Knowledge Engineering for Very Large Decision-analytic Medical Models

Marek J. Druzdzel, Ph.D.,[1] Agnieszka Onisko, M.S.,[1,2] Daniel Schwartz, M.D.,[1]
John N. Dowling, M.D.[1] and Hanna Wasyluk, M.D., Ph.D.[3]

[1] Decision Systems Laboratory, School of Information Sciences, Intelligent Systems Program, and Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA,
marek@sis.pitt.edu, aonisko@sis.pitt.edu, pathdan@cbmi.upmc.edu, dowling+@pitt.edu
[2] Institute of Computer Science, Bialystok University of Technology, Bialystok, 15-351, Poland,
aonisko@ii.pb.bialystok.pl
[3] Medical Center of Postgraduate Education, Warsaw, 01-813, Marymoncka 99, Poland,
hwasyluk@cmkp.edu.pl

Graphical decision-analytic models, such as Bayesian networks, are powerful tools for modeling complex diagnostic problems, capable of encoding subjective expert knowledge and combining it with available data. Practical models built using this approach often reach the size of tens or even hundreds of variables. Constructing them requires practical skills that go beyond simple decision-analytic techniques. These skills are difficult to gain and there is almost no literature that would aid a modeler who is new to this approach.
Drawing on our experiences in building large medical diagnostic models, we list typical problems encountered in model building and illustrate the knowledge engineering process with examples from our networks.

INTRODUCTION

Decision analysis has had a major influence on computer-based diagnostic systems. The field of Uncertainty in Artificial Intelligence, through which this influence was funneled, has developed practical modeling tools, the scale of which goes far beyond simple models developed in decision analysis. Bayesian networks[1] are one such tool. Also called causal networks, they are directed acyclic graphs modeling probabilistic dependencies among variables. The graphical part of a Bayesian network reflects the structure of a problem, while individual, local interactions among its variables are quantified probabilistically. One of the main advantages of Bayesian networks over other schemes used in computer-based diagnostic tools is that they readily combine existing frequency data with expert judgment within the probabilistic framework. They have been employed in practice in a variety of fields, including engineering, science, and medicine (for examples of successful real world applications of Bayesian networks, see March 1995 special issue of the journal *Communica-*

tions of the ACM) with some models reaching the size of hundreds of variables.

Bayesian networks are powerful tools for modeling complex problems and can be extremely valuable in medical diagnosis, where observations rarely imply a diagnosis with absolute certainty. Often, hospitals and clinics collect patient data, which over time allow for discovering statistical dependencies and potentially improving the overall quality of diagnosis. When data sets are sufficiently large and complete, the structure of Bayesian networks can be learned purely and directly from the data (see (2) for an overview of the available methods). Unfortunately, most data sets are small and contain many missing values. In practice model building has to be based on a combination of expert knowledge and statistics extracted from the data.

Large models pose a number of problems in terms of knowledge engineering. The sheer scale of the model requires numerous interactions with the experts, eliciting the structure of the model and obtaining the probabilities that quantify the interactions. A model of a hundred variables, for example, may require several thousand numerical parameters. If traditional decision analytic techniques are applied, both the structure and the numbers require countless sessions with an expert. This is practically impossible given the value of a medical expert's time. Furthermore, even if the expert were available, trusting the quality of several thousand numbers would be naïve. Skills and experience in constructing large Bayesian networks are difficult to gain and there is almost no literature that would aid a knowledge engineer, who is new to decision-analytic modeling.

The goals of this paper are to describe problems that occur in building large medical Bayesian network models and to illustrate some practical techniques to overcome them. We draw on our own collective ex-

perience in knowledge engineering of large diagnostic models, from the point of view of both knowledge engineers (Druzdzel and Onisko) and experts (Schwartz, Dowling and Wasyluk). We have collaborated in building diagnostic systems for diagnosis of liver disorders, processing of liver pathology data, and various epidemiological models. In our work, we often had access to medical data sets, although these were not large or complete enough to fully automate the model building process. An additional complication was that we collaborated across two continents and faced differences in terminology, patient population, and medical practice. In some cases, we constructed the structure of a model and used an existing clinical database to learn its numerical parameters. In other cases, we obtained both the structure and the probabilities from our experts. In the remainder of this paper, we assume that the reader is familiar with the basic concepts of Bayesian networks.

MODEL STRUCTURE

The graphical structure of a Bayesian network models important and fairly robust structural properties of the domain – direct interactions among the domain variables and, indirectly, conditional independencies among them. There are two important reasons why it is important to devote much attention to the structure of the network. The first is that there is much anecdotal and some empirical evidence[3] that structure is more essential than numbers for the model's performance. The second, arguably more important, reason has to do with the human side of the modeling process. The graphical structure reflects the modeler's understanding of the domain. If encoded well, this structure will facilitate obtaining numerical parameters, and in creating future extensions and revisions of the model. The structure is also important to the ultimate user of the system, as it can be examined in case of disagreement between the user and the system. The structure of the model can help its user in understanding the relevance of various findings in individual patient cases. A decision support system based on decision-analytic methods can, thus, in addition to aiding expert diagnosticians, play a role in medical training.

