Journal of Machine Learning Research 4 (2003) 295–316 Submitted 11/01; Published 7/03

Combining Knowledge from Different Sources in Causal Probabilistic Models

Marek J. Druzdzel MAREK@SIS.PITT.EDU
Decision Systems Laboratory
School of Information Sciences
and Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260, U.S.A.

Francisco J. Díez FJDIEZ@DIA.UNED.ES
Dept. Inteligencia Artificial
Universidad Nacional de Educación a Distancia
c/ Juan del Rosal, 16
28040 Madrid, Spain

Editors: Richard Dybowski, Kathryn Blackmond Laskey, James Myers and Simon Parsons

Abstract

Building probabilistic and decision-theoretic models requires a considerable knowledge engineering effort in which the most daunting task is obtaining the numerical parameters. Authors of Bayesian networks usually combine various sources of information, such as textbooks, statistical reports, databases, and expert judgement. In this paper, we demonstrate the risks of such a combination, even when this knowledge encompasses such seemingly population-independent characteristics as sensitivity and specificity of medical symptoms. We show that the criteria "do not combine knowledge from different sources" or "use only data from the setting in which the model will be used" are neither necessary nor sufficient to guarantee the correctness of the model. Instead, we offer graphical criteria for determining when knowledge from different sources can be safely combined into the general population model. We also offer a method for building subpopulation models. The analysis performed in this paper and the criteria we propose may be useful in such fields as knowledge engineering, epidemiology, machine learning, and statistical meta-analysis.

Keywords: Probabilistic models, Bayesian networks, numerical probabilities, elicitation, selection biases, learning, combining knowledge.

1. Introduction

Development of the theory of directed probabilistic graphs, notably Bayesian networks (Pearl, 1988) and influence diagrams (Howard and Matheson, 1984), has caused a considerable interest in applying probability theory and decision theory in intelligent systems—see Henrion et al. (1991) for an accessible overview of decision-theoretic methods in artificial intelligence. Directed graphical probabilistic models have been successfully applied to a variety of problems, including medical diagnosis, prognosis, and therapy planning, epidemiology, machine diagnosis, user interfaces, natural language interpretation, planning, vision, robotics, data mining, and many others.

One of the most serious hurdles in practical application of probabilistic methods is the effort that is required of model building and, in particular, of quantifying graphical models with numerical probabilities. To make this task doable, typically knowledge engineers rely on a variety of sources that include expert knowledge, literature, available statistics, and databases of relevant cases. Very often, the structure of the model is elicited from experts and the numerical probabilities are learned from databases. Lack of attention to whether the sources are compatible and whether they can be combined can lead to erroneous behavior of the resulting model. While most knowledge engineers realize the danger of misapplication of data that describe different population groups, they often fail to appreciate purely statistical effects that play a role in probabilistic information.

In this paper, we first demonstrate the problem by showing that even such seemingly population-independent characteristics as sensitivity and specificity of medical symptoms can vary significantly between hospital patients and the general population. Although this variability has been reported in the medical literature for decades—see, for instance Ransohoff and Feinstein (1978) and Knottnerus (1987)—many of today's epidemiogical studies on the assessment of diagnostic tests fail to mention it, and, to our knowledge, researchers in the area of artificial intelligence have never considered it when building probabilistic models. This entails a significant risk because, as we show in this paper, collecting these statistics in one setting and using them in another can lead to errors in posterior probabilities as large as several orders of magnitude. We use the framework of directed probabilistic graphs to systematize our observation, to explain the risks of naive knowledge combination, and to offer practical guidelines for combining knowledge correctly. The problems that we are referring to are due to purely statistical effects related to selection phenomena. They may occur when data or knowledge are collected from different subpopulations and subsequently combined into one model, or even when the parameters for a causal model are obtained from a the same subpopulation in which the model is applied. On the contrary, these problems have nothing to do with small databases, missing data, or unreliable expert judgment.

Our analysis was inspired by our practical experiences in building medical diagnostic systems in independently conducted projects, in which we typically build a causal graph from expert knowledge and obtain the parameters from data, textbooks and expert estimates. We encountered a puzzling phenomenon that led to an initial disagreement between us. We have subsequently analyzed the problem, gaining insight that escaped each of us despite our fairly solid theoretical preparation and considerable field experience. We suspect that many knowledge engineers, data miners, and epidemiologists who apply machine learning algorithms face similar problems, often not realizing them. Hence, the purpose of the current paper is not only to point out the risks of the unwary use of data sets, but also to offer criteria for deciding when one or several knowledge sources can be used in the construction of a probabilistic model. In particular, we are interested in profiting from subpopulation data—for instance, the records of a database or the set of past cases on which a human expert bases the probability estimates. This is of most practical interest because usually knowledge engineers and automatic learning systems do not have access to general-population data.

Because of our experience with medical models, we use in this paper clinical diagnostic examples, but the principles applied and the conclusions of this analysis are general. An isomorphic problem is, for example, machine diagnosis, where models are built based on a combination of field experience, device specification, and repair shop data. Yet another is fraud detection, where models are based on general population characteristics combined with customer transaction data.

The remainder of this paper is structured as follows. Section 2 presents the conceptual framework and, in particular, it offers an introduction to Bayesian networks. Section 3 presents two motivating examples which show the risks of naive use of data. We analyze these examples in depth and explain the statistical reasons for why those models are incorrect. In Section 4, which is the core of the paper, we offer two graphical criteria for combining knowledge from different sources and for building subpopulation models. Finally, Section 5 discusses the implications of our analysis for knowledge engineering, machine learning, epidemiology, and meta-analysis.

In the mathematical analysis, we will use upper case letters, such as V, to denote variables, and lower-case letters, such as v, to denote their outcomes. When a variable V is binary, we will use $+v$ and $\neg v$ to denote the truth and falsity of the proposition or its positive and negative value respectively. Correspondingly, $\mathrm{pa}(X)$ will denote the direct predecessors (parents) of a node in a graph that models a variable X, and $\mathrm{pa}(x)$ will denote a combination of values of parents of a variable X. In the same way, $\mathrm{anc}(X)$ will denote the set of ancestors of X. A node U is an ancestor of X if either U is a parent of X or there is a node V such that V is a parent of X and U is an ancestor of V.

2. Conceptual Framework

Directed graphical models are a prominent class of probabilistic modeling tools, arguably most widely applied in practice. Two most popular instances of this class are Bayesian networks (Pearl, 1988) and influence diagrams (Howard and Matheson, 1984). Influence diagrams can be viewed as Bayesian networks enhanced with an explicit representation of decisions and utilities over the outcomes of the decision process. While in the sequel we will focus on Bayesian networks, we would like to point out that our results apply equally well to influence diagrams.

2.1 Fundamentals of Bayesian Networks

Bayesian networks are acyclic directed graphs in which nodes represent random variables and arcs represent directed probabilistic dependencies among them. A Bayesian network encodes the joint probability distribution over a finite set of variables $\{X_1,\ldots,X_n\}$ and decomposes it into a series of conditional probability distributions over each variable given its parents in the graph. More specifically, for each configuration $\mathrm{pa}(x_i)$ of $\mathrm{pa}(X_i)$ (the parents of X_i), there is a conditional probability distribution $\Pr(x_i|\mathrm{pa}(x_i))$. The conditional probability of a parentless node is just its prior probability: $\Pr(x_i|\emptyset) = \Pr(x_i)$. The joint probability distribution over $\{X_1,\ldots,X_n\}$ can be obtained by taking the product of all of these conditional probability distributions:

$$\Pr(x_1,\ldots,x_n) = \prod_{i=1}^{n} \Pr(x_i|\mathrm{pa}(x_i)) \,. \tag{1}$$

It can be proven that this product is a probability distribution and that it satisfies the Markov condition, i.e., that the probability of each variable given its parents is independent of its non-descendants in the graph. For instance, if X_i is an ancestor of neither X_j nor X_k, then $\Pr(x_i|\mathrm{pa}(x_i),x_j,x_k) = \Pr(x_i|\mathrm{pa}(x_i))$.

