UNIVERSITÀ DEGLI STUDI DI FIRENZE

Dipartimento di Statistica 'G. Parenti'

Tesi di Dottorato in Statistica Applicata XVII Ciclo
di
Gianluca Baio

DECISION THEORETIC APPROACH TO CAUSAL INFERENCE
THEORY AND APPLICATIONS

Relatore: Clar.mo Prof. Fabio Corradi
Correlatore: Clar.mo Prof. Fabio Pammolli
Coordinatore: Clar.mo Prof. Fabrizia Mealli

Gennaio 2005

CONTENTS

PREFACE

The evaluation of causal mechanisms has always represented a relevant theoretical and applied research area in many disciplines, from Philosophy to Epidemiology, Economics and Econometrics, Law, Engineering and of course Statistics and Probability.

The objective of this work is to provide a brief review of some of the most relevant approaches to statistical analysis of causality, and to propose some methodological advances.

The approach that we decided to follow is the *decision theoretic*. Moreover, the statistical tool that we use is that of Bayesian Networks. The combination of these two instruments has been widely treated in the literature (Heckerman & Shachter 1994, Dawid 2002), although there seems that not many practical applied works have actually used it.

We propose two main streams of research: on the one hand, we work in a standard causal inference framework, where the typical objective is to estimate the causal effect of a treatment over a suitable response, discarding all the possible identifiable confounding factors, that can screen off such effect.

In this area, Dawid (2002) proposed a model based on a particular specification of graphical models that allows the representation and the estimation of problems where the objective is to use past information in order to estimate the future impact of a present decision (intervention). This kind of situation has been termed *Effects of Causes* problem (Holland 1986, Dawid 2000).

Our contribution is to try to 'expand the market' for this idea, building a methodology that aims at a twofold objective. First we propose an algorithm that, provided some necessary assumptions, can build a causal structure for which the standard causal constraints hold, which can help an experimenter structure a particular problem, starting from a comprehensive database.

Second, we define a procedure that automatically selects the admissible sets of *sufficient covariates*, which need be known, in order to correctly estimate the causal effect.

The typical application of this tool could be in clinical practice, aiming

at providing an estimation of the causal mechanism that relates the administration of a given treatment, a suitable clinical response, and a set of covariates of interest for the problem.

When a new patient comes into the practice, then the physician would be equipped with a computer assisted procedure that can suggest which variables are essential in order to assess the impact of the treatments, specifically for that individual.

We provide examples to show how two patients with different initial conditions can be associated to different sets of sufficient covariates that need be known, in order to find the best individual treatment. This work is still in progress, and we provide reference of what we reckon are the main problems to be solved, before an efficient algorithm can be programmed.

On the other hand, we tackle the causal problem from a rather different point of view: the idea is to utilise the standard representation based on the decision theoretic approach in order to evaluate situations in which some of the pieces of available evidence have not a clear origin.

The typical use of a Bayesian Network model is to make inference about some unobservable variables, using some pieces of related evidence. The underlying assumption is that the data generating process is known and static. However, if some of the available evidence is externally manipulated:

- On purpose, in order to mislead the inference on the unobservable variables;
- Because of non controlled (detected) measurement errors;
- Because of a mutation of the causal mechanisms that govern the experiment under study,

then the deriving inference would be flawed.

The objective of our model is to construct a method that can help the experimenter handle possibly manipulated evidence, in order to produce a correct inference. This goal will be pursued by means of some pieces of *control evidence*, suitably defined, that allow the probabilistic evaluation of the data generating process, provided some opportune underlying assumptions.

Some examples of research areas that can be interested in such a model are police investigation, Economics & anti-trust legislation, insurance activity and Genetics. However, in our views, even standard statistical methodologies can be a field for this procedure, in the analysis of conflicting evidence, and in outliers detection.

Finally, we present an application of a Bayesian Network model for Health Technology Assessment, with particular reference to the economic evaluation of influenza vaccination. Although from a further different perspective, this last work shares some common points with the ones described above.

In fact, the framework is explicitly that of decision theoretic analysis. In this case, the decision maker has to choose whether to vaccine a reference population or not, and this decision is evaluated with respect to both its clinical and economic implications.

However, in this case causal inference is not the main objective, although from the technical point of view, some common features can be identified with the models presented above, for example, the presence of 'switch' nodes.

ACKNOWLEDGEMENTS

At this stage, I would like to thank many people, that contributed in several, different ways to this work. First, I must mention Prof. Fabio Corradi, who supervised this PhD thesis, and gave his constant help and encouragement and suggestions. It has been a pleasure to work with him, and to improve our joint knowledge of this subtle subject, during the last two years.

Second, I must acknowledge Prof. Fabio Pammolli, the first reader of this thesis, for the help that he provided in focusing on substantial aspects of the research and for pointing out that our work could have some applications in areas that we were not aware of. Moreover, I have to thank Prof. Pammolli for giving me the possibility to visit the Harvard–MIT Division of Health, Science & Technology.

With respect to this matter, I must also thank Prof. Stan Finkelstein and his assistant Joanne McHughes, and Prof. Marco Ramoni, with whom I have been working in Boston. It was brilliant to experience such a prestigious University.

To the same extent, I have to thank Prof. Phil Dawid of University College, London, for his comments, and his interaction. Meeting him personally and having the chance to discuss my views on Causal Inference with him has been a great honour and a source of inspiration for me.

A special thank goes to Mr. Abercrombie and Mr. Fitch, back in Columbus, Ohio, and to Mr. and Mrs. Huxtable in New York, NY, and finally, I would like most to thank Marta for her love, understanding and support.

December, 2004

Gianluca Baio

CHAPTER 1

AN INTRODUCTION TO THE THEORY OF CAUSAL MODELLING

– *"So... you ever wondered which is worst? You know... going through labour or getting kicked in the nuts?*

– *What?!*

– *Oh, that's interesting! Because no one'll never know, because no one can experience both! One of life's great unanswerable questions!"*

Chandler Bing in "Friends". The last episode

1.1 Introduction

Probabilistic causation designates a group of philosophical theories that aim at characterising the relationship between *cause* and *effect* using the tools of probability theory.

