AIME 89

Second European Conference on Artificial Intelligence in Medicine

Lecture Notes in Medical Informatics

Edited by O. Rienhoff and D.A.B. Lindberg

38

J. Hunter J. Cookson J. Wyatt (Eds.)

AIME 89

Second European Conference on Artificial Intelligence in Medicine
London, August 29th–31st 1989
Proceedings

Springer-Verlag
Berlin Heidelberg New York London Paris Tokyo Hong Kong

Editors

Jim Hunter
Department of Computing Science
University of Aberdeen, King's College
Old Aberdeen AB9 2UB, United Kingdom

John Cookson
Hill Centre, The London Hospital Medical College, University of London
Turner Street, London E1 2AD, United Kingdom

Jeremy Wyatt
Department of Clinical Physiology
National Heart and Lung Institute, Brompton Hospital
Fulham Road, London SW3 6HP, United Kingdom

ISBN 3-540-51543-7 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-51543-7 Springer-Verlag New York Berlin Heidelberg

Printing and binding: Druckhaus Beltz, Hemsbach/Bergstr.
2127/3140-543210 – Printed on acid-free paper

Proceedings

AIME 89

Second European Conference on Artificial Intelligence in Medicine

London, August 29th - 31st 1989

Organised by:

AIME
European Society for Artificial Intelligence in Medicine

Organised in cooperation with:

BMIS
British Medical Informatics Society

Proceedings editors

Jim Hunter, John Cookson, Jeremy Wyatt

International Programme Committee

Chair: Jim Hunter, University of Aberdeen

I Bratko, Ljubliana
R Engelbrecht, Munich
M Fieschi, Marseilles
A Hasman, Maastricht
J-L Renaud-Sallis, Bordeaux
M Stefanelli, Pavia
T Wetter, Heidelberg
S Cerri, Milan
J Fox, London
T Groth, Uppsala
W Horn, Vienna
D Spiegelhalter, Cambridge
J Van Bemmel, Rotterdam

Local Organising Committee

Chair: John Cookson, London Hospital Medical College

E Carson
J Morgan
M Leaning
J Wyatt

Tutorial Chair: Jeremy Wyatt, Brompton Hospital, London

REFEREES OF PAPERS

K-P Adlassnig	University of Vienna
P Alvey	Royal Free Hospital School of Medicine, London
S K Andersen	University of Aalborg
S Andreassen	University of Aalborg
P Barahona	Universidade Nova de Lisboa
B Becker	GMD Schloss Birlinghoven, St. Augustin
F Beltrame	Universita' di Genova
I Bratko	E Kardelj University, Ljubliana
E Carson	The City University, London
D Cramp	Royal Free Hospital School of Medicine, London
P F de Vries Robbe	Academisch Ziekenhuis Groningen
R Engelbrecht	GSF-Medis Instit, Neuherberg
M Fieschi	Hopital de la Conception, Marseilles
J Fox	Imperial Cancer Research Fund, London
T Groth	Uppsala University
F Harvey	St. Thomas Hospital Medical School, London
A Hasman	University of Limburg
R Haux	University of Tubingen
W Horn	University of Vienna
M S Leaning	Clinical Operational Research Group, London
I Magnin	INSA, Villeurbanne
G Molino	Universita' di Torino
A Rector	University of Manchester
J-L Renaud-Salis	Fondation Bergonie, Bordeaux
A-L Rosenfalck	Nordjysk Udviklingscenter, Aalborg
A Rossi Mori	Inst Tecnologie Biomediche C N R, Rome
N Saranummi	Technical Research Centre of Finland, Tampere
D J Spiegelhalter	MRC Biostatistics Unit, Cambridge
M Stefanelli	Universita degli Studi di Pavia
J L Talmon	University of Limburg
P Torasso	Universita' di Torino
M Veloso	Hospital de Egas Moniz, Lisbon
J L Willems	Univers Hospital Gasthuisberg, Leuven
R P Worden	Logica, Cambridge
J Wyatt	Brompton Hospital, London

CONTENTS

CLINICAL APPLICATIONS (1)

METHODOLOGY (1)

REASONING BASED ON PHYSIOLOGICAL MODELS

CLINICAL APPLICATIONS (2)

ARCHITECTURES FOR MEDICAL KNOWLEDGE-BASED SYSTEMS (2)

UNCERTAINTY

GENERAL SESSION

METHODOLOGY (2)

Knowledge Elicitation and Acquisition

THE ROLE OF CLINICAL JUDGMENT ANALYSIS IN THE DEVELOPMENT OF MEDICAL EXPERT SYSTEMS.

