PERGAMON

Expert Systems with Applications 16 (1999) 379–384

Expert Systems with Applications

Validating SPONGIA, an expert system for sponge identification

M. Domingo[a], M. Martín-Baranera[b], F. Sanz[b], C. Sierra[a,*], M.J. Uriz[c]

[a]*Artificial Intelligence Research Institute, IIIA, Spanish Council for Scientific Research, CSIC, 08193 Bellaterra, Barcelona, Spain*
[b]*Institut Municipal d'Investigació Mèdica, Universitat Autònoma de Barcelona, Dr Aiguader, 80. 08003 Barcelona, Spain*
[c]*Centre d'Estudis AvanÇats de Blanes, CEAB, Spanish Council for Scientific Research, CSIC, Camí de Santa Bàrbara, 17300 Blanes, Spain*

Abstract

In this article we present the validation of SPONGIA, an expert system to help in the identification of marine sponges. Validation was performed by using data from 82 randomly selected literature descriptions of sponge species. The data gathered by SPONGIA to identify each specimen were obtained from the bibliographical description. The set of cases, in which each case was described by the data gathered by SPONGIA, was presented to five internationally recognised experts in sponge systematics. The identifications generated by SPONGIA were compared with the identifications of these human experts by means of a cluster analysis. The similarity between SPONGIA and human identifications was assessed using four measures: Euclidean distance, City-block distance, Mahalanobis distance and Kappa index. In this article we show that SPONGIA obtains similar quality results to the experts in Porifera systematics.

Keywords: Marine sponges and systematics; SPONGIA; Cluster analysis

1. Introduction

Although formal evaluation is a fundamental step in expert systems development (Geissman and Schultz, 1988), the multiple approaches found in literature have not yet been systematically reviewed and standardised (Gennip and Talmond, 1995). As a part of evaluation, the validation of the expert system performance has to face several major problems (O'Keefe et al., 1987). Decisions about what to validate, what to validate against, what to validate with or when to validate find not ready-made guidelines to support them.

In our previous work (Hernández et al., 1994b; Verdaguer et al., 1992), we have specifically addressed the issue of expert systems validation when no gold standard is available, as it is the case in a wide variety of settings (Wyatt and Speigelhalter, 1990). A gold standard is a reliable experiment or test that gives the correct solution to each problem, unequivocally and objectively. As far as an Expert System (ES) deals with complex domains, it is difficult, or impossible in many cases, to be completely certain of what the right solution is. Even a solution given by an expert is usually graded with uncertainty. The majority of validation studies solve such a situation by considering, in rather different ways, the opinion of several experts (Fieschi, 1990; Martín-Baranera et al., 1996; Redier et al., 1995).

We have addressed the validation problem in the context of SPONGIA, an ES dealing with marine sponges (Porifera) from the Atlanto-Mediterranean bio-geographical province (Domingo, 1994, 1995; Domingo and Uriz, 1998). SPONGIA can identify the orders and families of the class Demospongiae. Specimens of the classes Hexactinellida and Calcarea are identified, but the system is not programmed to continue their identification in lower ranks. Specimens of the family Geodiidae (Demospongiae: Astrophorida) can be identified to the species level.

SPONGIA, as most expert systems, is based on a question–answer protocol, that is, the system, when executing, puts questions to the user who gives appropriate answers to them. The outcome of SPONGIA is the identification of every taxonomic rank from class to species following down the porifera's taxonomic tree. SPONGIA is a type of expert system that permits users to reply questions with uncertain answers. The more uncertain the answers are the more imprecise the classification obtained by SPONGIA will be.

There are rather few examples of expert systems devoted to taxonomy (Conruyt et al., 1993; Wooley and Stone, 1987; Thonnat and Gandelin, 1988). Among them, virtually no report of validation has been found. Ideally, we could validate SPONGIA from real cases. That is to say, collecting a number of specimens randomly in the biogeographic area and processing the samples to gather the data that is required

* Corresponding author. Tel.: +34-935-809-570; fax: +34-935-809-661.
E-mail address: sierra@iiia.csic.es (C. Sierra)

