in Artificial Intelligence (L. N. Kanal, J. Lemmer and T. S. Levitt, eds.) **3** 199–208. North-Holland, Amsterdam.

SPIEGELHALTER, D. J. and COWELL, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 447–466. Clarendon Press, Oxford.

SPIEGELHALTER, D. J., DAWID, A. P., HUTCHINSON, T. A. and COWELL, R. G. (1991a). Probabilistic causality assessment after a suspected adverse drug reaction: a case study in Bayesian network modelling. *Philos. Trans. Roy. Soc. London Ser. A* **337** 387–405.

SPIEGELHALTER, D. J., HARRIS, N. L., BULL, K. and FRANKLIN, R. C. G. (1991b). Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. Technical Report 91-4. MRC Biostatistics Unit, Cambridge.

SPIEGELHALTER, D. J. and LAURITZEN, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20** 579–605.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search.* Springer, New York.

SRINIVAS, S. and BREESE, J. (1990). IDEAL: a software package for the analysis of influence diagrams. In *Uncertainty in Artificial Intelligence* (L. N. Kanal, J. Lemmer and T. S. Levitt, eds.) **6** 212–219. North-Holland, Amsterdam.

TARJAN, R. E. and YANNAKAKIS, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.* **13** 566–79.

THIESSON, B. (1991). (G)EM algorithms for maximum likelihood in recursive graphical association models. Master's thesis, Dept. Mathematics and Computer Science, Aalborg Univ.

THOMAS, A., SPIEGELHALTER, D. J. and GILKS, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 837–842. Clarendon Press, Oxford.

THOMPSON, E. A. (1986). Genetic epidemiology: a review of the statistical basis. *Statistics in Medicine* **5** 291–302.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M., HABBEMA, J. D. F. and GELPKE, G. J. (1981). Comparison of discrimination techniques applied to a complex data-set of head-injured patients (with discussion). *J. Roy. Statist. Soc. Ser. A* **144** 145–175.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

VAN DER GAAG, L. (1991). Computing probability intervals under independency constraints. In *Uncertainty in Artificial Intelligence* (P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer, eds.) **6** 457–466. North-Holland, Amsterdam.

WALLEY, P. (1990). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

WARNER, H. R., TORONTO, A. F., VEASEY, L. G. and STEPHENSON, R. (1961). A mathematical approach to medical diagnosis—application to congenital heart disease. *Journal of the American Medical Association* **177** 177–184.

WERMUTH, N. (1976). Model search among multiplicative models. *Biometrics* **32** 253–263.

WERMUTH, N. and LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70** 537–552.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Analysis.* Wiley, Chichester.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5** 161–215.

ZADEH, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems* **11** 199–228.

Comment: Assessing the Science Behind Graphical Modelling Techniques

A. P. Dempster

These papers, labelled here CW (Cox and Wermuth) and SDLC (Spiegelhalter, Dawid, Lauritzen and Cowell), are welcome reviews of extensive collaborations. CW are the more limited of the pair in their aims, making a few points convincingly, most notably (1) that covariance-based regression models are conceptually distinct from the simultaneous causal models of econometrics, even when both varieties are expressed through identical linear equations, and (2) that models with covariance matrices corresponding to restricted graphical structures often give good fits to empirical matrices. The SDLC paper by contrast is a tour de force that aims to leave no relevant topic unmentioned.

Both sets of authors intend their formal models and computations to speak to issues of scientific knowledge and science-based decision making, and in particular both are concerned about the informal scientific understanding that motivates their formal models. CW are reluctant to use the term "causal," viewing it as too ambiguous, but the authors substitute nonspecific language such as "appropriate subject matter considerations." SDLC, in contrast, discuss "influence" and "relevance" that take "account of one's understanding of causal structure." The difference appears to be that CW wish to hold to the idea that informal prior knowl-

A. P. Dempster is Professor of Statistics, Harvard University, Statistics Department, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

edge of the phenomena is limited to descriptive aspects, while SDLC accept that understanding of causal processes is part of normal scientific thinking. My own view is closer to the latter: I do not see that informal causal understanding can or should be suppressed (Dempster, 1990). On the other hand, the formal statistical models of these papers need not to be interpreted causally, because the essential role of probability models is to produce inferences and predictions derived from uncertainty relations. Probabilistic causation strikes me as an oxymoron, since probability quantifies progressions of internal uncertain knowledge while causation identifies external mechanics of change. The graphical model restrictions proposed in both papers depend for support and credibility on the quality of the underlying science, including causal interpretations, and on how aptly the formal models capture that science, but the uses of the models are mainly inferential and decision oriented.