Figure 1 shows an example Bayesian network modeling four variables: a disease D, a symptom S, a medical test result T, and admission to hospital H. A direct arc between D and S denotes the fact that whether or not an individual is a carrier of the disease D will impact the likelihood of her showing the symptom S. Similarly, an arc from D to T denotes that presence of D influences the test result T. Arc $S \rightarrow H$ means that the probability that an individual goes to hospital depends on whether she shows symptom S.

Lack of directed arcs is also a way of expressing knowledge, notably assertions of conditional independence. For instance, the absence of arc $T \rightarrow H$ means that the patient's decision to go to

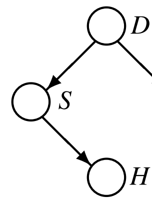

Figure 1: Example Bayesian network consisting of four variables: D, S, T, and H.

hospital depends only on whether she observes symptom S. It does not depend directly on D because she does not know whether she has the disease or not, nor on T because the test is not performed until she is admitted to the hospital. These causal assertions can be translated into statements of conditional independence: H is independent of D and T given S. In mathematical notation,

$$\Pr(h|s) = \Pr(h|s,d) = \Pr(h|s,t) = \Pr(h|s,d,t) \ .$$

Please note that this a particular instance of Markov condition: H is independent of its non-descendants (D and T) given its parents (S).

In the same way, the absence of link $S \rightarrow T$ means that the result of the test does only depend on D and not on the presence of the symptom, and the absence of link $T \rightarrow S$ has a similar interpretation. Again, this causal knowledge translates into probabilistic assertions: S and T are conditionally independent given D:

$$\Pr(s,t|d) = \Pr(s|d) \Pr(t|d)$$

or

$$\Pr(s|d,t) = \Pr(s|d) \ ,$$

which is another instance of Markov condition.

These properties imply that

$$\begin{aligned} \Pr(d,s,t,h) &= \Pr(h|d,s,t) \Pr(s|d,t) \Pr(t|d) \Pr(d) \\ &= \Pr(d) \Pr(s|d) \Pr(t|d) \Pr(h|s,t) \ , \end{aligned}$$

i.e., the joint probability distribution over the graph nodes can be factored into the product of the conditional probabilities of each node given its parents in the graph. Please note that this expression is just Equation 1 applied to this particular example network.

The assignment of values to observed variables is usually called *evidence*. In medical diagnosis the evidence is made up by the patient's antecedents, symptoms, signs, and test results. The most important type of reasoning in a probabilistic system based on Bayesian networks is known as *belief updating* or *evidence propagation*, which amounts to computing the probability distribution over the variables of interest given the evidence. In the example model of Figure 1, the variable of interest could be D and the focus of computation could be the posterior probability distribution over D given the observed values of S and T, i.e., $\Pr(d|s,t)$. In the case of influence diagrams, the focus of computation is identifying a decision option that gives the highest expected utility (possibly given some observations).

2.2 Subpopulations and Selection Variables

It is important at this point to realize that the probability of any event e, $\Pr(e)$, is expressed within some context ξ and should be formally written as $\Pr(e|\xi)$. When a model is constructed, it is by default assumed that the joint probability distribution over its variables is conditioned on some context. Since this implies that every prior, conditional, and posterior probability in that model is conditioned on this context, the context is omitted from the formulae for the sake of clarity and notational convenience.

It is useful to realize that in addition to numerical properties of the model, its structural properties are also conditional on the context. This means that the entire structure of the graph is conditioned on the context and, as a consequence, whether any two variables are dependent or independent is a property of the context in which the model was built. For instance, two diseases that are a priori independent in the general population, will be in general correlated in the context of a hospital. This phenomenon, known as Berkson bias (Berkson, 1946), is a particular case of a selection bias.

The example model in Figure 1, for example, might have been built for the general population of Pittsburgh, Pennsylvania. A model for patients of the University of Pittsburgh Medical Center might look structurally different and be quantified by a different set of numerical parameters. Conditioning has clear implications on data that we collect for different populations. And so, we will in general observe different frequencies and possibly different dependencies in data collected for the general population of Pittsburgh than we will in data collected at the University of Pittsburgh Medical Center.

In a probabilistic model, context can be expressed by means of observed (instantiated) model variables. In this way, we can build a model that is general and make it later applicable to a subpopulation by instantiating an appropriate selection variable. For example, if there is a data set collected at the hospital referred to in the example model of Figure 1, the data set will be conditioned on hospital admission H. We can indicate this by setting the variable H to $+h$ right in the model. Often such a conditioning variable is referred to as a *selection variable*, and $+h$ as the *selection value*. When a certain conditional probability differs from that of the general population, for instance, when $P(d|+h) \neq P(d)$, we say that there is a bias. We realize some obvious biases, such as those due to the fact that the data is collected at a hospital, but often forget that biases (or context variables) can be very subtle. Thus, while it is often acknowledged that post-mortem data are expected to be biased, it is often ignored that data collected from alive patients are in general biased as well.

In combining knowledge (or data) from two different sources, the most important factor are the selection variables and the corresponding biases in the knowledge (or data). The foremost question is whether these selection variables are compatible with one another.

2.3 Building Bayesian Networks

The construction of a Bayesian network consists typically of three phases: (1) selecting the model variables and their possible values, (2) determining the structure of the graph, and (3) obtaining the conditional probability distribution over each of the model variables.[1]

1. In practice, most Bayesian networks use discrete variables, due to the difficulty of eliciting and representing conditional probabilities involving arbitrary continuous variables—except for some typical distributions, such as Gaussian—and the difficulty of propagating evidence in networks with arbitrary continuous distributions.

The most straightforward way of building a Bayesian network consists of choosing a database, taking its variables as the variables of the Bayesian network, and applying some of the Bayesian network learning algorithms available in the literature (e.g., Cooper and Herskovits, 1992, Pearl and Verma, 1991, Spirtes et al., 1993). Most of these algorithms require that the database contain a sufficiently large number of cases, that there are no missing values or, if there are, that they are missing at random, that there are no selection biases, etc. However, databases almost always refer to subpopulations, whose probabilistic/statistical dependencies and independencies may be quite different from those of the target population for which the model is built. We have already mentioned in the previous section *Berkson bias* (Berkson, 1946), which is one of the many selection biases studied in the literature. A learning algorithm applied to a hospital database will most likely draw links between diseases which are probabilistically independent in the general population. Another serious problem with learning probabilistic models from data is the following. It is not uncommon that some of the configurations required for conditional probability distributions—which in the case of discrete variables are conditional probability tables (CPTs)—are only represented by a small number of cases, if any. For instance, given a symptom whose parents in the graph represent several diseases, it is quite likely that the database does not contain any patient suffering from all the diseases, and therefore it is impossible to estimate the corresponding conditional probabilities.