0957-4174/99/$ - see front matter
PII: S0957-4174(99)00013-5

Table 1
Validation sample composition

Taxon name	Number of cases
Order: Homosclerophorida	2
Order: Astrophorida	33
Order: Spirophorida	2
Lithistida group (incertae sedis)	2
Order: Chondrosida	1
Family: Epipolasidae (incertae sedis)	1
Order: Hadromerida	8
Order: Axinellida	5
Order: Poecilosclerida	9
Order: Haplosclerida	7
Order: Petrosida	1
Order: Halichondrida	3
Order: Verongida	1
Order: Dictyoceratida	3
Order: Dendroceratida	3
Order: Merliida	1

by the ES. In practice, the amount of money necessary to collect the samples would be much higher than the whole ES development budget, and clearly unaffordable by the involved research groups. Thus, validating in retrospect seemed to be the right compromise. That is, taking the case problems from literature descriptions.

The species name that is found along with each description in the literature could have been taken as a gold standard. This was not considered a right decision in a domain where the name assigned to a description is not always agreed upon by all of the field experts and the species delimitation are not definite due to their high biological variability. Moreover, this would have forced us to choose only the well-described cases, thus biasing the validation sample. Moreover, by selecting well-described cases, the capability of SPONGIA for dealing with missing or uncertain data would not be evaluated.

Also, we could take as gold standard the identification of the validation sample given by an expert. This would not constrain the selection of cases, i.e. we could select discussed species or incomplete descriptions. However, this validation procedure would introduce the bias of that expert's opinion.

We have validated SPONGIA by comparing the identifications given by the ES over 82 specimen descriptions with those identifications given by several experts. This kind of methodology has two important advantages: it is possible to randomly choose the case problems among the literature descriptions and it provides a balanced assessment of the system. The only drawback is the time that experts have to spend in carrying out a non-creative task, that is, to identify the descriptions of species. In our case this was not a problem because five internationally recognised experts in sponge taxonomy were interested in the development of SPONGIA (but not involved in) and they undertook to do this task.

Table 2
Linguistic truth values and their numerical equivalence

Linguistic term	Numerical value
Impossible	0.0000
Hardly possible	0.0330
Slightly possible	0.1077
Moderately possible	0.2416
Possible	0.4500
Quite possible	0.6500
Very possible	0.8486
Sure	1.0000

2. Material and methods

2.1. Taxonomic tree

The Porifera taxonomic tree used in the validation of SPONGIA includes a total of 45 taxa: all the orders of the class Demospongiae, all the families of the order Astrophorida and the genus and species of the family Geodiidae (Demospongiae: Astrophorida).

2.2. Validation sample

Eighty-two case samples were taken from bibliographical sources according to the following procedure. First, descriptions of species were collected without taking into account the quality of the description. Then, they were divided in two groups: complete descriptions and incomplete descriptions. We considered a description to be incomplete if there was not sufficient information to answer all queries of SPONGIA. As the frequency of appearance of taxa in the bibliography is not uniform, we performed a stratified sampling among the available bibliographical descriptions to obtain a representative sample of the validation domain (see Table 1). To validate the level of order (16 orders) the number of cases per order was decided according to its frequency of occurrence in the biogeographic province and to the number of genera contained in each order, which reflects the order's diversity. About 30% of the cases were incomplete. To validate the identification of the family Geodiidae down to the level of species (5 genus and 13 species), three cases per species (two complete and one incomplete cases, if available) were chosen. In the remaining families of the order Astrophorida one case per taxon was taken.

2.3. Methodology

Five independent experts, internationally recognised, were asked to identify each case of the validation sample. Experts were required to identify the cases as precisely as possible. An identification consisted in the qualification of each possible taxon with one of the ordered linguistic truth labels used in SPONGIA (see Table 2). The information used

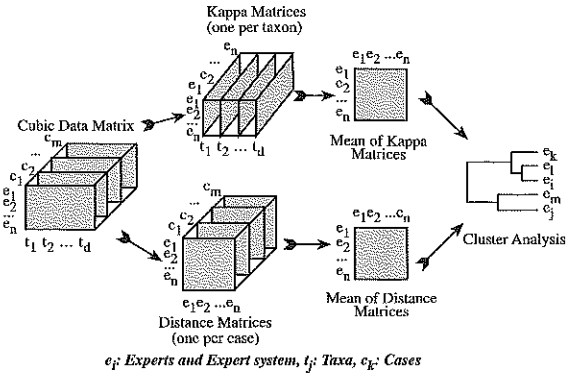

Fig. 1. Cluster analysis process for the Kappa similarity measure and the three distances.

by experts in order to make the identification was those data items gathered by SPONGIA during its identification process.

