J.C. Wyatt

University College,
London, UK

Commentary

The Promises and Perils of Modelling Medical Reasoning

Reflections on E.H. Shortliffe and B.G. Buchanan's paper:
A model of Inexact Reasoning in Medicine

1. Introduction

Work by Shortliffe and Buchanan on MYCIN, key parts of which are described in the paper reprinted in this Yearbook [1], was the first detailed case study of a large medical rule-based system which handled uncertainty [2]. By publishing their insights and making the EMYCIN tool widely available to their R & D community [3], the MYCIN investigators promoted widespread experimentation with "expert" systems in academic and commercial settings in the early 1980's (e.g. [4-6]). This in turn led to important new insights into artificial intelligence (AI) in general and medical AI in particular [7], as well as a thriving commercial sector. Despite their drawbacks, such rule-based systems remain influential to the present day.

1.1 Some Influential Aspects of this Paper

The key insight described in this paper was combining qualitative knowledge represented as IF...THEN rules with quantitative knowledge, represented as certainty factors, to build a hybrid system, MYCIN. This system was the forerunner of later techniques, such as causal probabilistic networks [8], which have proved rigorous, efficient methods [9] for propagating uncertainty.

In their paper, the authors explored some of the similarities and differences between the statistical concept of probability and experts' use of intuitive notions of belief and certainty in clinical decision making. They then identified some of the ideological and practical problems posed by the "Idiot Bayes" approach. In response to these difficulties, the authors developed methods for capturing knowledge about beliefs and disbeliefs from domain experts using a 1-10 scale, and described the "paradox of belief". For example, an expert may elaborate a rule which, if true, carries a certainty in the conclusion of 0.7. However, even if all the rules conditions are met, this does not necessarily mean the expert's belief in the conclusion being false is 0.3; their disbelief may be more or less. To overcome this paradox, the team developed a calculus which they called the Certainty Factor (CF) mechanism. For a given rule, the Certainty Factor is equal to how much the evidence in the left hand side of the rule increases an expert's belief in the conclusion minus how much it increases their disbelief in the same conclusion.

The authors explored some of the properties and implications of this calculus and demonstrated rigorous mathematical proofs of some of its key features. Having thus specified their certainty calculus, they developed a range of robust tools to propagate uncertainty using CFs, and incorporated these into the generic EMYCIN expert system shell [3]. Finally, using this novel calculus, the investigators implemented the large-scale MYCIN advisory system. Although never used routinely in clinical practice and archived in the early 1980s, the MYCIN rule base was the substrate for several other significant AI and AI-in-medicine research projects over the next few years, such as work on the generation of explanations [10,11], generic prototypes [12] and task models.

Many of the insights originating from the MYCIN project remain valid today, though with 25 years of progress in computing hardware, software and medical informatics [13], it would be surprising if some had not been challenged.

2. Current Position and Validity of those Insights

2.1 Broad Comments

To a clinician, it is curious in retrospect that the investigators chose diagnosis for a multi-year research project. They state in their paper that the *"Potential clinical significance* (of diagnosis) *is apparent"*, but in fact many patients have either an established diag-

nosis of a chronic disease, or the diagnosis can be readily deduced from specific tests [14]. By studying the questions and dilemmas arising in routine consultations, several workers have concluded that physicians need assistance not with diagnosis but with choice of therapy, monitoring of disease progress and interpretation of test results [15,16]. To support this, Haynes discovered that, when physicians were provided with on-line access to MEDLINE to facilitate patient management, only 6% of queries were about diagnosis while 41% concerned choice of therapy [17]. It is, therefore, likely that diagnostic decision-support systems (DSS) will prove less effective than systems which target other kinds of decisions [18]. This prediction was confirmed by a recent authoritative systematic review of 68 randomised controlled trials of DSS [19]. While 33 of the 45 trials studying DSS, which advised on therapy or preventive care (73%) showed clear improvements in clinical practice, diagnosis was the least fruitful area for decision support, with only one of 5 trials (20%) showing an improvements [19].