Iterative character of the modeling process
Model structuring is a complex task and is best done iteratively. In our experience, regular short sessions with the expert (or experts) work well. Between these sessions, the knowledge engineer can refine the model in order to improve the quality of the questions brought to the expert. This saves valuable expert time and prevents expert exhaustion. Sessions with the expert have then more of a verification and re-

finement character. It helps when the knowledge engineer understands the domain, at least roughly. Our advice to the knowledge engineers is to read at least a relevant section of a medical textbook on the topic of the model so that he or she is roughly familiar with the vocabulary, the variables involved, and the interactions among them.

Where to start?
It is useful to start model building from the main foci of the system under construction. In case of a diagnostic system, these foci are variables representing the disorders in question and we suggest that the model building process start there.

In our work, we noticed an initial confusion on the part of some physicians as to whether the Bayesian network variables corresponding to diseases model the actual presence of the disease(s), or merely the arrival at a diagnosis indicating their presence. There is no doubt that it is the former. Decision making under uncertainty requires that we model uncertainty related to the true state of the world. Only the true state of the world, usually unknown at the time of decision-making, in combination with the decision, determines the outcome of decision-making. In case it is important to model diagnostic decisions and utility related to being right or wrong, we advise that an influence diagram, rather than a Bayesian network, be used.

We encountered one particular problem fairly often. Very often medical practice makes an important simplifying assumption that diseases are mutually exclusive. In other words, it is assumed that a patient has at most one disorder. This assumption, reflected in a number of medical data sets, has serious implications on the model structure. For practical purposes, this implies that all diseases can or should be modeled by a single node, whose mutually exclusive states are the various possible disorders. We would like to stress that this assumption is unnecessary when using Bayesian networks, which are able to model the simultaneous presence of multiple disorders correctly.

How to proceed from there?
Once we have our foci, we can proceed with adding other nodes. This may turn out to be challenging. A practicing diagnostician attempts to consider every relevant finding that may have a bearing on the diagnosis. A diagnostic model needs to encode each of these findings explicitly in order for them to be considered in the diagnosis. When creating a Bayesian network model, it is tempting to include every finding that is potentially relevant and to add connections between variables, even if they are very weak. For example, in our model for the diagnosis of liver disorders, a non-hepatic cancer node could be argued to influence a large number of the nodes in the network. Since in the real world everything is connected with

everything, the effect is a spaghetti of nodes and connections among them. This very often obscures the model and introduces unnecessary complexity. It also often backfires in terms of the elicitation effort needed when quantifying the network. It is, therefore, crucial to focus on the most important variables. We have found out that focusing on a few of the medical findings that the expert judges to be important works well, especially if they form a sub-model, a piece of the network that can be reasonably studied in separation from other parts of the model. Adding one new variable at a time and asking the question *"How does this variable relate to what we already have in the model?"* seems to be a good heuristic. The software tool that we use in our modeling efforts, `GeNIe`, allows for structuring complex models into sub-models. This view allows one to double-check whether all important sub-models have been considered and how they interact.

The importance of causality

Probability theory, underlying Bayesian networks, does not put any restrictions on the directions of the arcs between nodes. In fact, arcs specify the direction of conditional probability and can be reversed by means of Bayes theorem. We have found in practice that it is very helpful to use a causal framework for modeling the interactions among the variables. There are several reasons for this. The foremost is that a causal graph is usually the easiest for an expert or a user to understand and conceptualize. In our experience, though, the above statement holds only for medical experts who are already somewhat familiar with directed graphs. We have noticed that during initial modeling sessions medical experts tend to produce graphs with arcs in diagnostic direction. This is a transient tendency that disappears after a few model-building sessions.

Causal graphs also facilitate interactions among multiple experts. Causal connections and physiological mechanisms that underlie disease processes are part of medical training and provide a common language among the experts participating in the session. We have observed that experts rarely disagree about the model structure. A brief discussion of the physiology of the disease usually leads to a consensus.

The second important reason for following the causal structure of the domain is that this usually ensures satisfaction of the Markov condition,[1,4] which ties conditional probabilistic independence with the structure of the graph. Testing for conditional independence is generally easier when the graph is causal. Finally, when the direction of arcs coincides with the direction of causality, it is usually (although not always!) easier to obtain probability judgments. Very often medical textbooks report conditional probabilities in the causal direction, for example in sensitivity and specificity data of medical tests.