The expert's identifications and those proposed by SPONGIA were gathered in a data matrix and compared by carrying out a hierarchical cluster analysis on the basis of four dissimilarity (Euclidean distance, City-block distance, Mahalanobis distance) or similarity measures (Kappa index). See Fig. 1.

Experts were also asked to assess their colleagues' proficiency. Confidentially, they scored one another on a numerical scale (from 1 to 10. 1 meaning low, 10 meaning high). The scores were averaged and a ranking of the experts' proficiency was computed. The challenge was to obtain a balanced measure of proficiency to discuss the relative position of SPONGIA in the clustering. The average of the scores given by the human experts to each other resulted in the following ranking (here we introduce the human experts' naming convention to be used in the rest of the article, that is inspired by this ranking): expert *E*1 obtained a score of 9, both the experts *E2A* and *E2B* obtained scores of 8, *E*3 obtained a score of 6.5 and *E*4 obtained an average of 6.

2.3.1. Data matrix

Each case identification corresponded to an array of truth intervals. It was not mandatory for the experts to give a value for each possible taxon nor a limited number of taxon hypotheses for each case. Any taxon which is not mentioned in the expert identification is considered impossible, by default. For example, suppose that C1 is the description of case number 1, and *E*1, an expert in sponge systematics. *E*1 may consider that C1 corresponds definitely to the genus *Geodia* but it may be a specimen of *Geodia conchilega* or of *Geodia cydonium*. Such an identification would be expressed as: "*Geodia* is sure", "*Geodia cydonium* is possible" and "*Geodia conchilega* is possible".

Obviously, an expert's identification has to be interpreted in the context of systematic biology. We assume that organisms are classified in a hierarchical arrangement of taxa called taxonomic tree in which each taxonomic level represents an increasing degree of abstraction from species (leaves) to phylum (root). In a taxonomic level, taxa are exclusive classes, and each taxon must belong to a higher level taxon until the tree root. So, taxa corresponding to the higher taxonomic levels were inferred form the expert's identification, and were given the truth value of "sure".

For instance, in the latter identification for C1 we would infer: "class Demospongiae is sure, order Astrophorida is sure, family Geodiidae is sure". Then, the array that would represent this identification would have a linguistic truth value of "sure" in the components representing class Demospongiae, order Astrophorida, family Geodiidae and genus *Geodia*, a truth value of "possible" in the components representing *Geodia cydonium* and *Geodia conchilega* and a truth value of "impossible" (by default) in the rest of the array's components.

In the previous example, the identification arrived to the leaves of the taxonomic tree, that is, to the species level. It might happen that the identification stops at higher taxa, because there is not enough information. This situation means that there is not enough evidence about which particular lower taxon is preferred. In other words, it is "unknown" to which lower taxon the specimen belongs to. In our setting this ignorance is modelled by the whole interval [0,1] or, in linguistic terms, by the label "possible" which is the central one. Hence, to complete the array a value of "possible" will be assigned to all nodes in the subtree rooted at the taxon where the identification stopped, and "impossible" to the remaining nodes.

These identifications were gathered in a three-dimensional matrix indexed in the rest of the article by i, j, k, where index i runs over the five experts plus SPONGIA, $i \in \{E1, E2A, E2B, E3, E4, \text{SPONGIA}\}$, j runs over the taxa names (represented as integers) $j \in [1, 45]$ and k over the codified names of the case descriptions$k \in [1, 82]$. Each value x_{ijk}, is a real number representing the linguistic truth value given by expert i, to taxon j for case k; see for instance the Cubic Data Matrix of Fig. 1. (See Table 2 for the relation between linguistic terms and values in the interval [0,1].)

We split this complete data matrix into two submatrices. One, named *astro*, contained the identification of 33 descriptions of specimens within the order Astrophorida. The other subset, named *order*, contained 49 descriptions of specimens of all the orders of the class Demospongiae. The complete matrix and the two submatrices were the three validation scenarios over which we made a cluster analysis.