Another assumption which now seems less robust is that "Rigorous probabilistic analysis (is) the ideal standard by which to judge the rationality of a physician's decisions". Since the 1980's [20] there has been increasing realisation that appropriate evidence from rigorous clinical studies should not only guide clinical actions but also be used as the basis for judging the rationality of physicians' decisions [21,22]. Rigorous probabilities are now seen to be a component of this evidence, but must be combined with many other factors, including the local availability of diagnostic tests (which reflects society's values), the risk or discomfort of these tests, and evidence that stems from patients and their medical records [23]. Such evidence includes informal patient preferences, formal utilities, and an individual patient's ranking of which clinical and other outcomes are most important to them [24].

In addition to choosing between antibiotic therapies for a possible septicaemia (MYCIN's domain), clinical decision makers will also consider doing nothing, referral (nowadays perhaps by telemedicine [25]), and a therapeutic trial among their "diagnostic" or, more broadly, management options. In choosing between the options, they may wish to take account of a "regret" factor [26] which reflects the loss of current opportunities and narrowing of future options, and even medico-legal exposure [27]. Thus, using a single dimension such as "certainty" to describe the complex reasoning which leads to the identification of an infecting organism or a therapeutic regime seems unnecessarily restricted. In reality, physicians navigate a complex network of management options and methods for choosing a rational path between them, which invite a richer variety of decision mechanisms.

One challenging issue which many decision-support system builders encounter but few formalise is the distinction between modelling the real world and modelling policy [28]. For example, many models attempt to describe actual clinical decision making or patient physiology, such as insulin–glucose metabolism [29]. Such models always seem restricted in their scope and accuracy and, while being of educational value [30], usually have limited clinical impact. The alternative is to use computers to model individual or shared policies, such as practice guidelines [31]. Because simplifying assumptions have already been made by the authors of guidelines, they represent what ought to be done (the "normative" approach), and can be more convincingly modelled. It seems as though - if the aim is to build models which are near to the truth - we should adopt this normative approach and avoid modelling reality. This is especially important now that many clinical DSS have been implemented [19] and criteria for judging the success of a decision model have moved from its faithfulness to physicians' think-aloud protocols to its utility for generating alarms or reminders which improve clinical practice [32].

To a rational clinician, the decision to treat a patient or collect more data depends crucially on the baseline risk of serious outcome, the probability that the patient will respond to the candidate treatment, the risk of side effects and the value of any extra information, for example a special investigation [14]. Specifically, this decision depends on whether any extra information will modify the clinician's estimate of the probability of therapeutic success sufficiently to cross the test / treat threshold [22,33]. Such reasoning underscores the need for the "rational clinical exam" [34] and rigorous evaluation of the performance characteristics of clinical findings and laboratory tests, expressed as likelihood ratios [14,35]. Thus, in contrast to our understanding 25 years ago [1], there is no need to capture hundreds of clinical findings because most are irrelevant, reflecting clinical tradition rather than their value to informed decision making [36].

Most medical informatics workers invited to collaborate in a project now shy away from the technology-led approach [13,37] and start by analysing the clinical problem and information needs [38]. Before building a decision-support system, a baseline audit of current clinical decision making [39] is needed to evaluate which errors are being made and their causes. Such causes may include:

- A lack of clinical or other knowledge.
- Poor quality patient data, e.g., de-

- Delays in obtaining data from records [40] or laboratory results.
- An inability to synthesise the two.
- Simple action slips [41]
- Lack of motivation.
- Barriers to physicians taking the correct actions, originating within a peer group or the organisation (e.g., lack of time or drugs) [42].

An example of a technology-led project and the failure of a computer DSS to improve decision making concerns the management of chest pain patients in an emergency room [43]. It emerged that the considerable delays and inaccuracies in patient management were largely due to a shortage of beds on the Cardiac Care Unit, rather than poor clinical decisions [18]; as a result, the DSS hardly improved matters [43].

2.2 Comments on Knowledge Representation and Uncertainty Propagation in MYCIN

MYCIN was undoubtedly a landmark system in terms of the technical and other insights it embodied. However, as mentioned earlier, it would be a serious mistake to build decision-support systems using the same techniques now, for major reasons.

2.2.1 Uncertainty Representation

First, considering the representation and propagation of uncertainty in advisory systems, the authors of the classic paper claimed that the Bayesian approach "becomes unworkable" in any realistic system. However, de Dombal's simple Bayesian Leeds abdominal pain system using 55 indicants halved the rates of serious errors and unnecessary surgery in 12 hospitals [44]. The authors expanded on this point by stating that the "Extent to which numbers can be manipulated as probabilities is unclear". It is certainly true that obtaining point estimates for probabilities from experts is hard, and it seems to be better to ask them to state a likely range of probabilities. One benefit of this approach is that it reveals an implied sample size, which in turn can be used to prime a system which combines subjective and objective probabilities [45]. However, the most significant development is that since 1988 we have been able to manipulate Bayesian probabilities accurately in a multiply connected graph or causal probabilistic network using the Lauritzen-Spiegelhalter algorithm, even in the presence of multiple diagnoses [8].