For these reasons, automatic learning methods alone are often insufficient in practice, and it is necessary to resort to human experts' knowledge. When a knowledge engineer relies purely on an expert, the structure of the network is determined by drawing causal links among nodes, and probabilities are obtained by subjective estimates. While building the structure of a model is in itself a challenging task that needs much care, most practitioners consider it doable. In our experience, most medical experts, for example, either give similar graphical structures or converge on the same structure after some discussion (Díez et al., 1997, Druzdzel and van der Gaag, 2000, Oniśko et al., 2001). Directed graphical models built in practice usually mimic the causal structure of a domain, which, given the fundamental role of causality in scientific understanding, explains expert agreement on the structure of models. The main drawback of this method is that sometimes there is not enough causal knowledge to establish the structure of the network with certainty. The second step in building a model based on expert opinion is quantifying the structure of a directed graphical model with numerical probabilities. Estimation of probabilities required for a typical real-world application is a tedious and time-consuming task because of the number of parameters required (typically hundreds or even thousands of numbers). Since the expert time is scarce and, therefore, costly, knowledge engineers utilize various sources of information. These may include, for example, textbooks, epidemiological studies and databases.

While the automatic-learning approach is attractive because it spares a lot of tedious elicitation effort, extracting conditional independencies from a small data set, necessary for learning the network structure, may be unreliable. A hybrid approach for the construction of Bayesian networks consists in building the structure of a causal graph with the help of human experts, who in turn rely on their experience and available literature, and combining this structural information with quantitative estimates of conditional probabilities obtained from a database. Elicitation of the structure from experts turns out to be doable in practice, even for large networks, and this approach is popular in practice. This hybrid approach is also the base of causal inference in epidemiology, in which causal graphs built from expert knowledge guide the analysis of epidemiologic data (Greenland et al., 1999, Pearl, 2000, Hernán et al., 2002).

3. Problems of Naive Use of Data

3.1 The Risk of Combining Knowledge from Different Sources

We show through an example[2] how unwary combination of knowledge from different sources may lead to severe deviations in the estimation of probability. Let us imagine that two internal medicine residents have decided to build a simple diagnostic decision support system for a certain disease D. In the first version of the system, they decided to model only D and its most important symptom S. They started by creating the model structure, consisting of two nodes, D and S (Figure 2).

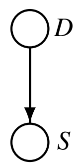

Figure 2: Example model under construction.

In the second stage, the residents focused on obtaining numerical parameters for their network. These parameters consist of $\Pr(+d)$, the prevalence of D, and the conditional probabilities of the symptom given the disease, $\Pr(+s|+d)$ and $\Pr(\neg s|\neg d)$, also known as *sensitivity* and *specificity* of the symptom, respectively. They decided to obtain these parameters from a data set of previous patient cases collected at their hospital. While there was no disagreement about sensitivity and specificity of the symptom, the residents had different opinion about the prevalence. One of them said that they need to take the prevalence as observed in their hospital, while the other suggested that they should take the prevalence for the general population, so that their system remains unbiased. After all, the second resident argued, one of the reasons why people were admitted to the hospital was because of the presence of symptom S, so if they used the hospital prevalence rate, the evidence would be double-counted. They ended up using the latter.

While the reader may disagree with the arguments made by either of the residents, it is easy to imagine obtaining the same model by the sheer fact that prevalence of a disease in the general population is often easy to find in a statistics yearbook or a morbidity table and the sensitivity and specificity may be in practice obtained from hospital records or elicited from an expert with clinical experience, i.e., one who has seen a large number of cases in clinical settings.

Let the prevalence of the disease, $\Pr(+d)$, taken from an epidemiological study performed in the town in question, be $\Pr(+d) = 0.01597$. Let the hospital data be summarized by Table 1.

N	$+d$	$\neg d$	Total
$+s$	729	63	792
$\neg s$	1	174	175
Total	730	237	967

Table 1: Distribution of the disease and test result in the hospital population.

2. We have recently discovered that a version of this example has been presented previously by Knottnerus (1987).

Sensitivity and specificity extracted from this table are

$$Sens = \Pr(+s|+d) = 729/730 = 0.99863 \tag{2}$$

$$Spec = \Pr(\neg s|\neg d) = 174/237 = 0.73418 \tag{3}$$

In our example model, the variable of interest is D and the focus of computation is the posterior probability distribution over D given an observed value of S. According to the thus constructed model, the possibility that a patient presenting with symptom S suffers from D is

$$\Pr(+d|+s) = \frac{\Pr(+s|+d)\ \Pr(+d)}{\Pr(+s|+d)\ \Pr(+d) + \Pr(+s|\neg d)\ \Pr(\neg d)} = 0.05748\ . \tag{4}$$

Leaving aside a possible error in estimating the probabilities from the database, the procedure followed seems to be correct. Nevertheless, we are going to show in the next section that this model and the posterior probability computed by it, $\Pr(+d|+s) \approx 6\%$, are incorrect.

Analysis of the problem To understand this problem, we should model the variable H, hospital admission, explicitly. Figure 3 shows a graph modeling variables D, S, and H. Admission to the hospital depends directly only on observing the symptom S, i.e., H is independent of D given S.[3] In other words, $\Pr(h|s,d) = \Pr(h|s)$. Given the symptom S, knowing whether the patient is in the hospital does not influence our belief in the presence of the disease, i.e.,

$$\Pr(d|s,h) = \Pr(d|s)\ , \tag{5}$$

which means that once we know about the presence or absence of the symptom S, the information that the patient has been admitted to the hospital does not affect the diagnosis.

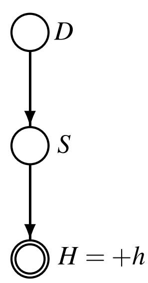

Figure 3: A causal model for the hospital data set. Note that since each entry in the data is collected for a hospital patient, the model is effectively conditioned on $H = +h$, presence in the hospital, which is the selection variable of the hospital data set. We encoded this by a double circle in the model.

The second resident in our example suggested that using the prevalence of D observed in the hospital would not be appropriate because it would double-count the evidence from observing S. In order to demonstrate that the argument behind the resident's reasoning is fallacious, we will first assume that the population of the town in question is distributed as shown in Table 2.

3. Please see Section 2.1 for a discussion of the meaning of links $D \rightarrow S$ and $S \rightarrow H$ and the relevance of the lack of link $D \rightarrow H$.

N	$+d$	$\neg d$	Total
$+s$	972	84	1,056
$\neg s$	532	92,568	93,100
Total	1,504	92,652	94,156

Table 2: Distribution of the disease and the symptom in the general population.

If a patient presenting with S is admitted to the hospital with probability $\Pr(+h|+s) = 0.75$ and a patient not presenting with S is admitted with probability $\Pr(+h|\neg s) = 1/532 = 0.00188$, the frequencies captured in the database (shown in Table 1) are consistent with the probabilistic model of the general population.

The prevalence of D is $\Pr(+d) = 0.01597$, in agreement with the result of the epidemiological study. Nevertheless, sensitivity and specificity of symptom S in the general population are

$$\begin{aligned} Sens &= \Pr(+s|+d) = 972/1,504 = 0.64628 \\ Spec &= \Pr(\neg s|\neg d) = 92,568/92,652 = 0.99909 \end{aligned}$$

which are quite different from the sensitivity and specificity among the hospital patients (see Equations 2 and 3). This difference can be attributed purely to the effect of conditioning on the patient population, i.e., looking only at those patients who are in the hospital. Given our assumption that, apart from the causal links shown in Figure 3, random variation was the only factor influencing presence or absence of the symptom and admission to the hospital, these patients may be in every respect identical to individuals in the general population. So, there are no genetic, cultural, or dietary reasons why these patients show a different sensitivity or specificity of S. We can compute $\Pr(+d|+s)$ by applying Bayes theorem or by reading the proportions in question directly from Table 2:

$$\Pr(+d|+s) = \frac{972}{1,056} = 0.92045 \; . \tag{6}$$

This result differs over an order of magnitude from the value $\Pr(+d|+s) \approx 0.06$ obtained in Equation 4. What is the explanation of this apparent paradox?

