive diagnostic test or finding at autopsy or a consensus of experts when no definitive test could confirm the diagnosis). Each of the four systems was used to produce a ranked list of possible diagnoses for each patient, as did the group of experts. Scores were then calculated for each system based on several performance factors. The scores for the number of correct diagnoses given by the systems were not particularly encouraging, ranging from 0.52 to 0.71, and failing to include the correct diagnosis in 9 cases. More worrying is the fact that the knowledge bases of the four systems were incomplete, the proportion of the primary case diagnoses included ranging from 0.73 to 0.91, with three diagnoses not included in any of the knowledge bases. As the paper does not comment further on the nature of these three missing diagnoses, it is difficult to assess the implications of these omissions.

The range of relevant diagnoses was even lower, ranging from 0.19 to 0.37, and, as the authors note, this arouses concern that important diagnostic considerations may be so obscured by other diagnoses that the value of a system may be significantly decreased, or that it could lead to execessive or costly interventions in inexperienced hands. However, the systems did suggest some diagnoses, though not highly likely ones, that the experts later agreed were worthy of inclusion in the differential diagnosis.

The results of Berner et al.'s study would not encourage one to use any of the four systems as they exist at present, nor is it obvious how they would help one choose between them. The authors make the point that the developers of the four systems intend them to serve a prompting function, reminding clinicians of diagnoses they may not have considered, or triggering their thinking about related diagnostic possibilities. If these systems are intended to act in a cooperative rather than expert role, the underlying models they are based upon do not explicitly take account of users' problem solving abilities, and perhaps system designers should look towards current work in the area of cognitive psychology for more appropriate models [3]. Furthermore, evaluating systems that act as reminders is more complex then simply assessing diagnostic accuracy, and involves looking at the effects on clinican education and performance in addition to clinical outcomes.

Stamper et al.'s study describes a

system which provides diagnostic assistance based on the retrieval of similar previously diagnosed cases from a database [4]. The aim of the study is to compare statistical approaches to retrieval, in terms of their diagnostic accuracy and also their user interface and explanation capabilities. The "nearest neighbours" approach (using Hamming distance as a metric) was found to have a low accuracy. However, the authors claim that the approach was highly accountable, in that the users can inspect the subset of cases upon which the system diagnosis is based, and make their own judgement concerning the closeness of the match.

The 'independence Bayes' approach gave better accuracy, but was less accountable. In order to maximise accountability and accuracy, the authors propose a hybrid approach, which combines nearest neighbours and some other statistical technique to define a more accurate metric. In this study they evaluated this hybrid approach using independence Bayes to define the metric for nearest neighbours.

Whilst the hybrid approach was not found to be significantly more accurate than direct use of independence Bayes, the authors claim that the approach gives improved accountability. This study is encouraging in that it acknowledges the fact that clinicians are more likely to use a system when they can understand and examine the basis on which its advice is derived. However, it fails to provide any definition of accountability or any measurements to support its claim of improved accountability.

The paper by Dybowski et al. discusses three statistical approaches for use in a decision-support system for the management of septicaemia [5]. The authors' intention is to improve upon the approach taken by the MYCIN system, based upon their preference for objective probabilities provided by a database of septicaemia episodes, rather than the subjective measures of belief employed by MYCIN. Whilst this study is theoretically interesting, it fails to establish a genuine need for such a system at the given hospital site, and does not conduct any background study of the current situation in relation to the diagnosis and treatment of septicaemia. Whilst the authors do indicate their plans to evaluate the system once developed, the lack of baseline measurements will make it difficult to determine the extent to which any improvements in diagnostic accuracy can be attributed either directly to the system, or to other factors such as the Hawthorne effect [6].

Clinical decision support systems do not operate in a vacuum; rather they are situated within a complex environment where technical, social, political and organisational factors interact in unpredictable ways. Even if we are assured of the diagnostic accuracy of a system, this is not sufficient to conclude that a system will have a positive impact on users and healthcare. Studies which address these wider issues are much needed if decision-support research is to progress from theoretical work and prototypes towards clinically useful systems.