Of course, there are many other methods to model and represent uncertainty, including:

- Standard statistical approaches such as CART [46].
- Multiple logistic regression [47].
- Dempster-Schafer methods.
- Decision analysis.
- Cognitive modelling.
- Qualitative approaches, such as counting the arguments for and against a proposition [48].
- Non-monotonic logics, such as the deontic logic of obligation.
- Machine-learning methods, such as neural nets and genetic algorithms, which seem best suited for domains where we have no qualitative model, such as bioinformatics [49]

Some of these uncertainty methods have led to successful probabilistic systems, such as the spectacular discrimination and calibration of the Apache III system for predicting mortality in intensive care [50]. This is based on standard statistical modelling techniques applied carefully to large high-quality patient databases acquired in many hospitals using rigorous definitions.

2.2.2 Lack of Modularity of Rules

A second insight was the realisation that we cannot readily split off "discrete packets of knowledge" [1] as rules, "the myth of modularity" [51]. To summarise, in a typical large rule-based system, the role of each rule in the consultation process and generation of advice depends critically on which other rules are present. Also, and most difficult to predict, the way in which the certainty factors propagate from rule to rule depends on the CFs in other rules. The resulting difficulties of maintaining large rule bases [5] has led to increasing disillusionment with the simple method of representing knowledge as IF... THEN rules.

2.2.3 Kinds of Knowledge Represented in MYCIN

The third major realisation was that in MYCIN, the "judgmental knowledge" in rules actually compiles at least two different kinds of knowledge into the one-dimensional association: IF a AND b, THEN c. Clancey observed that rules in MYCIN were directed at two tasks: abstracting from clinical and laboratory findings to intermediate conclusions (e.g., the identity of an infecting organism), then classifying the conclusions using simple heuristics (e.g., a suitable antibiotic regime to cover these organisms) [11]. This is why the "explanations" generated by MYCIN's simple rule traces were inadequate for most purposes except to help the knowledge engineer debug the system. The need to represent such "task" knowledge explicitly in AI systems, so that it in turn can be reasoned about or used to generate explanations, has been realised [52].

A further difficulty with MYCIN's IF... THEN rules was that they failed to capture explicitly our deeper knowledge about the entities being reasoned about, such as:

- The relationships, or ontology, of the organisms, cultures, infections, clinical findings, laboratory results, etc. (e.g., streptococcus is a kind of pathogenic bacterium).
- The temporal relationships between

phenomena such as clinical findings, disease processes, and laboratory tests (e.g., infection precedes clinical symptoms by hours or days).
- Detailed anatomical knowledge about the body (e.g., the meningeal space does not usually communicate with the arterial circulation).
- Detailed causal knowledge (e.g., bacteria become penicillin resistant by evolving an enzyme which degrades penicillin).

Recognition of the value of explicitly representing such knowledge has led to AI systems with higher performance, especially at their margins, greater ability to explain their behaviour, easier maintenance and optimism about re-using the knowledge they contain.

2.2.4 System Control and Interfaces

The final insight about MYCIN and similar backward-chaining rule-based systems arose from the observation that they patronised their users, and were unable to make use of data available in other forms. In a seminal paper, the Demise of the Greek Oracle model for advisory systems was welcomed [53] and principles set out for a more sympathetic, opportunist model of advice. One central feature of this is the substitution of forward for backward chaining, while another is linkage of DSS to existing systems, particularly the electronic patient record. Of course, such linkages require that clinical data are coded using a controlled vocabulary, leaving other issues concerning clinical cost-benefit to be resolved [54].

Conclusions

The MYCIN project was very significant in the 1970s, which was influential over the following 25 years, both in academic AI and in the uptake of expert systems in industry and commerce. Now it may seem less relevant, thanks to a variety of developments ranging from the invention of the Lauritzen-Spiegelhalter algorithm to re-appraisal of the role of evidence and of patient preferences in clinical decision making. However, it is improbable that some of these more recent developments would have reached their current maturity without the ground-breaking work on medical AI in the 1970s and early 1980s which this paper exemplifies.