Two ideas appeared to be central behind these theories: first, the assumption that causes raise the chance of occurrence of their effects, all else being equal. According to David Hume (1748, section VII), causes are invariably followed by their effects:

> *We may define a cause to be an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second.*

A great deal of the work that has been done in this area has been concerned with making the *ceteris paribus* clause more precise.

Second, the definition of a strict link between *causation* and *manipulation*. Following this approach, causes are to be regarded as handles or devices for manipulating effects (von Wright 1973). This theory has produced several links between philosophers, statisticians and econometricians: for instance, Holland (1986) reports the motto *'No causation without manipulation'* as representative of the work of Donald Rubin and himself.

Bearing these two central ideas in mind, the objective of this paper is to review some of the main features of causal modelling, with specific focus

towards their statistical implications and applications. In section 2, we describe the major framework of causal inference that we refer to, whereas in section 3 we give account of what we reckon are the most relevant approaches to Causality, and discuss the major points of accordance or disagreement among them. Section 4 presents two of the most important criteria defined to allow the estimation of causal effects, and finally some conclusions are drawn in section 5.

1.2 Causal Inference

According to the manipulation approach, we define causal inference as the process of evaluation of an external intervention on one or more variables within a stochastic system.

This feature characterises causal analysis and produces a huge difference with standard statistical modelling: in fact, on the one hand, Statistics is typically concerned with the identification of a probabilistic structure of *association* among some variables. On the other hand, causal inference is mainly focused on the estimation of the structure of *causal* relations among the variables within a given stochastic system.

Consequently, a basic difference is that, while Statistics tend to estimate some quantities of interest, under the assumption that the data generating process remains the same, Causality is rather aimed at evaluating dynamically the process that made the observed data arose.

Formal attempts have been made to describe this situation using the tools of probability; nevertheless, it is now well accepted that standard statistical methodologies are not reliable for causal inference since they are typically based on a conditional probability of observing a variable Y, after *observing* a variable X, $\Pr(Y \,|\, X = x)$.

Conversely, causal modelling should be concerned with a different quantity, i.e. the conditional probability of observing Y, after *setting* the value of X, $\Pr(Y \,||\, X = x)$, using the notation introduced by Lauritzen (2000)[1]. Since, in general, these two distributions are different, in order to assess the latter using the former, it is necessary to make some suitable assumptions.

In fact, the objective of most causal analysis is that of estimating the

[1]Other notations are used in the literature: for instance, Pearl (1993) refers to the intervention distribution as $\Pr(Y \mid set(X = x))$ or $\Pr(Y \mid do(X = x))$, whereas Pearl (1995) describes it as $\Pr(Y \,|\, X = \check{x})$. However, the notation of Lauritzen (2000) seems more straightforward to us, as it clearly establish the difference with standard statistical conditioning. Therefore, we will use it through all this work

effect of a variable on an other within an observational setting. Rosenbaum (1995) defines an observational study as:

> *... an empirical investigation of treatments, policies, or exposures and the effect they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects.*

Consequently, the study of causal mechanism from observational studies relies on the additional investigation of the *potential confounders*, i.e. all those variables that can 'screen off' (Reichenbach 1956) the desired causal effect of a variable on a suitable response.

Moreover, it is possible to define two different frameworks for causal inference (Holland 1986, Dawid 2000):

- *Effects of causes* (EoC);
- *Causes of effects* (CoE).

The main difference stems in the dynamics associated to the causal analysis. In the EoC problem, starting from the observation of a set of variables, the objective is to analyse how an external (present) intervention might modify a suitable (future) response. As for the CoE problem, the typical framework involves the identification of a (past) casual process that led to the (present) observation of a given response variable.

As suggested by Holland (1986) and Dawid (2000), the EoC problem seems to be more specific to Statistics, dealing with some measurements used to forecast an unknown quantity. On the other hand, the CoE problem is related to a set of further assumptions on the actual causal mechanism associated to the system under study, which do not involve statistical considerations only.

In our opinion, this argument is highly valuable, and should be taken into account carefully, when dealing with statistical analysis of Causality.

1.3 Statistical approaches to causal modelling

1.3.1 Potential outcomes and counterfactual analysis

In the last three decades, the first major contributions to the statistical analysis of causation have perhaps been those of Rubin (1974, 1978). In his framework, the impact of a treatment variable on a suitable response is measured in terms of *potential outcomes*.

Assuming for the sake of simplicity that the treatment is binary (t, c), the causal effect is typically evaluated in terms of the difference between the response $Y_t(i)$ that *has actually been observed* on the $i-th$ individual for the assigned treatment t, and the response $Y_c(i)$ that *would have been observed*, should the treatment selected be c.

Since for an individual it is possible to be assigned only to one treatment at a given time, one of the two responses is a *counterfactual* quantity. Rubin's model considers the joint distribution of the potential outcomes (Y_t, Y_c), $p = \Pr(Y_t, Y_c)$, and for each individual the counterfactual response is regarded as a missing value, to be estimated, possibly with respect to a set of suitable covariates.

Most of the criticism to this approach concerns the fact the joint distribution of the potential outcomes, involving counterfactuals, is against De Finetti's observability principle, which dictates that probability statements are possible only on quantities that are *at least in principle* observable. By its definition, a counterfactual does not verify this condition, after the assignment of the treatment. Consequently, the potential outcomes approach is charged of focusing on a 'metaphysical', non scientific quantity (Shafer 1996, Dawid 2000).

However, other authors such as Greenland (2004) regards counterfactuals as a natural way of defining causal mechanisms, or rather as a key aspect in causation, and hence produce causal evaluations by means of this theoretical construction.

The evaluation of causal effect is typically based on the difference:

$$\tau = Y_t - Y_c. \tag{1.1}$$

Nevertheless, it can be shown that the *Individual Causal Effect*

$$ICE(i) = \tau(i) = Y_t(i) - Y_c(i) = f(p)$$

being a function of the joint probability distribution of the potential outcomes is not directly identifiable, unless further restrictions are assumed.