D. Mark Chaput de Saintonge and M. J, Cookson
Departments of Clinical Pharmacology and Therapeutics, and The Hill Centre,
The London Hospital Medical College, Turner Street, London E1 2AD

ABSTRACT.

Clinical Judgment Analysis (CJA) has been used to generate statistically firm models of expert judgment. These studies show that experts have poor insight into the basis for their decisions. In addition expert performance has been shown to be inconsistent and often surprisingly poor compared with non-experts in the domain of expertise. CJA studies have shown that the frequently used methods of knowledge elicitation for IKBS based on the analysis of a single expert's behaviour and his interaction with a knowledge engineer may be seriously compromised by sources of error rarely considered by IKBS implementers. CJA reveals important differences between experts, the causes of these differences and ways of achieving consensus between experts. Possible applications of CJA techniques in the development of IKBS are discussed.

Introduction

The classical model of the development of an expert system application is that an expert is identified, and his 'expertise' is extracted by a process which has been compared to 'mining'. The majority of applications are developed by an individual - a knowledge engineer - interacting with the domain expert. Preliminary study - domain analysis - identifies the role of the prospective system and the tasks it has to perform (1,2). To reach this stage a model has to be produced of the domain being examined - an interpretation model. When such a model has been constructed this then guides the subsequent elicitation of expertise.

Expertise is commonly obtained from the study of the expert in action. Often verbal data is gathered from interviews which range from the informal to the highly structured or focussed. Introspection may be used, where the expert reports verbally on a typical case which may be synthetic rather than actual. The expert may review protocols derived from experimental study. A major problem is that the analysis of verbal data is difficult (3), there may be significant problems in interpreting these data consistently and they are always incomplete (4). An underlying assumption of the IKBS (Intelligent Knowledge-Based Systems) tradition is that validity of the experts judgment is taken for granted.

The knowledge acquisition process for an expert system is often protracted and typically takes months or years. Given the difficulties of extracting expertise from even a single expert, it is hardly surprising that comparatively little work has been performed using several experts systematically. Where there is more than one expert there are the additional problems of conflicting opinions and the problem of identifying who has the appropriate expertise in the domain. This area is one that most of the current generation of expert systems developers have ignored.

In fact the analytical study of groups of experts attempting the same task - research on Judgment and Decision Making (J/DM) - has an extensive literature dating back 60 years to Thorndike's study of the selection of army officer candidates (5). The isolation of the J/DM and the Artificial Intelligence (AI) research streams has recently been reviewed and their similarities and differences explored (6). Results reported in the J/DM literature raise fundamental questions concerning the very nature of the expertise that experts are claimed to possess. If medical expert systems are ever to grow from limited experimental systems to routine medical tools, the problem of identifying and acquiring 'real' expertise must be solved.

Models of Expert Judgment

With the odd exception (7) there is overwhelming evidence that experts from non-medical fields have poor self-insight when they are asked to describe their judgment policies. The policy models prove to be poor predictors of future decisions (8,9,10). Furthermore there is even some evidence that the accuracy of insight decreases with seniority and experience (11). Studies with experienced general practitioners (12), rheumatologists (13) and psychiatrists (14) have confirmed that doctors perform no better. The use of methods which rely on introspection to capture expert knowledge are thus open to criticism unless the models or expert systems constructed can be validated against observed performance. This stage is too time-consuming to be acceptable to most clinicians so it is usually omitted. These observations have serious implications for the methods of selection of the expert to provide the expertise. The intuition of the knowledge engineer would be to work with the most senior and experienced clinician, yet the evidence is that this policy may be inappropriate. Self-confidence appears to be characteristic of expert decision-makers (15) yet the most confident diagnosticians may also be the least accurate (16). Some studies have shown that the judgments of experienced clinicians may be no better than those of graduate students (17)