The answer is that the frequencies contained in the database do not reflect the probabilities $\Pr(d,s)$ but rather $\Pr(d,s|+h)$. For this reason, Equations 2 and 3 are wrong: they do not represent the true sensitivity and specificity, $\Pr(+s|+d)$ and $\Pr(\neg s|\neg d)$, but rather $\Pr(+s|+d,+h)$ and $\Pr(\neg s|\neg d,+h)$ respectively. A proper application of Bayes theorem is then

$$\Pr(d|s,h) = \frac{\Pr(s|d,h)\ \Pr(d|h)}{\Pr(s|+d,h)\ \Pr(+d|h) + \Pr(s|\neg d,h)\ \Pr(\neg d|h)} \; .$$

From the hospital database, we obtain the prevalence of D among the hospital patients $\Pr(+d|+h) = 730/967 = 0.75491$. Hence

$$\Pr(+d|+s,+h) = \frac{0.99863\ 0.75491}{0.99863\ 0.75491 + 0.26582\ 0.24509} = 0.92045 \; .$$

Comparing this result with Equation 6, we verify that $\Pr(+d|+s) = \Pr(+d|+s,+h)$, in agreement with Equation 5.[4]

4. We found that a similar analysis based on odds-likelihood ratios was offered by Knottnerus (1987).

In summary, when building the above model, it is correct to take the values of prevalence, sensitivity, and specificity either from the general population or from the hospital data. In both cases, the model predicts correctly the posterior probability of the disease D. But if we mix data from these two sources, the model and the resulting diagnosis can be completely wrong.

However, the next example shows that a model can be wrong even if all its conditional probabilities are obtained from the same data source.

3.2 A Wrong Subpopulation Model

Let us consider two symptoms, S_1 and S_2, of a certain disease D, such that both S_1 and S_2 may make the patient go to hospital (Figure 4).

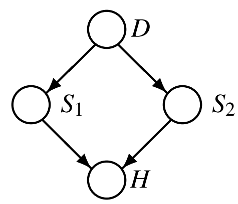

Figure 4: Example model with two symptoms.

A knowledge engineer builds a model for the hospital population with three nodes, $\{D, S_1, S_2\}$, and two causal links $\{D \to S_1, D \to S_2\}$. Variable H is implicit in this model, because all the probabilities are conditioned on $+h$. This model seems correct because the causal graph represents all the causal mechanisms between these variables and all the probabilities have been obtained from the subpopulation to which the model will be applied. However, we prove that this model is wrong by giving numerical values to the parameters and comparing its results with those of the general-population model.

We assume that $\Pr(+d) = 10^{-5}$, $\Pr(+s_i|+d) = 0.9$, $\Pr(\neg s_i|\neg d) = 0.999$, $\Pr(+h|+s_1,+s_2) = 1$, $\Pr(+h|+s_1,\neg s_2) = \Pr(+h|\neg s_1,+s_2) = 0.7$, and $\Pr(+h|\neg s_1,\neg s_2) = 0.001$. According with the causal graph given in Figure 4, the join probability is

$$\Pr(d,s_1,s_2,h) = \Pr(d)\ \Pr(s_1|d)\ \Pr(s_2|d)\ \Pr(h|s_1,s_2)\ . \tag{7}$$

In contrast, in the hospital model, whose graph does not contain H,

$$\Pr_H(d,s_1,s_2) = \Pr_H(d)\ \Pr_H(s_1|d)\ \Pr_H(s_2|d)\ . \tag{8}$$

The probabilities that the knowledge engineer obtains from hospital data are:

$$\Pr_H(+d) = \Pr(d|+h) = 0.019\ ,$$
$$\Pr_H(+s_i|+d) = \Pr(+s_i|+d,+h) = 0.927\ ,$$
$$\Pr_H(\neg s_i|\neg d) = \Pr(\neg s_i|\neg d,+h) = 0.708\ .$$

The substitution of these values into Equation 8 leads to

$$\Pr_H(+d|+s_1,+s_2) = \frac{\Pr_H(+d,+s_1,+s_2)}{\Pr_H(+d,+s_1,+s_2)+\Pr_H(+d,+s_1,+s_2)} = 0.166\ . \tag{9}$$

In contrast, the general population model (Figure 4 and Equation 7) leads to

$$\Pr(+d|+s_1,+s_2,+h) = \Pr(+d|+s_1,+s_2) = 0.976 \; . \tag{10}$$

Analysis of the problem Why is the hospital model wrong? Because the general population model states that

$$\Pr(d,s_1,s_2|+h) \neq \Pr(d|+h)\ \Pr(s_1|d,+h)\ \Pr(s_2|d,+h) \; ,$$

and for this reason Equation 8 is incorrect. Put another way, the causal graph for the hospital model correctly represents the causal mechanisms by means of links $D \to S_1$ and $D \to S_2$, but the probabilistic model corresponding to this simplified causal graph states that S_1 and S_2 are conditionally independent given D, which is false for the hospital subpopulation, because the selection value $+h$ introduces a correlation between the symptoms.

We might think that the cause of the problem is that the knowledge engineer did not include H in the causal graph for the hospital. As a remedy, we might try to use instead the graph in Figure 4, which contains H, and build a Bayesian network by adding the hospital probabilities. The join probability would then be

$$\Pr_H(d,s_1,s_2,+h) = \Pr_H(d)\ \Pr_H(s_1|d)\ \Pr_H(s_2|d)\ \Pr_H(+h|s_1,s_2) \; .$$

But when both symptoms are present $\Pr_H(+h|+s_1,+s_2) = 1$, and this leads us back to Equation 8, which gave the wrong result $\Pr_H(+d|+s_1,+s_2) = 0.166$.

It would seem that it is not possible to build a subpopulation model, not even by representing explicitly the selection variables. The remainder of this paper will show that this conclusion is false.

3.3 Discussion

In the first example we have seen that the combination of knowledge from different sources may lead to wrong results. Therefore, one rule for building Bayesian networks might be: "do not combine knowledge from different sources." Since the knowledge must be obtained from a single source, a refined version of this rule might be: "obtain all the data from the subpopulation to which the model will be applied." In fact, medical literature on the assessment of diagnostic tests often recommends to select "the adequate patient population" and to make sure that the patient belongs to the same subpopulation for which the test was assessed (see, for instance, van der Schouw et al., 1995, or Sackett et al., 1997, page 83).

However, neither of these rules is necessary or sufficient. They are not sufficient because, according to the previous example, a model may give wrong results even if the causal graph were correct and all the probabilities were obtained from the population to which the model were to be applied. They are not necessary because, as we will show in Section 4.1, in some cases it is correct to combine data from different subpopulations. An additional reason why the second rule is not necessary is that, as we saw in the example in Section 3.1, in some cases the general-population model gives the correct result for the hospital subpopulation. Furthermore, the general population model gives the correct posterior probabilities for all possible subpopulations, while subpopulation models are often wrong. This may be the case even when they are applied to the populations from which they were built—see the example in Section 3.2. As an alternative, in Section 4.2 we will offer a sufficient criterion for the correctness of a subpopulation model, and from this criterion we derive a method for building causal Bayesian networks based on selected data.