On the contrary, under (1.1), the *Average Causal Effect*:

$$ACE = \mathrm{E}_p[ICE(i)] = \mathrm{E}_p[Y_t(i) - Y_c(i)].$$

is completely identifiable, as:

$$\mathrm{E}_p[Y_t(i) - Y_c(i)] = \mathrm{E}[Y_t(i)] - \mathrm{E}[Y_c(i)],$$

i.e. it is a function of the marginal distributions only, which are estimable from observed data.

Obviously, the linearity of (1.1) is crucial: for instance, in case the individual effect is evaluated by means of a non linear function, then even the average effect might turn to be not identifiable. Yet, for fairly general situations, plausible ranges have been identified (Balke & Pearl 1994, Dawid 2000) that allow the estimation of the effect.

As reported before, the situation changes, at least from the practical point of view, when some further assumptions are sustainable. For example, in case that the reference population is *homogenous*, each unit i is essentially identical to the other individuals. Hence, although on different units, it is possible to observe both $Y_t(i)$ and $Y_c(i)$, so that both ICE and ACE are identifiable from empirical data.

This situation highly resembles that of randomised trials, where all the potential confounding factors are randomised within the individuals, so that it is possible to assume that they are homogenous, and that the differences in the outcome are actually attributable to the treatment assigned.

A less restrictive assumption is that of SUTVA (*Stable Unit Treatment Value Assumption*), originally introduced by Rubin (1980). In this case, it is assumed that the potential outcomes for the $i-th$ unit just depend on the treatment that the $i-th$ unit received.

In other words, there is no interference between units and there are no different versions of treatments. Consequently, all potential outcomes for the N units can be represented by an array with N rows and two columns, each unit being a row with two potential outcomes, $Y_t(i)$ and $Y_c(i)$.

The stability assumption is almost always made in epidemiological work, even though it is not always appropriate. For example, consider a study of the effect of vaccination on a contagious disease. The greater proportion of the population that gets vaccinated, the less any unit's chance of contracting the disease, even if not vaccinated. In this case, the units are said to interfere with each other.

A more sustainable assumption is that of TUA (*Treatment Unit Additivity*), Rubin (1978). In this case, it is supposed that the individual effect of treatment, $\tau(i)$ is constant for each unit. Assuming a normal bivariate model for the potential outcomes $(Y_t(i), Y_c(i))$ with means (θ_t, θ_c), common variance ϕ_Y and correlation ρ, TUA is equivalent to impose that the potential outcomes are perfectly correlated, $\rho = 1$.

This hypothesis is quite relevant, as the average causal effect becomes meaningful to *all* the units in the reference population. The plausibility of TUA can be *partially* tested, partitioning the population in subsets, where homogeneity is more likely to hold, with respect to the stratification variables. Consequently, TUA can be seen as a weakening of the assumption of unit homogeneity (Holland 1986).

However, according the the approach of Karl Popper (1959, 1983), Dawid (2000) argues that this methodology cannot be termed 'scientific', as it involves non falsifiable assumptions.

Alternative approaches have been proposed to try to tackle this drawback, which will be shown in the next sections.

1.3.2 Bayesian Networks and Graph Models for causal inference

During the 1990s, a new stream of work has witnessed a massive development. Newly available computer technologies fostered the utilisation of advanced probabilistic models, and the 'neo-Bayesian revival' (Dawid 2004) reached its acme with the formal creation of Bayesian Networks (BNs).

Also Causality literature was interested by the new methodology, mostly since the works of Pearl (1993) and Spirtes et al. (1993, SGS hereafter), the first to define causal inference problems in terms of a BN model.

Formally, a BN is represented by $\mathcal{B} = \{\mathcal{G}, \mathbf{P}\}$, where $\mathcal{G}$ is a Directed Acyclic Graph (DAG), and $\mathbf{P}$ includes the conditional probability distributions for the nodes in $\mathcal{G}$. The simplest kind of conditional distribution is a Conditional Probability Table (CPT), i.e. a multi-dimensional array, which is suitable when the nodes involved are discrete-valued.

From the technical point of view, a DAG is a graphical structure $\mathcal{G} = (\mathbf{X}, \mathbf{E})$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is the set of relevant nodes, each of which is associated to one of the random variables in the domain problem, and $\mathbf{E}$ is the set of edges connecting the nodes.

The set $\mathbf{X}$ includes both *unobservable* (such as working hypotheses) and *observable* variables, which become pieces of evidence, once actually observed.

The set $\mathbf{E}$ specifies the alleged relations among the variables in $\mathcal{G}$. A node that 'points' to another is said to be a *parent*, whereas the node that is reached by the arrow is a *child*. The set of the parents of a node X is indicated by pa(X), and the set of its children is ch(X). The nodes in the directed path leaving X, named *descendants*, are grouped in the set de(X), while those preceding it in a directed path are named *ancestors*, an(X).

Without any further assumption, a direct arrow drawn from the node X_1 towards the node X_2 does not imply any causal effect, but only means that the probability distribution of X_2 is modified according to the value assumed by X_1.

More specifically, this circumstance expresses the fact that by means of that graphical representation we are willing to: *a)* establish an explicit association between X_1 and X_2, and *b)* declare a preference in providing

the joint distribution $\Pr(X_1, X_2)$ through the factorization $\Pr(X_1) \times \Pr(X_2 | X_1)$, over any other alternative specifications.

On the contrary, the absence of a direct link from X_1 to X_2 encodes the assumption that the conditional distribution of X_1 is not directly dependent on the possible values that X_2 can take on. Nevertheless, observing X_1 can produce an undirect change in the probability distribution of X_2.

An important feature of a BN is the specification of the joint probability distribution of all the random variables involved. In fact, on the one hand it is always possible to represent a full joint probability distribution of a (high dimensional) set of variables recursively applying the definition of conditional probability (chain rule):

$$\begin{aligned} \Pr(X_1, \ldots, X_n) &= \Pr(X_1 | X_2, \ldots, X_n) \times \Pr(X_2 | X_3, \ldots, X_n) \times \ldots \\ &\quad \times \Pr(X_{n-1} | X_n) \times \Pr(X_n). \end{aligned}$$

Yet, on the other hand the vector $\mathbf{X}$ can be arranged in $n!$ possible ways and many of them could involve difficult specifications of the conditional probabilities required.