It is sometimes assumed that the policies of most experts in a given field will be similar, so it is sufficient to examine a single individual in depth. This assumption is not born out by the evidence in medicine where systematic conflicts of judgment appear to be common (12,18,19,20). Since most expert systems developments involve a single human expert they are vulnerable to error.

In medicine, the symptoms and signs which form the basis of diagnosis bear a partial and uncertain relationship to the underlying disease. Where large statistical data bases and independent diagnostic criteria exist, the information content and discriminant power of clinical data can be established statistically - the diagnosis of acute abdominal pain is a good example (21). Such situations are still uncommon, diagnostic knowledge must therefore be obtained by probing experts.

Psychologists have provided a model of cognition (or knowing) which has provided a useful tool for modelling expert judgments.

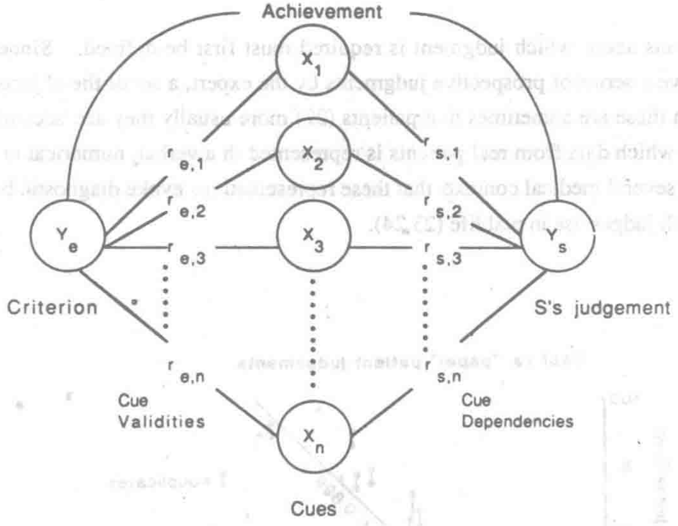

Figure 1. Brunswik's lens model applied to disease assessment. The criterion to be judged might be the activity of a disease. This is manifest as any number of cues or indicants which traditionally take the form of symptoms, signs, laboratory variables etc. The relationships between them and the criterion are usually indicated by the correlation cofficients ($r_{e,n}$). It is these cues that the doctor (S) takes into account when making a judgment about disease activity. His pattern of cue utilisation is apparant from the correlations they have with his judgments ($r_{s,n}$). This lens model paradigm allows systematic differences between doctors' judgmants to be dispalyed in terms of differences in importance attached to the various cues and differences in the combination rule used to arrive at a final judgment.

Expert judgment may be seen as drawing a conclusion about something the expert cannot see (the criterion) from something he can see (the cues). Cues which have a high degree of covariation with the event to be judged have a high degree of utility (or ecological validity). It would be most appropriate for the judge to weight the cues in proportion to their covariance with the criterion. When this does not happen and his cue utilisation weights differ significantly from the ecological validities, his overall achievement (ability to make a correct judgment about the criterion) is reduced. Inconsistent application even of an appropriate judgment policy will also reduce achievement. In situations where the criterion can be approached independently or there is some 'gold standard' the ecological validity of the cues can be assessed directly and used to construct the most efficient model, an approach used in the construction of data-based systems (DBS). When, as is frequent in medicine, the criterion is inaccessible only the utilisation weights are available. Clinical Judgment Analysis has been used to generate statistically firm models of expert judgment by eliciting the utilisation weights and combining them using the simplest rule.