4. How to Use Subpopulation Data

The main problem faced in quantifying a Bayesian network for some general-population is that typically we do not know many required general-population parameters. We only know some probabilities corresponding to one or several subpopulations, and a key question is how these can be utilized in the construction of a model for the general population. We will focus on two important practical questions: (1) how to utilize available subpopulation data in building a causal model for the general population, and (2) how to build a model specific to a certain subpopulation characterized by a known selection variable X, assuming the selection value $X = x_s$.

The starting point of our approach is a causal graph that explicitly represents the general population and, within this graph, the selection variable for the available subpopulation. We assume that it is possible to construct such a graph based on the scientific literature and knowledge elicited from experts. We will be using this graph as a guide for determining how to combine data from different sources.

Our first result allows to identify which conditional probabilities in a certain subpopulation are unbiased, i.e., unaltered by the selection process. This allows us to estimate such parameters from the subpopulation and introduce them into the general-population model.

Our second result allows us to build a model specific for a certain subpopulation characterized by a selection value, x_s. The parameters of this model are taken from the subpopulation, even if they are different from those of the general population. However, this method requires that the graph satisfies a certain condition, namely that it is linearly ordered for X_s (see Definition 2 and Theorem 4). We also prove that it is always possible to make a graph linearly ordered for X_s by adding new links.

4.1 Introducing Subpopulation Data in a General-Population Model

Theorem 1 *Given a selection variable X_s in a Bayesian network and a node X_i (other than X_s), such that X_i is not an ancestor of X_s, the conditional probability distribution of X_i given* $\mathrm{pa}(X_i)$ *is the same in the general population and in the subpopulation induced by value x_s, i.e.,*

$$\Pr(x_i|\mathrm{pa}(x_i), x_s) = \Pr(x_i|\mathrm{pa}(x_i)) \tag{11}$$

Proof The theorem is just a special case of Markov condition, satisfied in directed probabilistic graphs: X_i is conditionally independent of its non-descendants, in particular X_s, given its parents, $\mathrm{pa}(X_i)$. ■

To illustrate this, let us consider again the example discussed in Section 2.1 (Figure 1). In this network, node T is not an ancestor of the selection variable H. According to Theorem 1, the conditional probability of T given D is the same for the general population and for the hospital subpopulation:

$$\Pr(t|d) = \Pr(t|d,h)\ . \tag{12}$$

(Please note that this equation is a particular instance of Equation 11.) As a consequence, it is possible to estimate the sensitivity and specificity of T with respect to D at the hospital and introduce them into the general-population model.

In contrast, node S is an ancestor of the selection variable H, and this implies that the sensitivity and specificity of S with respect to D will differ between the general population and the hospital subpopulation.

Finally, we would like to note that in this example the probability $\Pr(h|s)$ is irrelevant to the posterior probability distribution of D. To verify this, please note that $\Pr(d|s,h) = \Pr(d|s)$ and $\Pr(d|s,t,h) = \Pr(d|s,t)$. Both $\Pr(d|s)$ and $\Pr(d|s,t)$ can be obtained from $\Pr(d,s,t)$, which is represented in the model by the following factorization: $\Pr(d,s,t) = \Pr(d)\ \Pr(s|d)\ \Pr(t|d)$.

4.2 Building a Subpopulation Model

In the previous section, we have shown that it is possible to use subpopulation probabilities, but only when they coincide with the general population probabilities. In this section we show how to build a model based on subpopulation conditional probabilities, even if they differ from those of the general population. The criterion for testing whether it is possible is again given by a corresponding causal graph.

Definition 2 *A graph is linearly ordered for X_s iff*

$$\forall X_i, X_i \in \{X_s\} \cup \mathrm{anc}(X_s),\ \exists X_j, X_j \in \mathrm{pa}(X_i),\ \exists X_k, X_k \in \mathrm{pa}(X_i)$$
$$\Longrightarrow (X_j = X_k) \vee (X_j \in \mathrm{pa}(X_k)) \vee (X_k \in \mathrm{pa}(X_j))$$

This property can be phrased as follows: if X_s or an ancestor of X_s (say X_i) has two parents (X_j and X_k), then one of the two must be a parent of the other. Obviously, if each ancestor of X_s has only one parent, then the graph is linearly ordered for X_s. The propositions in the Appendix give more insight into the properties of linearly ordered graphs.

Definition 3 *A causal Bayesian network is linearly ordered for X_s if its graph is linearly ordered for X_s.*

Theorem 4 *Given a Bayesian network that is linearly ordered for X_s, for each configuration $\mathbf{x}_R$ of the variables in $\mathbf{X}_R = \mathbf{X} \backslash \{X_s\}$, it holds that*

$$\Pr(\mathbf{x_R}|x_s) = \prod_{i \neq s} \Pr(x_i|\mathrm{pa}(x_i), x_s)\ .$$

The proof is given in the Appendix.

For instance, the graph in Figure 3 is linearly ordered for H because the selection variable H has only two ancestors, D and S, each having only one parent. Theorem 4 asserts that it is possible to build a model whose graph is given by removing H (see Figure 2) and the conditional probabilities for D and S can be taken from the hospital subpopulation—for instance, from a database.

The causal graph in Figure 1 is also linearly ordered for H, for the same reason, and again it is possible to remove H and take the three conditional probabilities from the hospital subpopulation. Furthermore, the causal graph asserts that the conditional probability for T is the same in the general population as it is at the hospital (see the previous section, in particular Equation 12), and can be taken from any of them. In summary, in this example it is possible to estimate the conditional probabilities for D and S either from the general population or from the hospital subpopulation (but never to combine them) and in the same way it is possible to take the conditional probability for T from any of both populations, because it is the same in both.

The reader might ask: "Why is it not possible to apply this method when the graph is not linearly ordered for X_s?" The answer is that if the graph is not linearly ordered for X_s, then X_s will have at least two ancestors X_i and X_j in $\mathbf{X}$ such that neither one is an ancestor of the other. It means that even if X_i and X_j are (a priori or conditionally) independent in the general-population model, this independence is lost in the subpopulation—this phenomenon is known as Berkson bias (Berkson, 1946). The example in Section 3.2 showed that the introduction of subpopulation data in a graph that is not linearly ordered for the selection variable H led to wrong computations of probability.

However, it is always possible to make a graph linearly ordered for X_s by applying the following algorithm:

1. make an ordered list of $\mathbf{X}$ such that $\forall i$, $\mathrm{pa}(X_i) \subseteq \{X_1, \ldots, X_i\}$;[5]
2. $A \leftarrow X_s$;
3. **while** A has parents
   (a) $B \leftarrow$ last node in $\mathrm{pa}(A)$, according to the list created in step 1;
   (b) for each node C in $\mathrm{pa}(A)\backslash\{B\}$,
   if link $C \rightarrow A$ is not in the graph, add it;
   (c) $A \leftarrow B$

   end while

Please note that the new graph is acyclic (because it does not introduce any link $X_i \rightarrow X_j$ with $i > j$) and the ancestors of X_s in the new graph are the same as those in the original graph (because the algorithm only introduces links between the ancestors of X_s).

