J. Wyatt

Biomedical Informatics Unit, ICRF
London, UK

Synopsis

Decision Support Systems

This section of the Yearbook includes eight papers describing *Decision Support Systems*, a term implying a welcome lack of commitment to any specific technology. A further welcome revelation is that four papers appeared in clinical journals: medical informatics is definitely coming out of the closet at last. However, even if a paper appears in a clinical journal and describes a system which might support decisions, this does not mean that there is always a clinical need for decision support (Heathfield et al. [1]); more on this below.

Hovorka et al. [2] have written a useful paper with a clear five-page introduction to causal probabilistic networks (directed acyclic graphs). They describe an application to the prediction of blood glucose profiles in diabetics, and the selection of suitable therapy. The most novel feature of their work is their use of 24 time slices to give a prediction for the whole day. They also introduce an asymmetric weighting function to penalise low and high predicted blood glucose values; their system selects the optimal therapy on this basis. Those concerned about the notional threats to clinical freedom posed by clinical decision-support systems may be reassured to note that this system manipulates a 50Mb data set to look only 24 hours ahead, placing considerable demands on both hardware and software.

Another paper (Farr and Schuchter [3]) describes the principles behind, and the implementation of a tool for capturing an expert's utilities for possible outcome states for later use in advising on ventilator adjustment. The authors make a strong claim: that all decision support systems need an explicit representation of utilities - but they accept that conventional methods for assessing utilities, such as simulated gambles or lotteries, often give inconsistent results. Their tool may well be a solution for acquiring utilities, but their statement that it is "accurate and efficient", repeated twice in the paper, is sadly not substantiated by any data. Of course, it is hard to know how one would substantiate such a claim, since by definition the Gold Standard for utilities varies from individual to individual. They also give no indication of how they "covered the range of sample cases of interest", to ensure that the utilities elicited would be relevant to the cases in which the decision support system would be used. A similar problem arises in the selection of training and test cases for neural nets (see later): just which cases will people use these systems for, and how many of each kind of case are needed? Further research is needed on requirements analysis, assessing the impact of decision support systems on simulated decisions in psychological experiments, and on real decisions in field trials (Wyatt et al. [4]).

Freedman et al. [5] provide an excellent, lucid discussion of interim analysis in clinical trials and of conventional and Bayesian approaches to stopping rules. They show a welcome commitment to declaring, justifying and examining the implications of their assumptions, which is echoed in their use of sensitivity analysis when comparing the three major approaches. They even carry their existential doubts as far as to question the bedrock underlying their own and all competing approaches: the proportional hazards model. Their suggestion that physicians should always declare the circumstances under which a trial should be prematurely halted, and the language which they provide to formalise these, has similarities to the position adopted by Farr and Shachter [3]. They also remark that the current paucity of Bayesian trial analyses may result more from a lack of tools implementing the principles than from investigator 's disagreement. This is a general problem: when tools are available, we tend to use them, without questioning the principles on which they are based (Heathfield et al. [1]). What trialists are clearly waiting for is the launch of an irresistible Bayesian Analysis and Interpretation Toolkit (BAIT).

Chen et al. [6] describe a Chinese study of a numerical approach to psychiatric diagnoses, with a laboratory evaluation. They show care in defining the problem and the clinical data items for input, and assess the inter-rater reliability and repeatability of their instrument, though the long interval in the latter case (15-25 days), during which patients were on treatment, suggests that they might have underestimated the true repeatability. Unlike the neural network approach, their scoring system was built using the insights of experienced physicians, and its accuracy is further evidence that subjective scores or probabilities may be useful (Spiegelhalter et al. [7]). One criticism is that they compare their model to the decisions made by physicians currently treating the patients, implying that these are a Gold Standard. However, they do not explain why the decisions cannot continue to be made by the physicians, unaided by the decision support system. It would also perhaps be interesting to build a decision support system that predicts which therapy patients would respond to, derived from an analysis of cases that received various therapies. Evaluation would then be a straightforward question of checking whether the decision support system's predictions were vindicated after a specific period of patient follow-up.

There is evidence of another fruitful collaboration between physicians and scientists in a paper by Leaning et al. [8]. They describe a novel system, based on time-series analysis and chemometric methods, to predict resistance to chemotherapy in patients with trophoblastic tumours. However, in their evaluation the mean interval between the system predicting resistance and the physicians changing therapy ignores the 24% of cases in which the system identified drug resistance later than the physicians. They also quote the mean interval, which relies on the intervals being normally distributed, when the median often gives a better picture of the true effect. Finally, the clinical relevance of the time advantage is not entirely clear: will the patients attend every three days for blood tests, as they did in their example, to allow predictions to be made, and if resistance is successfully detected earlier, will the change in therapy be reflected in better survival or quality of life for the patient? From a clinical perspective, this is preliminary work, and we await the results of their field trial with interest.

Not surprisingly, with neural networks being the latest addition to the classificationist's armament, there were three papers describing such systems. Edenbrandt et al. [9] conducted a rigorous study to determine the potential of neural nets for classifying ECG ST-T segments. This paper deserves careful reading by all with an interest in neural nets: not only do the authors give a clear introduction to the subject, they go on to analyse which factors may be relevant to generating a successful network, including the number and granularity of input data items, the number of hidden nodes, variability in results following different training runs, the use of random subsets of the training data and of equal numbers of training cases in each outcome category, and the number of output nodes. Sadly, they did not re-examine the effects of early termination of training or different strategies for coding missing data (Hart et al. [10]). I was also a little concerned at the clinical plausibility of their results, as they failed to distinguish between ST-T segments from chest leads V1 to V4, and even treated four leads from 500 patients as if they were 2000 independent cases. Finally, their Gold Standard consisted of one cardiologist looking at only the four chest leads (not even the whole ECG), and it is unclear if relevant clinical data such as age and race were taken into account. Nevertheless, this is an excellent and thoughtful study.