Johnston et al.'s paper is an interesting and comprehensive study that looks at the effects of computer-based clinical decision-support systems on clinician performance and patient outcome [7]. Of 793 citations (dated January 1983 to February 1992) of controlled trials that were examined, only 28 were found that met predefined criteria, reflecting a lack of scientific rigour in the majority of evaluation studies.

The decision-support systems considered were grouped into four application areas, and the number of systems showing an improvement in clinician performance given as follows: three of 4 applications of computer-assisted dosing, 1 of 5 applications of computer-aided diagnosis, 4 of 6 applications of computer-aided quality assurance for active medical care. Only three of 10 studies that assessed patient outcomes reported significant improvements. The results for computer-aided diagnosis uphold previous evidence that questioned the appropriatness of the preoccupation of the medical informatics community with developing diagnostic systems [1]. Unfortunately, this preoccupation with diagnostic systems continues in 3 of the 5 papers in this section.

The remaining paper in the decision-support section is concerned with extending computer science techniques to cope with the needs of medical informatics. In the paper "A Temporal Query System for Protocol-Directed Decision Support", Das and Musen describe a novel technique which supports temporal extensions to the Structured Query Language (SQL) for relational databases [8]. The work is based upon a need to store and query both instant-stamped data and interval-stamped data, at varying granularities in clinical databases. Evaluation of the method shows that it can express sufficiently all required temporal queries and that the search time of such queries is similar to that of standard SQL. This work has a wider applicability than simply satisfying the particular needs of protocol-decision support systems, and will be valuable in medical information systems that use commercial relational databases and SQL servers, and need to deal with temporal queries. It should also be of interest to the wider computer science community and one would hope that it will also be presented to this forum, as medical informatics and computer science often overlook the work of each other.

Much of the current evaluation work in clinical decision support is driven

by the idea that by demonstrating diagnostic accuracy, we will convince clinicians to routinely use systems.

However, this is blatantly not the case. Several notable systems have been repeatedly shown to improve diagnostic accuracy, yet do not get used in routine clinical practice. The problem of this lack of system utilisation is complex and requires further attention. Two important directions emerge. Firstly, there is a need to study the effects of social and organisational factors on system acceptance and impact, and one would hope to see socio-technical evaluations being undertaken [9]. Secondly, we might raise questions concerning the lack of professionalism in medical informatics and the manner in which this has affected clinicians' confidence in decision support systems [10].

References

1. Heathfield HA, Wyatt J. Philosophies for the design and devlopment of clinical decision-support systems. Meth Inf Med 1993;32: 1-8.
2. Berner ES, Webster GD, Shugerman AA et al. Performance of four computer-based diagnostic systems. N Eng J Med 1994; vol 330; No 25: 1792-96.
3. Coiera E. Question the assumption. In: Knowledge and decisions in health telematics. P Barahona and JP Christensen (eds). IOS Press, 1994: 61-6.
4. Stamper R, Todd BS, Macpherson P. Case-based explanation for medical diagnostic programs, with an example from gynaecology. Meth Inf Med 1994; 33: 205-13.
5. Dybowski R, Gransden WR, Phillips I. Towards a statistically oriented decision support system for the management of septicaemia. AI in Med 1993; 5: 489-502.
6. Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Med Inform 1990; 15: 205-17.
7. Johnston ME, Langton KB, Haynes B et al. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. Ann Intern Med 1194; 120: 135-42.
8. Das AK, Musen M. A temporal query system for protocol-directed decision support. Meth Inf Med 1994; 33: 358-70.
9. Forsythe DE, Buchanan BG. Broadening our approach to evaluating medical information systems. In P Clayton, ed., Proc 15th SCaMC 1993, McGraw-Hill 8-12.
10. Heathfield HA, Wyatt J. Towards professionalism in medical informatics. Draft paper to be submitted to Meth Inf Med.

Address of the author:
Heather H.A. Heathfield,
Computing Department,
Manchester Metropolitan University,
Manchester, UK.