As an example, the graph in Figure 4 can be made linearly ordered for H by building an ordered list of nodes $\{D, S_1, S_2, H\}$. Since $\mathrm{pa}(H) = \{S_1, S_2\}$ and S_2 is the last node in $\mathrm{pa}(H)$, the algorithm draws a link $S_1 \rightarrow S_2$. The hospital model has then three nodes, $\{D, S_1, S_2\}$, and three links $\{D \rightarrow S_1, D \rightarrow S_2, S_1 \rightarrow S_2\}$. The conditional probabilities $\Pr(d)$, $\Pr(s_1|d)$, and $\Pr(s_2|d, s_1)$ can be estimated from the hospital data, because Theorem 4 guarantees that

$$\Pr(d, s_1, s_2| + h) = \Pr(d| + h)\ \Pr(s_1|d, +h)\ \Pr(s_2|d, s_1, +h)\ .$$

In particular, this model yields the correct value of $\Pr(+d| + s_1, +s_2, +h)$:

$$\begin{aligned}\Pr(+d| + s_1, +s_2, +h) &= \frac{\Pr(+d, +s_1, +s_2| + h)}{\Pr(+s_1, +s_2| + h)} \\ &= \frac{\Pr(+d, +s_1, +s_2| + h)}{\Pr(+d, +s_1, +s_2| + h) + \Pr(\neg d, +s_1, +s_2| + h)} = 0.976\ .\end{aligned}$$

Please compare it with Equations 9 and 10.

The general method for building a subpopulation model works as follows:

1. Build a causal graph that includes the selection variable X_s;

5. Condition $\mathrm{pa}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ means that the parents of X_i must be numbered before X_i. The list can be built by recursively removing a node without parents from the graph and putting it at the end of the list.

2. Make the graph linearly ordered for X_s;

3. Remove X_s and its links from the graph;

4. Estimate the CPTs from subpopulation data.

Theorem 4 guarantees that the probabilities computed from this model will be the same as those we would obtain from the general population model.

One limitation of this method is that the addition of new links increases the size of the CPTs. When the probabilities are obtained from a database, the reduction in the number of database cases available for estimating each parameter endangers the accuracy of the model. When the probabilities are obtained from subjective estimates, the difficulty is even bigger. For instance, in the above example it would virtually impossible for a human expert to give different values for $\Pr(s_2|d,+s_1,+h)$ and $\Pr(s_2|d,\neg s_1,+h)$, given that S_1 and S_2 are conditionally independent given D, and the impact of S_1 on S_2 is due to the conditioning on H. It would even be difficult to say whether $\Pr(s_2|d,+s_1,+h)$ is bigger or smaller than $\Pr(s_2|d,\neg s_1,+h)$, let alone to assign numerical values to these parameters.

A minor problem is the fact that the links introduced by the above algorithm are not causal, and this might complicate the process of generating explanations for the user (Lacave and Díez, 2002). When explaining the model to the user, it would be necessary to differentiate causal links, which represent dependencies induced by causal influences, from the links added in order to make the graph linearly ordered, which represent dependencies induced by selection mechanisms.

5. Conclusions

Knowledge engineers quantifying probabilistic models usually combine various sources of information, such as existing textbooks, statistical reports, databases, and expert judgement. However, lack of attention to whether the sources are compatible and whether they can be combined may lead to erroneous behavior of the model. For instance, an unwary knowledge engineer might combine the prevalence of a certain disease, obtained from a general-population study, with the sensitivity and specificity of a certain test obtained at hospital. This combination of information may lead to a several orders of magnitude error in the computation of the posterior probabilities of interest. While the reader might think that no experienced knowledge engineer would make such a mistake, the fact that sensitivity and specificity may be biased when obtained from a subpopulation has never been mentioned in Bayesian network literature. Even in medical literature, it is not uncommon to find values of sensitivity and specificity without an explanation of how they were obtained, because they are assumed to be invariant. After all, sensitivity and specificity do not depend on the prevalence. We realize that different population characteristics, such as sex, race, diet, etc., can influence both sensitivity and specificity, but we forget about purely statistical phenomena such as conditioning. Please note that in our first motivating example (Section 3.1), the population in the hospital consisted of identical individuals for all that matters. It was not the special characteristics of the hospital patients that made them develop the symptom more or less likely than the general population. Biases related to sensitivity and specificity of medical tests have been reported in epidemiological literature over the last two decades, although without the framework of directed probabilistic graphs the descriptions rely mainly on contingency tables and are somewhat obscure (Díez et al., 2003).

Our motivating examples were based on a medical data set, but the same argument can be made with respect to numbers obtained from human experts. Subjective probability judgments have been shown to rely on judgmental heuristics (Kahneman et al., 1982) and they are very sensitive to prior experiences (in fact prior experiences are often all that probability judgments are based on). Humans have been shown to be able to match the probability of observed events with an amazing precision in certain experiments (Estes, 1976). Physicians working in a hospital will tend to match the sensitivity and specificity of medical symptoms and tests that they observe in their practice. These are often determined by the circumstances, such as what brought the patients to the hospital or clinic in the first place. Physician experts will tend to at least adjust the parameters to what they observe in their practice. While their experience is valuable for building decision models for the particular clinics where they have worked, in general they cannot be readily used in other settings. Similarly, one cannot assume that this knowledge can be combined with data originating from other settings.

In our examples, the variable that led to selection biases was the fact that the patient was admitted to the hospital. There is a plenitude of other variables that might lead to a similar bias. In some cases, for example, a possible bias variable may be the very fact that a patient is alive. An important conclusion of our paper is that, contrary to the usual practice in knowledge engineering, such variables may not be ignored.

The assertion that combining data from two different sources is dangerous may seem trivial. However, to our knowledge, the literature on building probabilistic expert systems has never mentioned the risk of combining knowledge from textbooks, databases, and human experts. In fact, it has been known that conditioning affects qualitative, structural properties of models, such as probabilistic independence, but to our knowledge no attention has been paid to its impact on such seemingly robust local properties as conditional probabilities.

On the other hand, an over-cautious position of never combining numerical data obtained from different sources would result in disregarding valuable information, which might be useful in model construction. In fact, we have shown in Section 3.3 that the criteria "do not combine knowledge from different sources" and "obtain all the data from the subpopulation in which the model will be applied" are neither necessary nor sufficient to guarantee the correctness of the model. For this reason, we have introduced a criterion for combining data from different sources, namely that the causal graph, built from expert knowledge, is linearly ordered (see Definition 2). We have also offered an algorithm for making the graph linearly ordered by adding links that represent the probabilistic dependencies induced by selection mechanisms. Knowledge engineers must not ignore this property, because the absence of those links may lead to important errors in the computation of the probabilities, even when all the probabilities were obtained from the subpopulation in which the model is applied (cf. Section 3.2).

In summary, there are two ways to build a probabilistic model based on a causal graph. The first one is to build a *general-population model*, in which some parameters may be taken from certain subpopulations only if the probabilities are not biased by selection mechanisms (cf. Section 4.1). The main difficulty of this approach is to make the experts estimate general-population probabilities, since, in general, human expertise is based on selected populations. The second approach is to build a *subpopulation model* in which the selection variables have been replaced by non-causal links (cf. Section 4.2). However, we argued that subjective estimates are even more difficult in this case, and for this reason we recommend knowledge engineers to build general-population models.