Sometimes following the causal structure is difficult because of lack of medical knowledge – we simply do not know more than that there is a correlation. At other times, it is possible to use proxy measures for variables that are hard or impossible to observe. For example, we used INR (International Normalized Ratio of prothrombin) as a proxy variable for hepatic failure and made it a parent node of such nodes as *Palmar erythema* and *Skin erythemic eruptions*. Strictly speaking, elevated INR does not cause either of them, but it reflects diminished clotting proteins in blood, a manifestation of hepatic failure. Elevated INR can therefore be used as a proxy for hepatic failure.

Modeling the practice or how things really are?

Sometimes, our common sense modeling efforts may clash with medical practice. Consider tests for (viral) hepatitis C. The test for anti Hepatitis C antibody (Anti-HCV) is usually considered to be a screening test because it is fast, inexpensive, and has a high sensitivity. However, it also has a high false-positive rate (around 20%). So, when the Anti-HCV is positive, a confirmatory test, usually the RIBA (Recombinant ImmunoBlot Assay) is performed. The RIBA is complicated, time-consuming and expensive, but has a low false-positive rate. Public Health Service Guidelines for the practicing physician state that when both are positive, the patient is essentially defined as having Hepatitis C, as there is essentially no gold standard that is more accurate than the two tests combined. From a purely statistical standpoint, however, there is still a very low, but definable, chance the patient does not have hepatitis C, even if both tests are positive. The correct model for this situation is one in which Hepatitis C is a parent of both tests. In that case, the model will show a non-zero chance for no disease, even if both tests are positive.
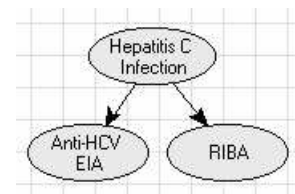
Clarity test

A useful concept that originates from decision analysis is the clarity test. Essentially, for each element of the model, such as a variable, its states, etc., we need to ask ourselves the question whether it has been clearly defined. Vague and imprecise definitions of model elements will generally backfire at some stage of modeling, certainly when it comes to elicitation of numerical parameters. We believe that the problem

of clarity is broader than just model building and applies, for example, to whenever medical data is collected. We have encountered many unclear definitions of variables included in patient records.

Cross-cultural differences

We had the opportunity to collaborate with expert diagnosticians in two different countries. Our experts had different kinds of expertise in terms of cultural and social environment. Dr. Wasyluk is a practicing hepatologist with over 20 years of clinical experience in Warsaw, Poland. Dr. Schwartz is a hepatic pathologist with over 10 years of pathology experience in the USA. We have observed that some of the tests used in Poland are not used in the USA and vice versa. For example, the *LE cells* test is not commonly performed in the USA, while it is assigned some diagnostic value in Poland. Another apparent difference is the reliance of our US-trained experts on objective measures and a considerable distrust for subjective measures (for example, self-reported data). Our experts have sometimes differed in their opinion as to what should be included in a model. We believe that in addition to differences in culture, a factor that might have played a role here is the difference in professional experience (clinic vs. laboratory).

Other useful guidelines

Sometimes we may decide to leave out a moderately important variable because of a complete lack of information about how it interacts with other variables in the model or because it is difficult to obtain its value reliably. For example, in the following model, the variable *History of viral hepatitis* was originally meant to be based on patient's self-report. Since this is not very reliable and has serious consequences for the model by "screening off" reliable results of antibody tests from the disorder nodes, we decided to give it the objective meaning and add a child node to it that reflected patient self-report. We decided not to
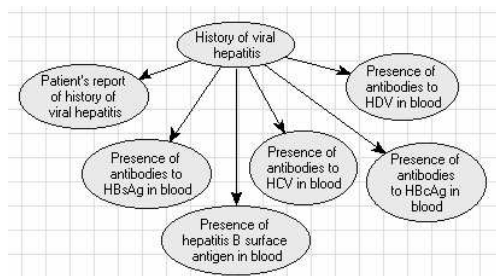
ignore self-reported data, as it is readily available at no expense, but to model structurally its inherently unreliable character.

While it is generally advisable to limit the variables that are modeled to those that are most important, there are exceptions to this heuristic. Sometimes, adding a variable that provides only weak diagnostic relevance is very inexpensive, computationally and in terms of expert time expenditure, and can improve the quality of the model. In general, variables whose values will be known at the outset of the diagnostic session, such as risk factors, are good to include in the model, especially if their interaction with the disorder node(s) is well known. Once the value of those variables is known, the patient enters a more specific prevalence group and this usually benefits the diagnosis.

The knowledge engineer should watch out for nodes with too many parents – every node is indexed by its parents and the size of its conditional probability table grows exponentially in the number of parents.