Sadly, the same cannot be said for the two remaining papers about neural nets. There is something about neural nets currently - it was rule-based expert systems five years ago - that encourages fundamentalist behaviour by enthusiasts who fail to look beyond the technique itself. This emphasis on pursuing a technique irrespective of its relevance to solving a clinical problem concerned Heathfield et al. [1], and has sparked a debate in *Methods of Information in Medicine* about whether some of the work on medical decision-aids is not a case of the Emperor's New Clothes (Heathfield et al. [11]). For example, Akay [12] proposes that his system could be used for screening for ischaemic heart disease, but fails to appreciate the major effects of disease incidence on the utility of screening techniques (Sackett et al. [13]). Thus, he fails to compare like with like, taking performance figures for Thallium scans for a low-risk population and comparing them with those from his own system on a population who were at such high risk they all had coronary angiograms! There is also no definition of the Gold Standard; it is implausible that a neural net could be trained on only 12 cases unless this is a trivial classification problem, and Akay fails to exclude the obvious hypothesis that it is the clinical data, not the sounds of turbulent flow in the coronary arteries, that is wholly responsible for the discrimination between cases. Kippenham [14], meanwhile, built a neural net to help diagnose dementia from PET scans, failing to appreciate that dementia is by definition an untreatable condition, and that what is needed is a system to differentiate the treatable diseases that may be confused with it from the untreatable. Instead, he joins these two clinically important subgroups into his "possible dementia" category. He also uses the neural net's performance on the "test" set to decide when to stop training, thus making it into a training set, while suggesting that the physicians had an unfair advantage since they learnt while classifying the test cases. Fortunately, he does make two useful comments: firstly, a reminder that poor data collection will decrease classification accuracy, which is true for all decision-support systems, and secondly that classifying cases into disease present / absent is much easier than performing a differential diagnosis between several possible conditions.

In conclusion, interesting work is being done, using a variety of reasoning methods, but determining the clinical relevance of much of it must wait until the rigorous evaluation methods that are already applied to conventional medical technologies (Sackett et al. [13]) are widely used for clinical decision support. Such scientifically rigorous evaluation studies of decision-support systems have been performed by independent institutions, such as the excellent European CSE study [15], and deserve closer study by all who build clinical desision-support systems.

References

[1] Heathfield HA, Wyatt J. Philosophies for the design and development of clinical decision-support systems. Meth Inf Med 1993;32:1-8.

[2] Hovorka R, Andreassen S, Benn JJ, Olesen KG and Carson ER. Causal probabilistic network modeling - An illustration of its role in the management of chronic diseases. IBM Systems Journal 1992;31:635-48.

[3] Farr BR and Shachter RD. Representation of preferences in decision-support systems. Comput Biomed Res 1992;25:324-35.

[4] Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: Potential problems and solutions. In: Clayton P, ed. *Proc. 15th Symposium on Computer Applications in Medical Care*. New York: McGraw Hill 1991:3-7.

[5] Freedman LS and Spiegelhalter DJ. Application of Bayesian statistics to decision making during a clinical trial. Stat Med 1992;11:23-35.

[6] Chen HY, Luo HC and Phillips MR. Computerized psychiatric diagnoses based on euclidean distances: a Chinese example. Acta Psychiatr Scand 1992;85:11-4.

[7] Spiegelhalter D, Franklin R, Bull K. Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. In: Kanal L, Lemmer J, eds. *Proc 5th Workshop on Uncertainty in AI*. Amsterdam: North-Holland Publ Comp,1989.

[8] Leaning MS, Gallivan S, Newlands ES, Dent J, Brampton M, Smith DB and Bagshawe KD. Computer system for assisting with clinical interpretation of tumour marker data. BMJ 1992;305:804-7.

[9] Edenbrandt L, Devine B and Macfarlane PW. Neural networks for classification of ECG ST-T segments. J Electrocardiol 1992;25:167-73.

[10] Hart A, Wyatt J. Connectionist models in medicine: An investigation of their potential. Lecture Notes in Medical Informatics 1989;38:155-24.

[11] Heathfield HA, Wyatt J. Medical Informatics: Hiding our Light under a Bushel, or the emperor's New Clothes? Meth Inf Med 1993;32:181-2.

[12] Akay M. Noninvasive diagnosis of coronary artery disease using a neural network algorithm. Biol Cybern 1992;67:361-7.

[13] Sackett D, Haynes R, Guyatt G, Tugwell P. Clinical epidemiology: A basic science for clinical medicine (2nd ed). London: Little Brown & Co, 1991.

[14] Kippenhan JS, Barker WW, Pascal S, Nagel J and Duara R. Evaluation of a neural-network classifier for PET scans of normal and Alzheimer's disease subjects. J Nucl Med 1992;33:1459-67.

[15] Willems JL, Abreu-Lima C, Arnaud P, van Bemmel JH et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. New Engl J Med 1991;325:1767-73.