2.3.2. Cluster analysis

The hierarchical cluster analysis performed in this work is based on three distances: *Euclidean*, *City-Block*, *Mahalanobis*, and on a similarity measure called *Kappa index*. The cluster analyses used apply the weighted average linkage criterion (Vogt et al., 1987), also called UPGMA (*Unweighted Pair Group Method using arithmetic averages*). Assume that groups i and j form a new cluster t. The distance between the new formed group t and every other preexistent group r is calculated as the weighted

average of $d(i,r)$ and $d(j,r)$, using i and j group sizes (n_i and n_j) as weighs:

$$d(t,r) = \frac{n_i}{n_i + n_j} d(i,r) + \frac{n_j}{n_i + n_j} d(j,r).$$

In the cluster analysis from the matrix *astro* we aimed to compare the quality of the identification of the specimens of the order Astrophorida until the species level made by human experts and by SPONGIA. In the cluster analysis from the matrix *order*, we pursued the same goal related to the identification of any sponge to the level of order. Thus, we could assess the ability of SPONGIA in identifying the level of order, and the level of species independently.

In the remainder of this section we present the measures used.

2.3.2.1. Euclidean distance

$$d_k^{\mathrm{e}}(a,b) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (x_{ajk} - x_{bjk})^2},$$

where $d_k^{\mathrm{e}}(a,b)$is the Euclidean distance between two experts a and b for the validation case k, N is the total number of possible identifications (taxa), j is each one of the possible identifications, x_{ijk} is the numerical truth value assigned by expert i to taxon j for case k. This is a suitable measure to calculate distances between elements in an orthogonal space, that is to say, when variables are independent. In our case, variables are the possible identifications, namely the taxa.

2.3.2.2. City-block distance (Manhattan distance)

$$d_k^{\mathrm{c}}(a,b) = \frac{1}{N} \sum_{j=1}^{N} \left| x_{ajk} - x_{bjk} \right|.$$

This distance makes equivalent a high number of slight differences in several positions to a great difference in a single position; in contrast to the previous distance. Thus, it detects more clearly the disagreement between a couple of experts when the disagreement's origin is because the experts give slightly different possibilities to several taxa. The Euclidean distance would have given a relatively lower difference in this case.

2.3.2.3. Mahalanobis distance This generalised distance measure takes into consideration the correlations between identifications. In a taxonomic domain the possible identifications are correlated, e.g. the identification of a genus is positively correlated with the identification of the family, the order and the class to which this genus belongs, and negatively correlated with the remaining families, orders and classes. Also, if the identification arrives at the level of species, two species of the same genus are more likely to be labelled with a positive certainty value than two species of different genera. This correlation stems from the assumption that the answer is consistent with the characteristics of the taxonomic tree i.e. satisfying that taxa are exclusive and exhaustive classes. The calculation of Mahalanobis distances follows the expression:

$$d_k^{\mathrm{m}}(a,b) = \sqrt{\frac{1}{N} V^t * W^{-1} * V},$$

$$V = X_{a-k} - X_{b-k},$$

where $d_k^{\mathrm{m}}(a,b)$ is the Mahalanobis distance between two experts a and b with regard to a validation case, k, V is the N-dimensional column vector of the differences between those experts for each possible taxon in such a case, V^{t} is the transposed vector, and W^{-1} is the inverse of the variance–covariance matrix of the possible identifications that permits to take into account the correlation between the identifications.

2.3.2.4. Kappa index This similarity measure addresses the problem of the concordance between experts due to a random coincidence of their answers. A weighted Kappa, k_p, has been defined for every pair of experts, a and b, and every possible taxa j, taking into account the influence of random agreements among experts. $k_p(a,b,j) = 1$ if there is total agreement between experts a and b in all identifications of taxon j. $k_p(a,b,j) = 0$ if the agreement is only due to random coincidence, and $k_p(a,b,j) < 0$ if there is an agreement between a and b in the classification of taxon j lower than the random one. The definition of k_p is:

$$k_p(a,b,j) = 1 - \frac{\sum_{k=1}^{G} \sum_{l=1}^{G} O_{kl}(a,b,j) w_{kl}}{\sum_{k=1}^{G} \sum_{l=1}^{G} E_{kl}(a,b,j) w_{kl}},$$

where $O_{kl}(a,b,j)$ is the observed number of cases that expert a labels taxon j with linguistic label k and expert b labels the same taxon j with label l. $E_{kl}(a,b,j)$ is the expected number of observations due to random assigment of label k by expert a and label l by expert b, always with respect to taxon j, G the total number of linguistic labels, and w_{kl} the weigh assigned to cell kl, being $w_{kl} = |k - l|$. For further details refer to Fleiss, 1981.