Although the main focus or our paper is knowledge engineering, it may be shed light on other fields. From the point of view of machine learning, it emphasizes the importance of selection biases in the automatic construction of causal models from databases. It can also be useful when one or several agents look for information (for instance, by searching the Internet) and try to build a model by combining information extracted from several sources. In this scenario, the agent should use qualitative knowledge as a guide for combining numerical data. A particular case of this scheme would be the development of a tool for automated elicitation of knowledge through interaction with human experts, similar to those that exist for building rule-based expert systems. Finally, from the point of view of statistics, this paper can be useful for the application of causal models in epidemiology (Greenland et al., 1999, Pearl, 2000, Hernán et al., 2002), in which the analysis of data (in general, selected data) is based on a causal graph built from expert knowledge. Our analysis might also be applied to meta-analysis,[6] because both the data of each study and the collection of studies are prone to selection biases (see, for instance, Macaskill et al., 2001).

Acknowledgments

The first author was supported in part by the National Science Foundation under Faculty Early Career Development (CAREER) Program, grant IRI–9624629, and by the Air Force Office of Scientific Research under grant F49620–00–1–0112. The second author was supported by the Spanish CICYT, under grants TIC97–1135–C04 and TIC–2001–2973–C05. Our collaboration was enhanced by travel support from NATO Collaborative Linkage Grant number PST.CLG.976167. We would like to thank the reviewers and the participants of the *UAI–2000 Workshop on Fusion of Domain Knowledge with Data for Decision Support*, where we presented our initial ideas on what is the foundation of the current paper, for their comments, which improved the clarity of our presentation. Additional improvements in readability were prompted by the reviewers of the current journal.

Appendix. Proof of Theorem 4

Proposition 5 *Let* $\Pr(\mathbf{x})$ *be a probability distribution defined on a set of variables* $\mathbf{X} = \{X_1, \ldots, X_n\}$ *and* $\mathbf{x}_R$ *a configuration of the variables in* $\mathbf{X}_R = \mathbf{X} \backslash \{X_s\}$. *Then*

$$\Pr(\mathbf{x}_R | x_s) = \prod_{i \neq s} \Pr(x_i | x_1, \ldots, x_{i-1}, x_s) \, .$$

Proof We have

$$\Pr(\mathbf{x}) = \prod_{i=n}^{s+1} \Pr(x_i | x_1, \ldots, x_{i-1}) \prod_{i=s-1}^{1} \Pr(x_i | x_1, \ldots, x_{i-1}, x_s) \; \Pr(x_s) \, .$$

Please note that $s+1 \leq i \leq n$ implies that $X_s \in \{X_1, \ldots, X_{i-1}\}$, and then

$$\Pr(x_i | x_1, \ldots, x_{i-1}) = \Pr(x_i | x_1, \ldots, x_{i-1}, x_s) \, ,$$

which proves the proposition. ■

Proposition 6 *If a graph is linearly ordered for* X_s *and* X'_s *is an ancestor of* X_s, *then the graph is linearly ordered for* X'_s.

6. Meta-analysis, a technique that has become popular in the last years, especially in medicine, consists in extracting data from different epidemiological studies published in the literature and combining them in order to draw more reliable or more precise conclusions.

Proof It follows from Definition 2, because if a node X_i is an ancestor of X_s' then it is also an ancestor of X_s. ■

Proposition 7 *If a graph is linearly ordered for X_s and X_s has at least one parent, then there is a node* $\mathrm{pa}_L(X_s) \in \mathrm{pa}(X_s)$ *such that all the other parents of X_s are also parents of* $\mathrm{pa}_L(X_s)$.

Proof We can set a total ordering in $\mathrm{pa}(X_s)$ by this definition: $X_i < X_j$ iff there is a link $X_i \rightarrow X_j$. (The absence of cycles in the graph guarantees transitivity. It is a total order because there is a link for each pair of parents.) The last node in this ordering is $\mathrm{pa}_L(X_s)$. ■

Corollary 8 *If a graph is linearly ordered for X_s and X_s has at least one parent, then there exists a node* $\mathrm{pa}_L(X_s) \in \mathrm{pa}(X_s)$ *such that*

$$\mathrm{pa}(X_s) = \{\mathrm{pa}_L(X_s)\} \cup \mathrm{pa}(\mathrm{pa}_L(X_s))$$

and

$$\mathrm{anc}(X_s) = \{\mathrm{pa}_L(X_s)\} \cup \mathrm{anc}(\mathrm{pa}_L(X_s)) \tag{13}$$

where $\mathrm{anc}(X_i)$ *is the set of ancestors of X_i.*

Corollary 9 *If a graph is linearly ordered for X_s and X_i is a parent of X_s, then there exists a unique chain of nodes,* $\mathrm{chain}(X_i, X_s) = \{Y_1, \ldots, Y_m\}$, *such that* $\mathrm{pa}_L(X_s) = Y_m$, $\mathrm{pa}_L(Y_m) = Y_{m-1}$, $\ldots$, $\mathrm{pa}_L(Y_1) = X_i$. *Additionally, each parent of X_s other than X_i is either in* $\mathrm{chain}(X_i, X_s)$ *or in* $\mathrm{pa}(X_i)$*:*

$$\mathrm{pa}(X_s) = \mathrm{chain}(X_i, X_s) \cup \{X_i\} \cup \mathrm{pa}(X_i) \; .$$

Please note that $\mathrm{chain}(X_i, X_s) = \emptyset$ if and only if $X_i = \mathrm{pa}_L(X_s)$.

The following proposition generalizes this corollary to the case in which X_i is an ancestor of X_s (not necessarily a parent):

Proposition 10 *If a graph is linearly ordered for X_s and X_i is an ancestor of X_s, then there exists a unique chain of nodes,* $\mathrm{chain}(X_i, X_s) = \{Y_1, \ldots, Y_m\}$, *such that* $\mathrm{pa}_L(X_s) = Y_m$, $\mathrm{pa}_L(Y_m) = Y_{m-1}$, $\ldots$, $\mathrm{pa}_L(Y_1) = X_i$. *Additionally, each ancestor of X_s other than X_i is either in* $\mathrm{chain}(X_i, X_s)$ *or in* $\mathrm{anc}(X_i)$*:*

$$\mathrm{anc}(X_s) = \mathrm{chain}(X_i, X_s) \cup \{X_i\} \cup \mathrm{anc}(X_i) \; . \tag{14}$$

Proof If X_i is an ancestor of X_s, then either $X_i \in \mathrm{pa}(X_s)$—and Corollary 9 applies—or there exists a chain $\{Z_1, \ldots, Z_p\}$ such that $X_i \in \mathrm{pa}(Z_1), Z_1 \in \mathrm{pa}(Z_2), \ldots, Z_p \in \mathrm{pa}(X_s)$. Then,

$$\mathrm{chain}(X_i, X_s) = \mathrm{chain}(X_i, Z_1) \cup \{Z_1\} \cup \mathrm{chain}(X_i, Z_2) \cup \ldots \cup \{Z_p\} \cup \mathrm{chain}(Z_p, X_s) \; .$$

Equation 14 follows from the recursive application of Equation 13 to the nodes in $\mathrm{chain}(X_i, X_s)$, from X_s to X_i. ■

Proposition 11 *Let us have a Bayesian network linearly ordered for X_s, such that, for all j, $\mathrm{pa}(X_j) \subseteq \{X_1,\ldots,X_{j-1}\}$. If X_i is an ancestor of X_s, then*

$$\Pr(x_s|x_1,\ldots,x_i) = \Pr(x_s|x_i,\mathrm{pa}(x_i)) \; .$$

Proof Given chain(X_i,X_s) (cf. Proposition 10),

$$\begin{aligned}
&\Pr(x_s,y_1,\ldots,y_m,x_1,\ldots,x_i) = \\
&\qquad Pr(x_s|y_1,\ldots,y_m,x_1,\ldots,x_i) \prod_{j=1}^{m} \Pr(y_j|y_1,\ldots,y_{j-1},x_1,\ldots,x_i) \; \Pr(x_1,\ldots,x_i) \; .
\end{aligned}$$