CW choose to make their arrows point away from explanatory variables and toward response variables, whereas SDLC make their arrows point away from the disease node about which predictions are desired and toward the predictive variables. It is interesting that the medically driven "algorithmic approach" to the problem of telephone diagnosis of blue babies reported in Franklin et al. (1991) also reverses the statistically driven SDLC choice. My sense is that the CW and Franklin et al. strategies represent mainline scientific thinking, and that SDLC may weaken their claim to be able to represent genuine medical expertise by bucking the tide. It is normal for a clinician to absorb seriatim relatively simple pieces of information about a patient and to attempt to reason from these data to disease states. Disease states are relatively complex and sometimes amorphous constructs, but when they are well defined and separated, as in the blue baby example, it is relatively easy to mentally cycle through looking for a match of each disease to a list of symptoms. The cleverness of the Franklin et al. (1991) algorithm appears to derive from the skill of the senior coauthor in putting together a logical sequence of tests for matches involving subsets of the predictors, in such a way as to tease out, with a low error rate, which of 26 disease categories is the true one. Why do SDLC not attempt to directly probabilize the clinical expertise displayed in Franklin et al. (1991)?

The tree structure displayed in Franklin et al. (1991) is an event tree including decision nodes, of a kind commonly found in elementary decision analysis texts. By contrast, the expertise required in the modelling phase of the direct acyclic graph (DAG) approach of SDLC, with its conditioning on disease states, and its request for discernments of Markov structures, seems far from the practical expertise of clinicians. The technology seems rigid as implemented because the DAG structure is required to be the same whatever the disease. In principle, SDLC and Franklin et al. (1991) are attempting the same task of constructing a set of logical constraints on multivariate outcomes, and SDLC should have the advantage because they have the more powerful tools of probabilistic logic, whereas Franklin et al. (1991) use simpler and unrealistic deterministic logic with failures of diagnoses transparently labelled on the graph. Nevertheless the brief numerical comparisons in subsection 5.3 of SDLC indicate that in the current state of the art the deterministic approach does better. It is also troubling that in Table 6 the CHILD model with assessed conditional independencies is minimally better, if at all, than the naive model that is sometimes called idiot's Bayes. Is their technology the right medicine?

Both papers surprised me by a near absence of discussion of what is known about sample selection, in contrast with their softer heuristic assessments of relations or nonrelations among variables, whether derived from subject matter considerations or causal insights. In their examples, both papers ultimately assume that samples exist from which multivariate models can legitimately be estimated. Obviously statistical relations among variables are strongly influenced by processes that select the units making up a sample. These processes operate in a social realm operationally separated from the biological processes of disease, whence the selection mechanisms can be considered causally independent of the biological mechanisms. SDLC do mention at one point the effect of an original medical judgment on the flow of referrals in examples given, but such considerations appear to play no role in discussions of what conditional independence assumptions to adopt, at least in a first pass. Can this be justified? I think not. Technical papers in applied statistics by tradition and habit move quickly to formal assumptions and in doing so hide large areas of subjective choices of (unit, variable) pairs for consideration and analysis. By implication these choices are traditionally viewed as made for good reasons, usually by the statistician's client or substantive collaborator and hence are removed from the statistical analysis. Consequences are that statisticians tend to express distorted views of the mix of subjective and objective elements in the mosaics to which their analyses contribute and pay little heed to the creative experiences of carrying analyses back to their elemental sources in accrued informal scientific knowledge. My critical attitude about the scientific limitations of many statistical models and analyses rests on a perception that their ties to overall contexts are too often too loosely tied down.