The calculation of the measures and the cluster analyses were computed using the software POEMA (Hernández et al., 1994a). POEMA 1.0 is coded in Microsoft Visual Basic (version 2.0) for Windows. To compute these distances, the linguistic truth labels were replaced by their numerical equivalent (see Table 2). This conversion and a previous version of the software have already been used in the validation of medical expert systems (Hernández et al., 1994b; Verdaguer et al., 1992).

On the basis of each one of the measures, a 6 by 6 matrix of dissimilarity (distances) or similarity (Kappa) was obtained. The distance between two experts for a case identification results from the comparison of the two respective arrays of certainty values for that case. The distance

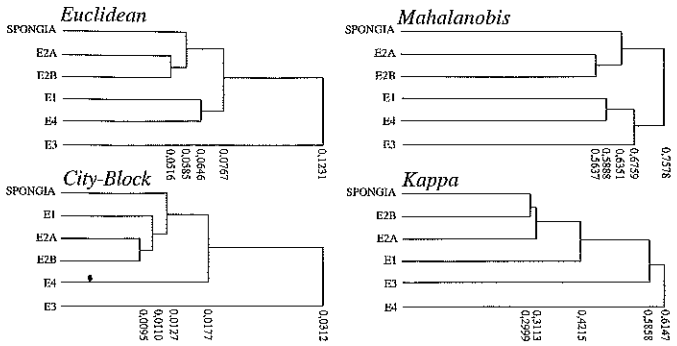

Fig. 2. Dendograms corresponding to the complete data matrix.

between two experts for the whole validation set is the average of the distance for each of the cases of the validation set.

3. Results

After performing the hierarchical cluster analysis we obtained a dendrogram representing the clustering of experts on the basis of each one of these distance matrices (Fig. 2). The complete set of dendograms cannot be presented here due to space limitations but can be found in Domingo, 1995.

Specifically, the results of the cluster analysis show that SPONGIA is always closer to a human expert than some human experts among them. None of the dendrograms has left SPONGIA isolated from the human experts. On the contrary, two of the experts, i.e. *E*3 and *E*4, often remained isolated in the dendrograms. SPONGIA forms a cluster with *E2B* and *E2A* in all dendrograms. In the dendrogram using the City-block distance with the complete data matrix and in the four dendrograms processing the *order* data matrix, *E1* also appears inside this cluster (*E2B*, *E2A* and SPONGIA). The uniformity of the dendrograms obtained with each measure and each data matrix (complete, *astro* and *order*) shows that the position of SPONGIA among the human experts does not depend on the measure we are using.

Taking into account the expert's ranking, we can state that SPONGIA occupies, in all dendrograms, a halfway position within the cluster formed by the three experts qualified with the highest proficiency. In particular, SPONGIA is very close to two human experts (*E2B* and *E2A*) that were both scored in second position in the average ranking of the inter-expert cross-evaluation.

4. Discussion

In the light of the results presented in this article we can state that the expert system SPONGIA provides identifications of sponge samples with reliability similar to the identifications given by qualified experts in the field. Thus, SPONGIA satisfies the requirements of an identification tool in the portion of the Porifera taxonomic tree which has been the object of this research.

Although in this validation project we did not aim to validate neither the performance nor the efficiency of the program some aspects may be discussed in the light of the comments written by the validators in the validation reports. Concerning performance, some of them pointed to redundant characters. In fact, part of the SPONGIA's performance is based on the collection of progressively more specific characters. In this way, an initial character may be seen as redundant when the user is asked for a more specific character e.g. the form of the megascleres after being asked for the number of axes in the megascleres.

If the user is able to answer the form of the megascleres, clearly the question about the number of axes becomes redundant. However, if the user is not a specialist, it may be easier for him to simply count the axes of the megascleres. The final ordination of the queries is a compromise between the proficiency of the potential user and the character set minimization to allow the identification of different taxonomic levels even with incomplete information.