Since $\mathrm{pa}(X_s) \subseteq \mathrm{anc}(X_s) \subseteq \{Y_1,\ldots,Y_m,X_1,\ldots,X_i\}$ (see Equation 14), the Markov property implies that

$$\Pr(x_s|y_1,\ldots,y_m,x_1,\ldots,x_i) = \Pr(x_s|\mathrm{pa}(x_s)) = \Pr(x_s|y_m,\mathrm{pa}(y_m)) \; .$$

For the same reason, the fact that the graph is linearly ordered for all Y_j implies that

$$1 \leq j < m, \; \Pr(y_j|y_1,\ldots,y_{j-1},x_1,\ldots,x_i) = \Pr(y_j|\mathrm{pa}(y_j)) = \Pr(y_j|y_{j-1},\mathrm{pa}(y_{j-1}))$$

and

$$\Pr(y_1|x_1,\ldots,x_i) = \Pr(y_j|\mathrm{pa}(y_j)) = \Pr(y_1|x_i,\mathrm{pa}(x_i)) \; .$$

Therefore

$$\begin{aligned}
\Pr(x_s|x_1,\ldots,x_i) &= [\Pr(x_1,\ldots,x_i)]^{-1} \sum_{y_1}\cdots\sum_{y_m} \Pr(x_s,y_1,\ldots,y_m,x_1,\ldots,x_i) \\
&= \sum_{y_1}\cdots\sum_{y_m} \Pr(x_s|y_m,\mathrm{pa}(y_m)) \; \prod_{j=1}^{m} \Pr(y_j|\mathrm{pa}(y_j)) \; .
\end{aligned}$$

In this equation, we apply recursively the properties that

$$1 \leq j \leq m, \; \sum_{y_j} \Pr(x_s|y_j,\mathrm{pa}(y_j)) \; \Pr(y_j|\mathrm{pa}(y_j)) = \Pr(x_s|\mathrm{pa}(y_j))$$

and

$$1 < j \leq m, \; \Pr(x_s|\mathrm{pa}(y_j)) = \Pr(x_s|y_{j-1},\mathrm{pa}(y_{j-1})) \; ,$$

in order to arrive at

$$\Pr(x_s|x_1,\ldots,x_i) = \Pr(x_s|\mathrm{pa}(y_1)) = \Pr(x_s|x_i,\mathrm{pa}(x_i)) \; .$$

■

Proposition 12 *Let us have a Bayesian network linearly ordered for X_s, such that, for all j, $\mathrm{pa}(X_j) \subseteq \{X_1,\ldots,X_{j-1}\}$. For each node X_i in $\mathbf{X}_R = \mathbf{X}\backslash\{X_s\}$,*

$$\Pr(x_i|x_1,\ldots,x_{i-1},x_s) = \Pr(x_i|\mathrm{pa}(x_i),x_s) \; .$$

Proof If X_i is not an ancestor of X_s, then

$$\Pr(x_i|x_1,\ldots,x_{i-1},x_s) = \Pr(x_i|\text{pa}(x_i)) = \Pr(x_i|\text{pa}(x_i),x_s)$$

Otherwise,

$$\Pr(x_i|x_1,\ldots,x_{i-1},x_s) = \frac{\Pr(x_i|x_1,\ldots,x_{i-1})\ \Pr(x_s|x_1,\ldots,x_i)}{\Pr(x_s|x_1,\ldots,x_{i-1})} .$$

We have that

$$\Pr(x_i|x_1,\ldots,x_{i-1}) = \Pr(x_i|\text{pa}(x_i)) .$$

From Proposition 11,

$$\Pr(x_s|x_1,\ldots,x_i) = \Pr(x_s|x_i,\text{pa}(x_i))$$

and

$$\begin{aligned}\Pr(x_s|x_1,\ldots,x_{i-1}) &= \sum_{x_i}\Pr(x_s|x_1,\ldots,x_i)\ \Pr(x_i|x_1,\ldots,x_{i-1})\\ &= \sum_{x_i}\Pr(x_s|x_i,\text{pa}(x_i))\ \Pr(x_i|\text{pa}(x_i))\\ &= \Pr(x_s|\text{pa}(x_i)) .\end{aligned}$$

Therefore,

$$\Pr(x_i|x_1,\ldots,x_{i-1},x_s) = \frac{\Pr(x_i|\text{pa}(x_i))\ \Pr(x_s|x_i,\text{pa}(x_i))}{\Pr(x_s|\text{pa}(x_i))} = \Pr(x_i|\text{pa}(x_i),x_s) .$$

■

We are now ready to prove the theorem.

Proof [Theorem 4] It is always possible to re-label the nodes in $\mathbf{X}$ in such a way that the parents of a node are numbered before that node: $\text{pa}(X_i) \subseteq \{X_1,\ldots,X_{i-1}\}$. From Propositions 5 and 12, we have

$$\Pr(\mathbf{x}_R|x_s) = \prod_{i=1}^{n}\Pr(x_i|x_1,\ldots,x_{i-1},x_s) = \prod_{i=1}^{n}\Pr(x_i|\text{pa}(x_i),x_s) .$$

■

References

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*, 2: 47–53, 1946.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

F. J. Díez, M. Druzdzel, and M. A. Hernán. Biases in the assessment of diagnostic indicants. Technical Report, UNED, Madrid, 2003. In preparation.

F. J. Díez, J. Mira, E. Iturralde, and S. Zubillaga. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10:59–73, 1997.

M. J. Druzdzel and L. C. van der Gaag. Building probabilistic networks: "Where do the numbers come from?" Guest editors' introduction. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):481–486, July–August 2000.

W. K. Estes. The cognitive side of probability learning. *Psychological Review*, 83(1):37–64, January 1976.

S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48, 1999.

M. Henrion, J. S. Breese, and E. J. Horvitz. Decision Analysis and Expert Systems. *AI Magazine*, 12(4):64–91, Winter 1991.

M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology*, 155:176–184, 2002.

R. A. Howard and J. E. Matheson. Influence diagrams. In Ronald A. Howard and James E. Matheson, editors, *The Principles and Applications of Decision Analysis*, pages 719–762. Strategic Decisions Group, Menlo Park, CA, 1984.

D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, 1982.

J. A. Knottnerus. The effects of disease verification and referral on the relationship between symptoms and diseases. *Medical Decision Making*, 7:139–148, 1987.

C. Lacave and F. J. Díez. A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, 17:107–127, 2002.

P. Macaskill, S. D. Walter, and L. Irwig. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20:641–654, 2001.

A. Oniśko, M. J. Druzdzel, and H. Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 27 (2):165–182, 2001.

J. Pearl. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

J. Pearl and T. S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, Cambridge, MA, 1991. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

D. F. Ransohoff and A. R. Feinstein. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, 299:926–930, 1978.

D. L. Sackett, W. S. Richardson, W. Rosenberg, and R. B. Haynes. *Evidence-Based Medicine. How to Practice and Teach EBM*. Churchill Livingstone, New York, 1997.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 1993.

Y. T. van der Schouw, R. van Dijk, and A. L. M. Verbeek. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *Journal of Clinical Epidemiology*, 48:417–422, 1995.