On the contrary, the conditional independence relations that characterise a BN induce the most essential factorization of the joint probability distribution, given the present knowledge of the problem:

$$\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i | \mathrm{pa}(X_i)), \tag{1.2}$$

where the conditional distributions on the right hand side are evaluated solely with respect to the variables that are indispensable, i.e. the parents. Condition (1.2) is know as *Markov Property*.

Figure 1.1: An example of DAG

The simplest example of DAG is shown in Figure 1.1. The set of nodes is $\mathbf{X} = \{X_1, X_2\}$, whereas the set of edges $\mathbf{E}$ is settled by the arrow that connects the two nodes: $\mathbf{E} = \{X_1 \rightarrow X_2\}$.

Considering the BN associated to the DAG of Figure 1.1, the set $\mathbf{P}$ has elements $\Pr(X_1)$ and $\Pr(X_2 \,|\, X_1)$, and the joint distribution of the system is factorized according to (1.2) as $\Pr(X_1, X_2) = \Pr(X_1) \times \Pr(X_2 \,|\, X_1)$.

From a purely probabilistic point of view, it would also be possible to express the joint distribution as $\Pr(X_1, X_2) = \Pr(X_1) \times \Pr(X_1 \,|\, X_2)$, or the trivial $\Pr(X_1, X_2) = \Pr(X_1, X_2)$, or even as $\Pr(X_1, X_2) = \Pr(X_1) \times \Pr(X_2)$, this last expression encoding the assumption of independence between the two variables. However, the graphical representation of Figure 1.1 suggests that the experimenter who is building the graph believes in the factorization of (1.2) instead.

Both Pearl and SGS agree that the standard DAG representation of a problem, i.e. a factorization of the joint probability distribution by means of conditional independence relationships, can be extended to give rise to a causal interpretation, provided some further hypotheses. This extension concerns the fact that in case a variable is forced to take on a given value by an *external intervention*, than the structure of the original DAG is modified, as shown in Figure 1.2.

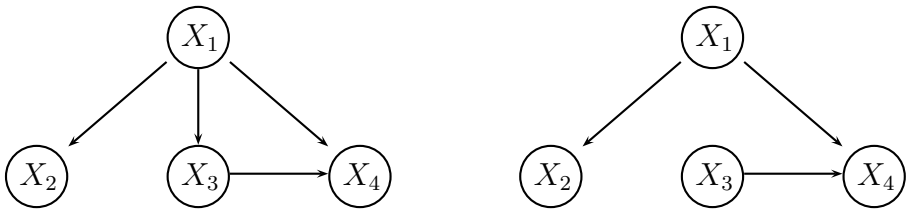


Figure 1.2: The DAG representation of the external intervention. In case the evidence arose by means of external manipulations, the direct connections between the intervened node X_3 and its parent is removed. The rest of the graph is unchanged

In particular, after an intervention, the DAG is 'manipulated', so that this intervention cannot influence the variables that come before it within a directed path (the *ancestors* of the intervened node).

In other words, the BN structure is such that each variable is associated to a suitable (conditional) probability distribution. In case of external intervention, this distribution is modified so that the variable is forced to assume a given value with probability 1. Each intervened variable can only modify its *descendants*.

If we make reference to the DAG of Figure 1.2, under the natural model, the observation of $X_3 = x$ modifies the distribution of X_4 both directly and through updating the distribution of the node X_1. Therefore, the most likely value of X_4 is the one that is most consistent with *a)* the observed

value of X_3, and *b)* the value of X_1 induced by $X_3 = x$.

However, if X_3 did not arise genuinely, the distribution of X_4 is only modified by X_3 itself, as the distribution of X_1 is not updated by the available evidence, since X_3 and X_1 are not directly connected in the intervention model.

While SGS's idea is to explicitly modify the structure of the graph, as shown in Figure 1.2, Pearl models the manipulation by means of an additional node, which is used to delete the connections with the parents of the intervened node.

Provided that the experimenter knows that a variable has been manipulated, it is possible to modify the graphical structure. In order to take into account the confounders and to analyse the effect induced on the response, suitable constraints have been defined, which will be described in section 4.

1.3.3 Causal inference via Structural Equation Models

A slightly different approach is that of Pearl (1995, 2000), in which still maintaining the DAG structure of the causal problem, the situation is modelled by means of some functional relationships between the variables, of the form $X_i = f_i(\text{pa}(X_i), \varepsilon_i)$. The causal mechanisms represented by the f_is are basically deterministic, even if they are perturbed by the random (mutually independent) ε_is.

This approach is quite similar to that of the *Structural Equation Models* (SEM), well known in the Econometrics literature (Haavelmo 1943), where a set of equations describes the impact of variables on each other.

Pearl (1995) names this representation *causal diagram*, and an example is that depicted in Figure 1.3.

The causal assumptions encoded in the model of Figure 1.3 correspond to the following equations:

$$Z_0 = f_0(\varepsilon_0),\ \ Z_2 = f_2(T, Z_1, \varepsilon_2),\ \ B = f_B(Z_0, \varepsilon_B),\ \ Z_3 = f_3(B, Z_2, \varepsilon_3),$$
$$Z_1 = f_1(Z_0, \varepsilon_1),\ \ Y = f_Y(T, Z_2, Z_3, \varepsilon_Y),\ \ T = f_T(Z_0, \varepsilon_T).$$

The external intervention is performed directly on the structural equations. More specifically, the intervention that sets Z_0 to the value z is mimicked replacing the structural equation with the assignment $Z_0 = z$ in all the equations that involve this variable, so that, for instance, the structural equation for the treatment becomes $T = f_T(z, \varepsilon_T)$.

In addition, the equations implicitly provide a clear correspondence with the counterfactual model. In fact, the potential outcomes (Y_t, Y_c) are

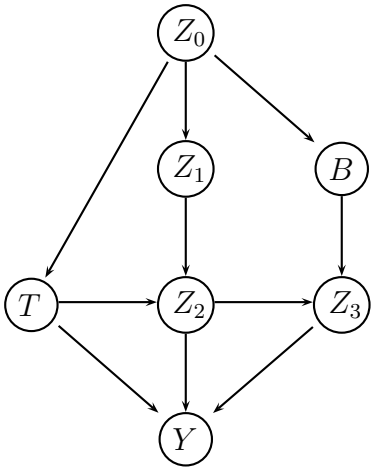


Figure 1.3: An example of causal diagram, representing the effect of T on Y and some confounders. Adapted from Pearl (1995)

directly calculated on the same individual by means of the deterministic function $f_Y(\{\text{pa}(Y) \setminus T\}, \varepsilon_Y)|_{T=t,c}$.

