How might we combine the information we know about a mass better? The use of mathematical models to handle medical data?

Peter Antal, M.Sc. [1], Herman Verrelst, M.Eng.[1], Sabine Van Huffel, M.Eng., Ph.D.[1], Bart De Moor, M.Eng., Ph.D.[1], Dirk Timmerman, M.D., Ph.D.[2], Ignace Vergote, M.D., Ph.D.[2]

Leuven, Belgium.

Department of Electrical Engineering (ESAT-SISTA/COSIC)[1], Katholieke Universiteit Leuven.

Department of Obstetrics and Gynecology[2], University Hospitals Leuven,

It is widely accepted that the combination of various sources of information can lead to better models for preoperative discrimination between malignant and benign adnexal masses. The advances in measurement techniques such as the visualization of the morphologic features of the mass, the assessment of the vascularisation of the mass and reliable serum tests provide a wide range of various observations to assist the doctor in this decision. Additionally the genetic background of relevant diseases and the role of other factors such as parity, age, lactation, contraceptives are better and better understood. Besides this large amount of medical background knowledge there is another type of information that gets more and more important with the spreading of information technologies (IT) services at clinics. The computer based documentation technologies of the patients provide cheap and natural access to a huge amount of past observations enhancing and facilitating the tedious and expensive data collection. Even if incompatibilities between the databases may exist, the use of computer based documentation technologies will cause an explosion in the near future with respect to the amount of available data for statistical analysis.

The growing amount of data and the ever more powerful and faster computers drastically changed the possibilities and methods in statistical data analysis. This trend is well characterized by the evolution of the techniques used in the assessment of the probability of malignancy of adnexal masses. At first various discrimination models were suggested mainly based on the medical background knowledge. Next, parametric models were used such as logistic regression models fitted to hundreds of observations. Subsequently more powerful non-parametric statistical models were used such as artificial neural networks using computer intensive optimization techniques. Recently adaptive probabilistic expert systems were suggested as a tool for integration of medical background knowledge and patient data. These models may require thousands of patient cases and use intensive and complex computations.

Of course, it would be a serious mistake to expect or to force the previous pattern everywhere since simple models (such as linear models) can provide a perfect solution in a specific problem. But the increasing number of patient data, the increasing computer power and more advanced statistical techniques make it possible to use more complex models. The following estimations

indicate the potential growth of medical data that will be available with the spread of the electronic patient records. The amount of networked data –dominated previously by geographical, astronomical or physical data - has increased by more than five orders of magnitude since 1980. Meanwhile the microprocessor power – the performance of a desktop computer – doubles every 2 years.

The proposed models for the classification of adnexal masses demonstrate the parallel development with respect to sample size, computer power and statistical techniques. The first discrimination models were based on single observations such as CA 125 - blood serum test -, pulsatility index – a characterization of vascularisation. The first multi-modal discrimination models were scoring systems constructed by leading experts in the field, tuned up and tested by observations, such as the RMI, which combines ultrasound properties, menopausal score and the value of serum CA 125. The statistical incorporation of the data in the models was achieved by applying multivariate logistic regression models. The logistic regression models require data sets with moderate size and standard statistical systems exist for model-fitting. Unfortunately it is not possible to incorporate medical background knowledge into the model and for complex problems the modeling capacity of the logistic regression models is not enough. Artificial neural networks provide a more powerful statistical model class, actually only the required computation and the sample size for model fitting restricts the modeling capacity. This method provides an ideal solution because the model complexity can be scaled with the sample size, thus the performance of such discrimination models in theory can approximate the optimal prediction.

However in practice the sample size and the model fitting set a hard limit to the complexity of the applicable neural network model, so to the performance of the model. Additionally this technique still does not provide a solution for the incorporation of the large amount of medical knowledge that is available about the nature of the adnexal masses. (These models are “black-boxes” in a sense that the model parameters cannot be interpreted for explaining the predicted probability of malignancy in medical terms.)

Recently the adaptive probabilistic expert systems were proposed as a potential candidate to solve the combination of large amount of background knowledge and statistical data. The modeling capacity of this method is similarly not limited in theory, but what is more important is that they can be balanced between the prior knowledge of a human expert and data integrating both of them in a single system. The following table summarizes the properties of the suggested methods.

Probabilistic approaches	Sample size	Statistical knowledge	Medical knowledge	Computation for model fitting	Computation for usage
Multivariate logistic regression	+	+	+	+	-
Artificial neural net	++	+++	+	++	-
Probabilistic expert models	++++	+++	+++++	++++	++++

Table: Properties of multi-modal models for the discrimination of malignant and benign adnexal masses.

These probabilistic methods can be used in the decision theoretic framework defining utilities or costs, so the decision-maker can use these models in a principled way either for the classification of the mass or to decide what additional measurement or test should be considered to get further relevant information.

In conclusion the importance of the integration of medical knowledge and statistical data should be emphasized. Successful mathematical models should not only be multi-modal (in the sense that they combine various inputs in the model), but also hybrid (in a sense that they combine effectively every relevant information into the model).

Timmerman D, Verrelst H, Bourne TH, De Moor B, Collins WP, Vergote I, Vandewalle J.: Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol 1999; 13: 17-25.

Antal P., Verrelst H., Timmerman D., Van Huffel S., De Moor B., Vergote I.: Bayesian networks in ovarian cancer diagnosis: potential and limitations, the 13th IEEE SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, June 23-24, 2000, Texas Medical Center, Houston, Texas

Castillo E., Gutiérrez J.M., Hadi A.S.: Expert systems and probabilistic network models, Springer 1997

Heckerman, D.: Learning Bayesian networks: The Combination of Knowledge and Statistical Data, Machine Learning, 20, 1995, pp. 197-243

Bishop, C.M.: Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995

Reichhardt T.: It's sink or swim as a tidal wave of data approaches, Nature, Vol. 399, 10 June 1999, www.nature.com

This work is supported by several institutions: the Flemish Government, Research Council K.U.Leuven: Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology); The FWO Research Communities:ICCoS (Identification and Control of Complex Systems) and ANMMM (Advanced Numerical Methods for Mathematical Modelling) and The Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs – Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997-2001): Modeling, Identification, Simulation and Control of Complex Systems; The Hungarian National Fund for Scientific Research (OTKA) under contract number T030586 and F-030763

The scientific responsibility is assumed by its authors.