References

1. Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Math Biosci 1975;23:351-79.
2. Shortliffe EH. *Computer-Based Medical Consultations*: MYCIN. New York: Elsevier North-Holland, 1976.
3. Van Melle W. EMYCIN - a domain independent production system for consultation programs. In: *Proceedings 6th IJCAI, 1979*. Dept of Computer Science, Stanford University, 1979: 923-25.
4. McDermott J. R1: the formative years. AI Mag 1981;2:21-9.
5. Bachant J, McDermott J. R1 revisited: four years in the trenches. AI Mag 1984;5:21-32.
6. Hayes-Roth F, Waterman D, Lenat D, eds. Building Expert Systems. Wokingham: Addison Wesley, 1983.
7. Clancey W, Shortliffe EH. *Reading in Medical Artificial Intelligence*. Wokingham: Addison Wesley, 1984.
8. Lauritzen SL, Spiegelhalter DJ. Local computation on graphical structures and their application to expert systems (with discussion). J Roy Stat Soc B Met 1988;50:157-224.
9. Beinlich IA, Suermondt HJ, Chavez RM, Cooper GF. The ALARM monitoring system: a case study with two probabilistic inference techniques for belief networks. In: Hunter J, Cookson J, Wyatt J, eds. *Proc Second European Conf on Artificial Intelligence in Medicine*. Heidelberg: Springer Verlag, 1989: 247-58.
10. Clancey WJ, Letsinger R. NEOMYCIN: reconfiguring a rule-based expert system for application to teaching. In: *Proc Seventh International Joint Conf on Artificial Intelligence*. Los Altos: William Kaufman Inc, 1981: 829-36.
11. Clancey W. The epistemology of a rule-based system: a framework for explanation. Artif Intell 1983;20:215-51.
12. Aikins JS. Prototypical knowledge for expert systems. Artif Intell 1983;20:163-210.
13. Wyatt JC. Medical informatics: artefact or science? Meth Inform Med 1996;35:1[illegible]200.
14. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature III. How to use an article about a diagnostic test: B. What are the results and will they help me in caring for my patients? JAMA 1994;271:703-7.
15. Timpka T, Arborelius E. The GP's dilemma: a study of knowledge need and use during health care consultations. Meth Inform Med 1990;29:23-9.
16. Smith R. What clinical information do doctors need ? BMJ 1996;313:1062-8.
17. Haynes R, McKibbon K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in a clinical setting. Ann Int Med 1990;112:78-84.
18. Heathfield HA, Wyatt J. Philosophies for the design and development of clinical decision-support systems. Meth Inform Med 1993;32:1-8.
19. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998;280:1339-46.
20. Sackett DL. Rules of evidence and clinical recommendations for use of antithrombotic agents. Arch Intern Med 1986;146:464-5.
21. Sackett DL, Rosenberg WM, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't (editorial). BMJ 1996;312:71-2.
22. Sackett DL, Richardson WS, Rosenberg WM, Haynes RB. Evidence-Based Medicine. London: Churchill Livingstone, 1997.
23. Wyatt JC, Wright P. Medical Records 1: Design should help use of patient data. Lancet 1998;1375-8.
24. Eddy DM. Anatomy of a Decision. JAMA 1990;263:441-3.
25. Wallace S, Taylor P, Wyatt J. Telemedicine in the NHS for the millenium and beyond. Postgrad Med J 1998, in press.
26. Doubilet P, McNeil BJ, Weinstein MC. Optimal strategies for the diagnosis and treatment of coronary artery disease: analysis using a microcomputer. Med Decis Making 1983;3:23-8.
27. Dyer C. Litigation for medical accidents. BMJ 1988;296:1058-9.
28. Spiegelhalter D, Knill-Jones R. Statistical and knowledge-based approaches to clinical decision support systems, with an application in gastroenterology. J Roy Stat Soc A Sta 1984;147:35-77.
29. Hovorka R, Andreassen S, Benn JJ, Olesen KG, Carson ER. Causal probabilistic network modelling - an illustration of its role in the management of chronic diseases. IBM Syst J 1992;31:635-48.