Pearl (1995) arguments that, as compared to the potential outcomes model,

> *...the functional characterisation $X_i = f_i(\text{pa}(X_i), \varepsilon_i)$ also provides a convenient language for specifying how the resulting distributions would change in response to external interventions.*

On the other hand, Imbens & Rubin (1995) argue that the standard causal model based on potential outcomes is preferable as it makes more explicit the distinction between assignment mechanisms, which is typically under the control of the experimenter.

Dawid (1995) argues that Pearl's approach based on manipulated graphs represents a more natural framework for causal analysis, and that the relation with counterfactuals is possible, but inessential, when facing EoC problems. The change in Pearl's view has been explained by the author himself (Pearl 1995, p. 706) as related to the necessity of treating CoE problems, for which the functional model approach and counterfactuals are necessary.

1.3.4 Decision Theoretic approach to Causal Inference

Among other contributions, and exploiting the approaches of Heckerman & Shachter (1994) and Lauritzen (2000), Dawid (2002) proposed an ad-

vanced representation of the causal problem, using a decision theoretic framework.

One relevant feature of this approach is that it is based on the evaluation of conditional probability distributions, rather than being focused on a joint distribution, as happens for the potential outcomes model. This characteristic is quite important, as it avoids problems of identifiability. Moreover, from the philosophical point of view, it is not influenced by *fatalistic* assumptions (Dawid 2000), such as TUA described above.

Within this framework, causal inference is openly modelled in terms of a suitable Augmented DAG (ADAG). This is a DAG including also an external *intervention* variable, F_T, which explicitly rules the behaviour of the treatment variable, T. A simple example of ADAG is depicted in Figure 1.4.

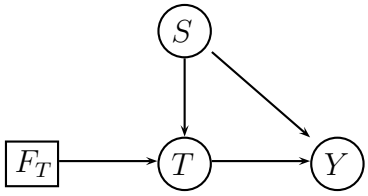


Figure 1.4: An ADAG representation. The variable T is subjected to an intervention, and hence associated to an external intervention node F_T

The possible external intervention is modelled as a *decision variable*, represented as a square. The variable F_T takes on the elements of the set $\{\mathcal{T} \cup \emptyset\}$, where $\mathcal{T}$ is the set of values that the possibly intervened node T may assume.

Unlike a random node, F_T is not associated to a CPT, as its state is always decided (known) by the experimenter. Therefore, it serves as a switch and it is used to allow the experimenter to activate a given scenario. When $F_T = \emptyset$, then the intervention is void, and hence T is a random variable governed by its conditional probability distribution $\Pr(T \,|\, \text{pa}(T))$.

Conversely, when $F_T = t, t \in \mathcal{T}$, then an intervention occurred. As a result, T becomes a degenerate variable, whence $\Pr(T = t \mid \text{pa}(T)) = 1$, for every configurations of the variables in $\text{pa}(T)$. As required, in case of external intervention, the parents are not updated by T.

The difference with the approach of Pearl (1993) is that the intervention is modelled in a more formal way, according to decision analysis, leading to a more straightforward definition of the causal problem. As for the approach of SGS, instead, the main difference is that they related their for-

mulation to counterfactual considerations, whereas Dawid's model, being based on identifiable conditional distribution does not.

In order to apply this model, the basic hypothesis is that the experimenter is able to identify among the set of covariates $\mathbf{C}$, a suitable subset $\mathbf{S}$ such that the distribution of the response Y, given T and $\mathbf{S}$ is the same, regardless on the way that T arose, either by external intervention F_T, or naturally (observational regime).

1.4 Criteria for the estimation of causal effects

The approaches to causal inference based on graphical representation have different philosophies, as described in the previous section. However, from the computational point of view, some criteria have been individuated, that allow the identification of the desired causal effect. In particular, the models of Pearl (1993) and Dawid (2002) share several common points, as compared that of Pearl (1995), whose focus is mainly on a different problem (that of CoE).

Two of the most relevant criteria proposed in the literature are described in the following, although others do exist (see Robins 1986), which are not reviewed here.

The main problem when trying to estimate a causal effect from observational data is that of reducing confounding bias due to spurious correlations between the treatment and the response.

Pearl (1993) provides two tests based on the topology of the graph, which allow the identification of the required conditional independence conditions that enable the experimenter to identify the causal effect.

1.4.1 The back door criterion

Suppose that a problem is suitably represented by a set of variables $\{\mathbf{C}, T, Y\}$, where the objective of the analysis is to estimate the causal effect of T on Y from observational data.

A subset $\mathbf{S} \subseteq \mathbf{C}$ is said to satisfy the *back door criterion* with respect to (T, Y) in a suitable graphical representation $\mathcal{G}$ if:

$i.$ no node in $\mathbf{S}$ is a descendant of T;

$ii.$ $\mathbf{S}$ blocks every path between T and Y which contains an arrow into T.

If these two conditions hold, then the causal effect of T on Y is identifiable (i.e. the confounders are correctly taken into account), and is computed as:

$$\Pr(Y \,|\, T = t) = \sum_{\mathbf{S}=\mathbf{s}} \Pr(Y \,|\, T = t, \mathbf{S} = \mathbf{s}) \Pr(\mathbf{S} = \mathbf{s}). \tag{1.3}$$

Formula (1.3) represents a weighted average, where the 'effect' of the confounders is marginalised off, in order to obtain only the relevant impact of T on Y.

Of course, the underlying assumption is that the covariates in $\mathbf{C}$ are exhaustive for the problem. Obviously, this hypothesis only approximates the real problem under study. Moreover, all the variables in $\mathbf{S}$ need be observed, in order to estimate the required causal effect.

Pearl (1993) points out that using the graphical representation of the problem provides a more straightforward solution to the problem of deriving the so called *ignorability conditions*, under which a causal model is identifiable.