The pessimistic cast of the foregoing remarks applies to the present state of applied statistics, not the future. The understanding of complex multivariate relations and of associated computational strategies, as detailed

especially by SDLC, bodes well for ongoing development. Although critics will point to the dangers, there is also great promise in the strategy of constructing stochastic models that adequately represent uncertainties about the states of hidden aspects of complex phenomena, and hence they are bases for credible probabilistic inferences. Since SDLC are more pointed to complexity than are CW, most of my subsequent remarks are directed to their enterprise. These experienced colleagues scarcely need my advice to push on to more ambitious and intensive modelling efforts, constructing formal models of population incidence, of disease signs and their rates of progression, and of prospects for intervention and cure, before proceeding to combine such submodels into larger systems capable of supporting informed and sound decisions. As part of this process, many specific concerns may arise and lead to profitable debates, some of which I will attempt to stimulate.

A major topic of concern is "eliciting subjective judgments." Expert judgment enters model construction at the successive stages of choosing variables, choosing graphical structures and choosing numerical probabilities. The soundness of each stage merits questioning, not least as they relate to social responsibilities and public purposes, but most critical attention is focused on the last stage of probability assessment that often draws on teams of experts selected and coached for the purpose, generally by those responsible for the preceding stages of choice. A fascinating case study of elicitation is to be found in the major "NUREG-1150" study of five large nuclear power plants (U.S. Nuclear Regulatory Commission, 1990). Extensive external reviews in this case forced a delay of more than a year while elicitations were redone to improve the quality of the panels and the credibility of their judgments. The point to stress is that subjectivity does not imply freedom to choose the first numbers that come to mind, nor to choose handily available local staff as experts. Analyses depend on subjective evidence to bridge gaps and supplement inadequate empirical data bases, but such evidence is not of a wholly different character from empirical evidence, as many discussions of systems and technologies for elicitation might suggest. Numbers provided by an expert are acceptable only if there is credible evidence that they are distillations of the expert's accumulated knowledge and experience. When final inferences have sensitive dependence on expertise, the information and analyses on which expert judgments depend need to be clearly set forth so they can be challenged, debated and revised, much as statistical data analysis and models are subject to criticism and model revisions. My sense is that much remains to be done by way of developing and testing quality control standards for evidential inputs from experts. There is no subjective nirvana, just as there is no objective nirvana, but real expertise exists and substantial payoffs can be expected from using it well.

An obvious way to decrease the influence of elicited expertise on diagnosis, or on risk assessment in general, is to increase the weight of documented empirical studies, including quality-checked data bases. Again it is instructive to compare and contrast the NUREG-1150 engineering example, where the sample size is 5 from a world-wide population of order 1,000, and where the physical description of each plant, though dauntingly complex, is more easily accessible and decomposable into independent subsystems than is the human body. My sense is that in the engineering analysis vastly more data sets on components, such as data on reliabilities of various types of critical elements, were assembled than is typical of medical expert systems. In place of data on physical components, medical statistics tends to rely on collections of studies and associated meta-analyses. These give clues to variation among different patient populations, and so may help inform expert prior assessments that depend on patient flows through treatment systems. A remaining difficulty with population thinking is that any specific patient under diagnosis automatically belongs to many cross classified subpopulations, depending on age, sex, ethnic identity, social class and so forth. Available data comes from margins and complex mixtures of these populations. There is large scope to jointly model networks and mixtures to facilitate combination of data from multiply interrelated varieties of populations.