MODEL PARAMETERS

Once the structure of the model is composed, the Bayesian network has to be quantified, i.e., all prior probabilities of parent-less nodes and all conditional probabilities of nodes given their parents have to be specified. For a large model this may imply thousands of numbers and this stage often provides a sobering experience to beginning modelers. It is often hard to split elicitation of parameters from elicitation of structure in practice and it is quite common to move back and forth between the two.

Decision analysis provides several time-proven techniques for probability elicitation. Unfortunately, most of these techniques require numerous preference judgments on the part of the expert and quickly become impractical for very large models.

Those modelers who have access to medical data sets can easily learn the network parameters from them. We were in this lucky situation in case of our model for diagnosis of liver disorders, where the data set comprised of several hundred records carefully maintained and extended over the period of several years by Dr. Wasyluk. Still, even though one might have access to good medical data, much care has to be exercised in adopting them. The data matches a specific patient population that is not necessarily the same as the population for which the model is being developed.

When there is no data available, large models can be quantified using parametric distributions, such as Noisy-OR gates.[1,5] Essentially, Noisy-OR gates assume that the parents of a node act through independent causal mechanisms that can each cause the effect but are noisy in the sense that their presence is not sufficient to cause the effect. Similarly, a Noisy-OR gate can be *leaky*[5] in which case even if none of the parents are active, the effect may still materialize.

An example of an interaction that can be approximated by a Noisy-OR gate is the interaction between *Liver disorder*, *Pancreatic cancer*, and *Jaundice*. Both *Liver disorder* and *Pancreatic cancer* can cause

Jaundice, and the mechanism by which they cause *Jaundice* is independent (Cancer involving the head of the pancreas tends to obstruct the common bile duct (CBD). This is the only route by which bile, which is manufactured in the liver, can leave it under normal conditions. So if the CBD is obstructed, the liver has no option but to release the bile into the bloodstream instead, resulting in jaundice.)
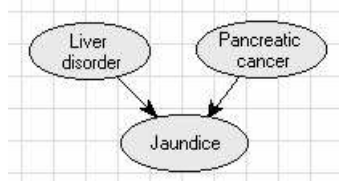
To give the reader an idea of the magnitude of savings in case of binary variables, Noisy-OR gates approximate the full conditional probability distribution by n numbers, instead of 2^n, where n is the number of parents of a node. Noisy-OR gates are clearly an approximation to an idealized probability distribution and may raise questions as to how good they are. Those modelers who hesitate to search for Noisy-OR-like interactions in their models are encouraged to consider what will be more precise: a Noisy-OR gate with 10 parameters elicited carefully from an expert or a full conditional probability distribution consisting of 1,000 numbers elicited from the same expert and within the same amount of time. While we do not doubt in the theoretical power of the generalized specification, we believe that the 1,000 numbers cannot be fully trusted because of the amount of effort that goes into obtaining them.

VALIDATION AND REFINEMENT

Finally, an important modeling step is validation. Models should ideally be built with validation in mind and validation should enter the modeling cycle rather than being its terminal step. When there are clinical data available, the model can be validated by learning its parameters from one part of the data set and using the remaining part to test the model predictions. In addition to such decision-analytic tools as sensitivity analysis and value of information, we have successfully applied scenario analysis. Scenario analysis in a diagnostic setting consists essentially of studying a patient case and analyzing model output, comparing it to the reasoning of an expert diagnostician. There is an interesting issue that we would like to raise at this point. It is not uncommon during the validation phase that the system recommendation conflicts with the intuition of the expert. How do we deal with this disagreement? It is not obvious who is right: the expert or the model. While in the model building phase, the odds are against the model, a clash of human intuition and the model output can

occasionally lead to important insights on the part of the expert.

Finally, it is important to remember that model refinement that originates from validation does not necessarily make the model more elaborate. Sometimes an important insight from validation is that some elements of the model are useless and can be reduced.

References

1. Judea Pearl (1998). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

2. Clark Glymour and Gregory F. Cooper, eds. (1999), *Computation, Causation, and Discovery*, AAAI Press

3. Malcolm Pradhan, Max Henrion, Gregory Provan, Brendan del Favero and Kurt Huang (1996), The Sensitivity of Belief Networks to Imprecise Probabilities: An Experimental Investigation, *Artificial Intelligence*, 85(1-2):363-397

4. Marek J. Druzdzel and Herbert A. Simon (1993), Causality in Bayesian Belief Networks, In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, Morgan Kaufmann Publishers, Inc., San Francisco, CA, pages 3-11

5. Max Henrion (1989), Some Practical Issues in Constructing Belief Networks, In L.N. Kanal, T.S. Levitt and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence 3*, pages 161-173, Elsevier Science Publishers B.V., North Holland