Some experts mentioned some characters that they considered indispensable to make a certain identification but were missing in the bibliographical description. When the missing characters were diagnostic characters, both the results of SPONGIA and those of the human experts were less precise either in certainty level or in taxonomic rank. Thus, the behaviour of both the experts and SPONGIA was modified in a similar way in the presence of incomplete and uncertain data.

It should be noted that, in some particular cases, both the experts and SPONGIA agreed in identifying a case description as a different taxon than the literature did. Although this situation cannot be used to draw any conclusion, it confirms the lack of gold standard in the field as discussed in the introduction.

Acknowledgements

This research has been funded in part by a Generalitat de Catalunya fellowship (FI/91-193) and the CICYT project

TESEU (TIC91-0430). We are especially grateful to the experts in sponge systematics that collaborated in this validation project: Nicole Boury-Esnault, Michelle Kelly-Borges, Maurizio Pansini, Jean Vacelet and Rob van Soest.

References

Conruyt, N., Manago, M., Renard & J.L., Lévi, C. (1993). Une méthode d'acquisition de connaissances pour la classification et l'identification d'objects biologiques application au domaine des éponges marines. *Proceedings of the 13th International Symposium on Expert Systems and Applications*, Number 1 (pp. 485–495). Avignon: France.

Domingo, M. (1994). Evaluating the expert system approach to biological identification through application to Porifera. In R. van Soest & T. van Kempen & J. Braekman (Eds.), *Sponges in Time and Space*, (pp. 75–82). Balkema.

Domingo, M. (1995). An expert system architecture for taxonomic domains. An application in Porifera: The development of SPONGIA. Ph.D thesis, Universitat de Barcelona, Barcelona.

Domingo, M., & Uriz, M. J. (1998). Design and development of SPONGIA, an expert system for sponge identification . *Sci. Mar.*, *62* (1–2), 45–57.

Fieschi, M. (1990). Towards validation of expert systems as medical decision aids. *Int. J. Biomed. Comput.*, *26*, 93–108.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Geissman, J., & Schultz, R. (1988). Verification and validation of expert systems. *AI Expert*, *3* (2), 26–33.

van Gennip, E. & Talmond, J. (Eds.). (1995). *Assessment and evaluation of information technologies in medicine.* Amsterdam: IOS Press.

Hernández, C., Martín-Baranera, M., Sancho J.& Sanz, F. (1994a). POEMA: a computer program to compare quantitative data vectors. *Proceedings of the 12th International Congress of the European Federation for Medical Informatics (MIE)* (pp. 22–26). Lisbon: Portugal.

Hernández, C., Sancho, J., Belmonte, M., Sierra, C., & Sanz, F. (1994). Validation of the medical expert system RENOIR. *Comput Biomed Res*, *27* (6), 456–471.

Martín-Baranera, M., Sancho, J., & Sanz, F. (1996). Simulation applied to medical expert systems validation in absence of gold standard. In J. Brender & J. P. Christensen & J. R. Scherrer & P. McNair (Eds.), *Proceedings of Medical Informatics in Medicine*, (pp. 506–510). Copenhagen: IOS Press.

O'Keefe, R., Balci, O., & Smith, E. (1987). Validating expert system performance. *IEEE Expert*, *2* (4), 81–90.

Redier, H., Daures, J., & Michel, C., et al. (1995). Assessment of the severity of asthma by an expert system description and evaluation. *Am. J. Respir. Crit. Care Med.*, *151*, 352–354.

Thonnat, M. & Gandelin, M. (1988). An expert system for the automatic classification and description of zooplanktons from monocular images. *Proceedings of the 9th International Conference on Pattern Recognition* (pp. 114–118). Rome.

Verdaguer, A., Patak, A., Sancho, J., Sierra, C., & Sanz, F. (1992). Validation of the medical expert system PNEUMONIA. *Comput. Biomed. Res.*, *25*, 511–526.

Vogt, W., Nagel, D., & Sator, H. (1987). *Cluster analysis in clinical chemistry: a model.* New York: Wiley.

Wooley, J., & Stone, N. (1987). Application of artificial intelligence to systematics: Systex a prototype expert system for species identification. *Sist. Zool.*, *36* (3), 248–267.

Wyatt, J., & Speigelhalter, D. (1990). Evaluating medical expert systems: what to test and how? *Med. Inf. (London)*, *15*, 205–217.