Modelling procedure in Clinical Judgment Analysis (CJA)

The problem or objects about which judgment is required must first be defined. Since the modelling procedure will involve a series of prospective judgments by the expert, a set of the objects must be made available. Although these are sometimes live patients (22) more usually they are 'scenarios', 'vignettes', or 'paper patients' in which data from real patients is represented in a verbal, numerical or pictorial form. It has been shown in several medical contexts that these representations evoke diagnostic behaviour which is similar to that which judges use in real life (23,24).

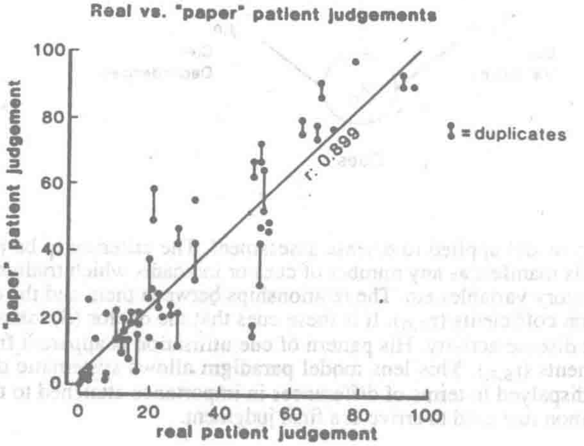

Figure 2. Correlation between clinical scores (0-100) for real and equivalent 'paper' patients when rheumatologists judge 'current disease activity' in patients with rheumatoid arthritis. (Reproduced with permission from 24)

Cases are usually chosen at random from the problem area. This ensures that the judgment task is representative and the resulting model will perform best under the commonest task conditions. If a model is required that is capable of handling 'unusual' or 'difficult' problems then their representation in the set to be judged should be proportionately increased. The judgment that is to be made about the objects is then defined and an appropriate scaling procedure agreed.

If the objects in the task set are patients, each could provide a vast number of cues, some important and some irrelevant, upon which a judgment of diagnosis could be based. Since human processing capacity is probably limited to about 7 variables presented simultaneously, the number of cues should be reduced accordingly. This may be achieved either by discussion with the judges or possibly by multidimensional scaling procedures. Criteria which cannot be represented by so few variables may need to be decomposed to simpler problems. The size of the task set which may be reasonably presented to the expert for his judgment is limited by human rather than statistical considerations. Among the general population of doctors in the UK few seem willing to consider more than 60-70 problems at a sitting. Each set should contain 15-20 cases presented in duplicate as a check on consistency.

When the expert has judged each of the cases his utilisation weights for each cue variable are identified by multiple regression analysis on the judgments using the cues as predictor variables. Provision may be made in the analysis for interactions and non-linear relationships between cues and the judgment. However in almost all studies of medical judgment such refinements have not significantly improved the fit of the equation. However finding a poorly-fitting model in a highly consistent judge would encourage a search for such refinements. Cue utilisation weights are usually expressed as their standardised regression cofficients or their percentage contribution to the overall multiple regression coefficient R^2.

We wanted to know how experts judged improvement or deterioration in patients with rheumatoid arthritis. A sample of rheumatologists were asked to judge the amount of improvement in 50 sets of data, each representing a single patient. Each data set gave 'before' and 'after' values for ten clinical variables (the cues) All possible subset regression was used, taking precautions to reduce the risk of overfitting. The cue utilisation weights differed considerably between the consultant rheumatologists taking the judgments (19). This was not unexpected since disagreement between doctors is a generally accepted fact of life and has been observed by others (25,22). The CJA procedure segregates the component of chance disagreement resulting from inconsistency from systematic disagreements which are the result of differing utilisation of cues. Expert policies modelled in this way have been shown to be stable over periods of 1 year (26) and to be sensitive to the effects of training (12).
An aggregate policy can be constructed by weighting each judge's policy by his consistency though there are obviously other approaches (27).

How do modelled policies compare with those generated by introspection?

Several studies have confirmed the observation that judges overestimate the importance of minor cues and underestimate their reliance on a few salient variables (10). Modelled policies are almost always considerably simpler than those generated by introspection.