SDLC correctly adopt a general approach to treating the complexity of an individual patient as demanding stochastic modelling of interacting biological systems such as heart and lungs. The currently fashionable alternative approach to complexity through deterministic chaos theory may be promising when one subsystem operates under controlled circumstances, but is scarcely capable of faithfully representing the complex social and biological processes routinely encountered by risk assessors (cf. Casdagli, 1992). I am uncomfortable, however, with the basic principle of simplicity used by both CW and SDLC that performs radical surgery on the proliferating parameter sets of highly multivariate statistical models. DAG models are mathematically transparent, have relatively few parameters and suggest elegant and fast computational strategies. Less felicitously, however, the models are expressed through variables that are rarely better than crude proxies for hidden variables that actually express underlying causal mechanisms, whence the substantive understanding that could justify DAGs if the hidden variables were observable is less than compelling and may be misleading when used to justify DAG assumptions for simple (e.g., dichotomous) variables. I believe that wholesale parameter reduction as widely practiced

in contemporary statistics is not the only way to achieve simplicity. The main lesson that I took away from Wermuth's doctoral research (cf. Dempster, Schatzoff and Wermuth, 1977) is that smooth systems of declining parameter values are usually a more efficient way to simplify statistical complexity than sharp cutoffs that set most parameter values to zero. Computational strategies of choice then become radically different. Classical estimation techniques that are adequate with relatively few parameters must be replaced with Bayesian or similar methods that reflect prior assessments of patterns of smooth decline. Donoho et al. (1992) illustrate a notable non-Bayesian approach. My own preference is for Bayesian models with many more hidden variables and many more dependence parameters than SDLC allow, to have a reasonable possibility of capturing actual mechanisms. I believe that rapidly developing computing power and algorithms that sample posteriors should be used to implement and test more complex Bayesian models.

Beyond the elicitation of priors and beyond the problem of simplifying the complex structures of highly multivariate and selectively filtered populations encountered in real practice, there remains a gray area that SDLC address briefly in two sentences as situations where "the number of assessments made is insufficient to specify a joint distribution uniquely." The use of maximum entropy or other arbitrary prior generation principles typically leads to exactly the unrealistic procedures that the smoothing of large parameter sets is designed to avoid. SDLC fail to mention the belief function approach (Shafer, 1976) that Dempster and Kong (1988) show fits naturally into network modelling built on decompositions of evidence into independent sources similar in spirit to the "graphical modelling" approach of SDLC. It is my view as a coinventor of the BEL theory that it is a near cousin of the Bayesian strategy that descends directly from classical subjective probability and is not a foreign interloper from distant tribes of semicoherent formal systems. Unlike the naive upper and lower probability models that have been studied by Good, Walley and others, the BEL system constructs models from judgmentally independent assessments on knowledge spaces and combines the components by a simple precise rule that reduces to the Bayesian rule for combining likelihood and prior in the special Bayesian case. The chief hindrance to developing and testing BEL models for probabilistic expert systems has been computational difficulties. Shafer, Kong and others showed in the mid-1980s how to decompose BEL computations coincidentally with the parallel demonstrations of Lauritzen and Spiegelhalter (1988) that SDLC feature. But these clever algorithms only stave off computational complexity temporarily. The future of both Bayesian and BEL approaches depends on the revolution that has been gathering speed for the past five years on Monte Carlo posterior sampling.

ACKNOWLEDGMENTS

I thank Emery Brown for helpful discussions on both medical and statistical issues. My work is partially supported by ARO Grant DAAL03-91-0089 and NSF Grant DMS-90-03216.

Comment: Conditional Independence and Causal Inference

Clark Glymour and Peter Spirtes

Fourteen years ago, in an essay on conditional independence as a unifying theme in statistics, Philip Dawid wrote that "Causal inference is one of the most important, most subtle, and most neglected of all the problems of Statistics" (Dawid, 1979a). Only shortly later, several statisticians (Wermuth and Lauritzen, 1983; Kiiveri and Speed, 1982) introduced frameworks that connect conditional independence, directed acyclic graphs (hereafter DAGs) and causal hypotheses. In these models the vertices of a DAG G represent variables, and a directed edge $X \rightarrow Y$ expresses the proposition that some change in variable X will produce a change in Y even if all other variables represented in G are prevented from changing. The power and generality of DAG models derive from their dual role in representing both causal or structural claims and

Clark Glymour is Alumni Professor of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, and Adjunct Professor of History and Philosophy of Science, University of Pittsburgh. Peter Spirtes is Associate Professor of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.