These causal constraints have been formalised in terms of conditional independence relationships by Dawid (2002):

$$\mathbf{S} \perp\!\!\!\perp F_T \tag{1.4}$$

$$Y \perp\!\!\!\perp F_T \,|\, \mathbf{S} \cup T \tag{1.5}$$

on the ADAG, or equivalently:

$$\text{an}(\mathbf{S}) \cap T = \emptyset \tag{1.6}$$

$$Y \perp\!\!\!\perp \text{pa}^0(T) \,|\, \mathbf{S} \cup T \tag{1.7}$$

on the corresponding unaugmented DAG.

By condition (1.4), Dawid's model assumes that the probability distribution of the observed confounders $\mathbf{S}$ must not depend on how T arose. In other words, if it is known that an intervention occurred on T, the distributions of the variables in $\mathbf{S}$ must not change with the value set for F_T. Moreover, these distributions must remain the same as the case in which the evidence is certainly genuine ($F_T = \emptyset$), but T has not been observed yet.

Assumption (1.5) instead indicates that the knowledge of T and $\mathbf{S}$ is *all* that is needed for Y to be independent on F_T, in which case the response is not modified by the way that the treatment arose. This situation basically amounts to the fact that the causal mechanism that relates T to Y is conveniently explained by the variables in the set $\{T, \mathbf{S}\}$, so that the

differences in the response Y can be directly attributable to T, once $\mathbf{S}$ is controlled for.

If both the assumptions hold, then the causal effect under experimental conditions $\Pr(Y \mid F_T = t, \mathbf{S} = \mathbf{s}) = \Pr(Y \mid\mid T = t, \mathbf{S} = \mathbf{s})$ equals the observational posterior probability $\Pr(Y \mid F_T = \emptyset, T = t, \mathbf{S} = \mathbf{s})$. Consequently, the experimenter can use the latter in order to correctly estimate the former.

A set $\mathbf{S}$ which verifies (1.4) and (1.5) is said to be a set of *sufficient covariates* for the estimation of the causal effect of T on Y. Notice however that the term 'sufficient' is not to be intended as in standard statistical analysis.

A basic distinction is that while a sufficient statistic is such that any larger statistic is sufficient *a fortiori*, this property does not hold for a set of sufficient covariates.

In fact, considering for example the ADAG of Figure 1.5, the node S clearly verifies conditions (1.4) and (1.5) and therefore is a sufficient covariate. However, the set $\mathbf{L} = \{C, S\}$, which includes S, and therefore is 'larger', is not a set of sufficient covariates, as clearly $\mathbf{L} \not\perp\!\!\!\perp F_T$, i.e. condition (1.4) does not hold.

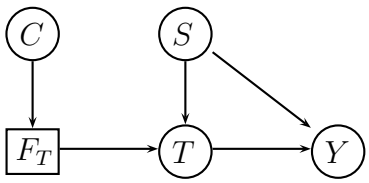


Figure 1.5: The node S is a sufficient covariate, whereas the set $\mathbf{L} = \{S, C\}$, despite including S, is not

1.4.2 The front door criterion

The back door criterion is useful to model situations where the 'treatment' has a direct impact onto the 'response'. However, some situations can occur where the impact of the treatment is induced by an 'active agent', using the terminology of Lauritzen (2000).

Suppose for example that a problem consists of the variables $\{U, T, Z, Y\}$, and the objective is to evaluate the causal effect of T on Y. Suppose further that a suitable graphical representation is that of Figure 1.6, where U represents a latent, unobservable variable.

The node Z clearly does not verify the conditions for the back door criterion, as it belongs to the set de(T). In order to estimate the desired causal

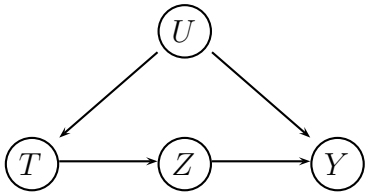


Figure 1.6: A diagram that represents the front door criterion. Source: Pearl (1995)

effect, the idea is then to decompose the joint probability distribution of the system into a set of factors which involve directly estimable probabilities.

In general, a set of variables $\mathbf{Z}$ satisfies the *front door criterion* with respect to (T, Y) if:

$i.$ $\mathbf{Z}$ intercepts all the directed paths from T to Y;

$ii.$ There is no back door path between T and $\mathbf{Z}$;

$iii.$ Every back door path between $\mathbf{Z}$ and Y is blocked by T.

From the graphical representation of Figure 1.6, the joint probability distribution of the system can be written as:

$$\Pr(U, T, Z, Y) = \Pr(U)\Pr(T\,|\,U)\Pr(Z\,|\,T)\Pr(Y\,|\,Z, U),$$

and, since in case U was observed, it would verify the back door criterion, then applying (1.3), the causal effect of T on Y would be given by:

$$\Pr(Y\,|\,T = t^*) = \sum_{u \in \mathcal{U}} \Pr(Y\,|\,U = u, T = t^*)\Pr(U = u),$$

where $\mathcal{U}$ is the domain of U.

The node U cannot be observed by definition, but using the conditional independence assumptions implied in the graphical representation of Figure 1.6, one can eliminate it to obtain:

$$\Pr(Y\,|\,T = t^*) = \sum_{t^* \in \mathcal{T}} \Pr(Z\,|\,T = t^*) \sum_{t' \in \mathcal{T}} \Pr(Y\,|\,T = t', Z)\Pr(T = t'). \quad (1.8)$$

Equation (1.8) is referred to as the front door formula.

1.5 Conclusions

In this paper we reviewed some of the major approaches to causal inference in the statistical literature. In particular, we give credit to the decision theoretic approach, which is based on the explicit reference to a decision framework, as implemented by the Augmented DAG.

Besides the philosophical advantage of dispensing of any counterfactual considerations, which is quite valuable in our view, the ADAG approach seems to be particularly useful, as it can be used in a quite straightforward way to solve problems of Effect of Causes, provided a suitable graphical representation of the system under study.

We also review two of the main graphical methods that are used to find out the conditions that allow the identifiability of the causal effect. These methods can be used to derive proper calculations of such effect.

In our view, some research areas remain open to investigation: first, Dawid's approach is purely subjectivist, as it involves the expert prior definition of the correct structural representation of the model.