30. Ingram D. Educational computing andmedicine. In: Dalton K, Chard T, eds. *Computers in Obstetrics and Gynaecology*. Amsterdam: Elsevier Science, 1990:313-28.
31. Ohno-Machado L, Gennari JH, Murphy SN, Jain NL, Tu SW, Oliver DE et al. The guideline interchange format: a model for representing guidelines. J Am Med Inform Assoc 1998;5:357-72.
32. Davis DA, Thomson MA, Oxman AD, Haynes RB. A systematic review of the effect of continuing medical education strategies. JAMA 1995;274:700-5.
33. Fagan TJ. Nomogram for Bayes' Theorem. New Eng J Med 1975;293:257.
34. Sackett DL. A primer on the precision and accuracy of the clinical examination. JAMA 1992;267:2638-44.
35. Friedman C, Wyatt J. *Evaluation Methods in Medical Informatics*. New York: Springer Verlag, 1997.
36. Wyatt JC. Clinical data systems, Part I: Data and medical records. Lancet 1994;344:1543-47.
37. Wyatt JC. Telemedicine trials: clinical pull or technology push ? BMJ 1996;313: 380-81.
38. Wyatt JC. Four barriers to realising the information revolution in health care. Chapter 5. In: Lenaghan J, ed, *Rethinking IT and Health*. London: Institute for Public Policy Research, 1998: 100-22.
39. Wyatt JC, Spiegelhalter D. Evaluating medical expert systems: what to test and how ? Med Inform 1990;15:205-17.
40. Nygren E, Wyatt JC, Wright P. Medical Records 2: Helping clinicians find information and avoid delays. Lancet 1998;352:1462-6.
41. Norman DA. *The Design of Everyday Things*. New York: Doubleday Inc, 1990.
42. Leape LL, Woods DD, Hatlie MJ, Kizer KW, Schroader SA, Lundberg GD. Promoting pateint safety by preventing medical error. JAMA 1998;280:1444-7.
43. Wyatt JC. Lessons learned from the field trial of ACORN, an expert system to advise on chest pain. In: Barber B, Cao D, Qin D, Wagner G, eds. *Proceedings of Sixth World Conference on Medical Informatics*. Amsterdam: North Holland, 1989:111-5.
44. Adams ID, Chan M, Clifford PC et al. Computer aided diagnosis of acute abdominal pain: a multicentre study. BMJ 1986;293:800-4.
45. Spiegelhalter D, Harris N, Bull K and Franklin R. *Empirical Evaluation of Prior Beliefs about Frequencies: Methodology and a Case Study in Congenital Heart Disease*. MRC Biostatistics Unit, Cambridge, England: BAIES Report BR-24, 1991.
46. Breiman L, Friedman JH, Olshen RA et al. *Classification and Regression Trees*. Belmont, Ca: Wadsworth International, 1984.
47. Spiegelhalter D, Knill-Jones R. Statistical and knowledge-based approaches to clinical decision support systems, with an application in gastroenterology. J Roy Stat Soc A 1984;147:35-77.
48. O'Neill M, Glowinski A. Evaluating and validating very large knowledge-based systems. Med Inform 1990;15:237-52.
49. Wyatt JC. Nervous about artificial neural networks ? (editorial). Lancet 1995; 346:1175-7.
50. Knaus WA, Wagner DP, Lynn J. Short term mortality predictions for critically ill hospitalised adults: science and ethics. Science 1991;254:389-94.
51. Heckerman D, Horwitz E. The myth of modularity in rule-based systems. In: Lemmer J, Kanal L, eds. *Uncertainty in AI 2*. Amsterdam: Elsevier Science, 1988:115-21,
52. O'Neill MF, Glowinski A, Fox J. A symbolic theory of decision-making applied to several medical tasks. In: Hunter J, Cookson J, Wyatt J, eds. *Proc AIME'89*. Heidelberg: Springer Verlag 1989;62-71.
53. Miller RA, Masarie FE. The demise of the Greek oracle model for medical diagnosis systems. Meth Inform Med 1990;29:1-2.
54. Wyatt JC, Keen JR. The NHS's new information strategy (editorial). BMJ 1998;317:900.

Address of the author:
J.C. Wyatt,
Knowledge Management Centre,
School of Public Policy,
University College London,
29 Tavistock Square,
London WC1E 7HN,
UK
E-mail: jeremy.wyatt@ucl.ac.uk