On the one hand, this is far from being a drawback in our view, leading to a model formulation that is explicitly open to criticism and that easily allows the introduction of newly available information.

On the other hand, it seems that this feature has been perceived as a limitation of the applicability of the decision theoretic approach. For this reason, we reckon that one open area of research is that devoted to the identification of some semi-automated procedures that allow some sort of data driven structural learning, given the need for some prior information and the verification of causal constraints.

Second, we reckon that the decision theoretic approach can have some natural extensions in the fields of economic evaluations, such as that of Health Technology Assessment (HTA). In a typical HTA problem, the objective is that of evaluating the cost-effectiveness of a given health resource, as compared to an alternative programme. Often, observational studies from clinical practice are used for the economic evaluation.

The ADAG framework could be applied to model the results in terms of causal effect (which can represent the measure of effectiveness) of a given treatment, controlling for biasing covariates.

Moreover, the natural extension to the Influence Diagram framework, implicit in the ADAG construction, can allow the formal modelling of the associated costs in terms of (dis)utilities associated to the different treatments (cfr. Baio et al. 2004, and Chapter 4).

Bibliography

Baio, G., Pammolli, F., Baldo, V. & Trivello, R. (2004), 'The economic evaluation of influenza vaccination in the elderly population: a model based on Bayesian Networks and Influence Diagrams', submitted to the *Journal of Health Economics*.

Balke, A. & Pearl, J. (1994), Counterfactual probabilities: Computational methods, bounds, and applications, *in* R. Lopez de Mantaras & D. Poole, eds, 'Uncertainty in Artificial Intelligence', Morgan Kaufmann, San Mateo, CA, pp. 46–54.

Cook, T. & Campbell, D. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Miflin, Boston, MA.

Cowell, R., Dawid, P., Lauritzen, S. & Spiegelhalter, D. (1999), *Probabilistic Networks and Expert Systems*, Springer-Verlag, New York, NY.

Dawid, A. (1979), 'Conditional independence in statistical theory (with Discussion)', *Journal of the Royal Statistical Society,* B **41**, 1–31.

Dawid, A. (2000), 'Causal Inference without Counterfactuals (with Discussion)', *Journal of the American Statistical Association* **95**, 407–448.

Dawid, A. (2002), 'Influence Diagrams for causal modelling and inference', *Statistical Review* **70**, 161–189.

Dawid, P. (1995), 'Discussion of 'Causal diagrams for empirical research' by J. Pearl', *Biometrika* **82**, 689–690.

Dawid, P. (2004), 'Probability, Causality and the Empirical World: A Bayes-de Finetti-Popper-Borel Synthesis', *Statistical Science* **19**, 44–57.

Greenland, S. (2004), An overview of methods for causal inference from observational studies, *in* A. Gelman & X. Meng, eds, 'Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspective', John Wiley and Sons, Chichester, UK.

Haavelmo, T. (1943), 'The statistical implications of a system of simulataneous equations', *Econometrica* **11**, 1–12.

Heckerman, D. & Shachter, R. (1994), Decision-Theoretic Foundations for Causal Reasoning, Technical Report MSR-TR-94-11, Microsoft Research Advanced Technology Division, Redmond, WA.

Holland, P. (1986), 'Statistics and Causal Inference', *Journal of the American Statistical Association* **81**, 945–968.

Hume, D. (1748), *An Enquiry Concerning Human Understanding*.

Imbens, G. & Rubin, D. (1995), 'Discussion of 'Causal diagrams for empirical research' by J. Pearl', *Biometrika* **82**, 694–695.

Klein, T. (2003), 'A Note on Causal Inference with Potential Outcomes and without Counterfactuals', *Unpublished manuscript*.

Lauritzen, S. (2000), Causal Inference from Graphical Models, *in* O. Barndorff-Nielsen, R. Cox & C. Klüppelberg, eds, 'Complex Stochastic Systems', Chapman and Hall, London, UK.

Pearl, J. (1993), 'Comment: Graphical models, causality, and intervention', *Statistical Science* **8**, 266–269.

Pearl, J. (1995), 'Causal Diagrams for Empirical Research', *Biometrika* **82**, 669–710.

Pearl, J. (2001), 'Causal Inference in the Health Sciences: A Conceptual Introduction', *Health Services and Outcomes Research Methodology* **2**, 189–220.

Popper, K. (1959), *The Logic of Scientific Discovery*, Hutchinson, London, UK.

Popper, K. (1983), *Realism and the Aim of Science*, Hutchinson, London, UK.

Reichenbach, H. (1956), *The Direction of Time*, University of California Press, Berkeley, CA.

Robins, J. (1986), 'A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect', *Mathematical Modelling* **7**, 1393–1412.

Rosenbaum, P. (1995), *Observational Studies*, Springer Series in Statistics, New York, NY.

Rubin, D. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **68**, 688–701.

Rubin, D. (1978), 'Bayesian inference for causal effects: the role of randomization', *Annals of Statistics* **6**, 34–68.

Rubin, D. (1980), 'Comment on 'Randomization analysis of experimental data: the Fisher randomization test' by D. Basu', *Journal of the American Statistical Association* **75**, 591–593.

Shafer, G. (1996), *The Art of Causal Conjectures*, MIT Press, Cambridge, MA.

Spirtes, P., Glymour, C. & Scheines, R. (1993), *Causality, Prediction and Search*, Springer-Verlag, New York, NY.

von Wright, G. (1973), *Explanation and Understanding*, Cornell University Press, Ithaca, NY.

CHAPTER 2

APPLIED CAUSAL INFERENCE FROM OBSERVATIONAL DATA: LEARNING CAUSAL STRUCTURE AND SUFFICIENT COVARIATES

2.1 Introduction

Causal inference has always been a major topic in a number of research areas, from Philosophy to Epidemiology, Economics and Econometrics, Engineering, Law, and of course Statistics.

Many attempts have been made in order to find a procedure that allows the researcher to discover causal relationships between two (sets of) variables.

Agreement on the method was hardly reached by the various scholars, and the debate is still going on. Yet, it is now well accepted that: *a)* causal inference is a different, more challenging task from standard statistical analysis, and *b)* in order to find some causal implication, a set of suitable assumptions, which may vary from one approach to another, must hold.

The reason why standard statistical methodologies are not reliable for causal inference is related to the fact that they are typically based on a conditional probability of observing a variable Y, after *observing* a variable X, $\Pr(Y \mid X = x)$. Conversely, causal modelling should be concerned with a different quantity, i.e. the conditional probability of observing Y, after *setting* the value of X, $\Pr(Y \mid\mid X = x)$, using the notation introduced by Lauritzen (2000). Since, in general, these two distributions are different, in order to assess the latter using the former, it is necessary to make some suitable assumptions.

Among the various approaches to causal modelling that can be found in statistical literature (Rubin 1974, 1978; Pearl 1993, 1995, 2000; Spirtes et al. 1993), we focus on the *decision theoretical*, described by Heckerman & Shachter (1994) and further formalised by Dawid (2000, 2002).

Within this framework, causal inference is explicitly modelled in terms of a suitable Augmented DAG (ADAG). This is a DAG including also an external *intervention* variable, F_T, which explicitly rules the behaviour of the treatment variable, T. Such a model is depicted in Figure 2.1.

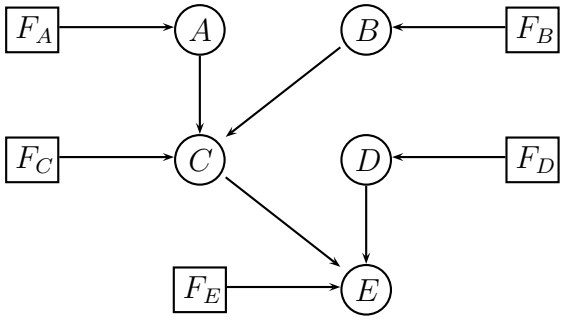


Figure 2.1: An example of Augmented DAG. Each node $\nu = \{A, B, C, D, E\}$ is subjected to intervention, and hence associated to an external decision node F_ν. Source: Dawid (2002)

The external intervention is modelled as a decision variable, whose possible values are $\{\mathcal{T} \cup \emptyset\}$, where $\mathcal{T}$ is the set of values that the variable T may take on, i.e. the support. When $F_T = \emptyset$, then the intervention is void, and hence T is observed in a natural regimen. Conversely, when $F_T = t, t \in \mathcal{T}$, then $\Pr(T = t \,|\, \text{pa}(T)) = 1$.

Notice that this procedure is such that intervened variables have no effect on their ancestors, but they do have an impact on their descendants. Moreover, it entirely dispenses with any counterfactual assumptions - see Dawid (2000) for a thorough discussion.

Moreover, the basic hypothesis is that the experimenter is able to identify among the set of covariates $\mathbf{C}$, a suitable subset $\mathbf{S}$ such that the distribution of the response Y, given T and $\mathbf{S}$ is the same, regardless on the way that T arose, either by external intervention F_T, or naturally (observational regime).

The sufficient covariates $\mathbf{S}$ are supposed to be independent on the way that the treatment is chosen, either by intervention, i.e. $F_T = t$, which forces T to take on that value too, or naturally, i.e. $F_T = \emptyset$, in which case T takes on a value in its support, according to its conditional distribution $\Pr(T = t \,|\, \text{pa}(T))$. As a result, in this case, causal inference is possible and unbiased starting from observational data.

Although difficult to empirically test, these assumptions have been formalised in terms of conditional independence relationships (Dawid 2002):

$$\mathbf{S} \perp\!\!\!\perp F_T \tag{2.1}$$

$$Y \perp\!\!\!\perp F_T \,|\, \mathbf{S} \cup T \tag{2.2}$$

on the ADAG, or equivalently:

$$\text{an}(\mathbf{S}) \cap T = \emptyset \tag{2.3}$$

$$Y \perp\!\!\!\perp \text{pa}(T) \,|\, \mathbf{S} \cup T \tag{2.4}$$

on the unaugmented DAG.

On the one hand, the ADAG formulation is more appealing, as it allows the straightforward evaluation of the response distribution, simply using its specialised semantics. This represents a great advantage of this approach, both under a philosophical and an operational point of view.

However, on the other hand, this procedure only works once that the structure of the relationships among the variables is known, e.g. if the experimenter takes the responsibility to build up the ADAG in a suitable way for the problem under consideration. Learning this structure from observed data seems to be cumbersome, as the intervention nodes distributions will always be degenerate, under observational conditions, where Nature is in control.

Conversely, since the node F_T is not formally considered, the unaugmented DAG representation does not allow for direct causal interpretation (as it only encodes conditional independence statements). Yet, for the same reason, it permits to perform the structural learning, using standard (constrained) algorithms.

A possible strategy that makes the most of the two alternatives could be that of learning the structure of the unaugmented DAG associated to a given problem, under the constraint that $T \rightarrow Y$. Once this procedure is performed, it is possible to augment the DAG, in order to work on a causal structure that is the most supported by the empirical data, among all the models that verify this condition. The set $\mathbf{S}$ for which conditions (2.3) and (2.4) hold can be defined, and causal effect can be calculated working directly on the ADAG.

Using this switch between the two graphical representation of the problem, the aim of this paper is to complement Dawid's model in order to:

a) define a specialised structural learning strategy that takes into account the fact that the search is not performed over the space of all possible models associated to the nodes in the analysis (as would happen for a normal DAG), but only over the subspace in which particular assumptions hold;

b) define also a strategy for the evaluation of the result obtained, that involves the analysis of the most supported graphical structure.

This approach is described in the following.

2.2 Learning the structure for causal discovery

2.2.1 Basic assumptions

Since we reckon that causality cannot be exploited simply by observational data, we aim at building a method that can help the experimenter put together all the available prior knowledge and to relax (though not to avoid) some of the assumptions needed for causal discovery.

In this spirit, we first focus on a specific situation, well defined by some initial conditions, and let more complicated definitions of the problem as a further topic of research.

The very basic assumptions underlying our approach are summarised by the following arguments.

A1 Suppose that an observational dataset about a treatment variable T, a response Y and a set of covariates of interest for the problem, $\mathbf{C} = \{C_1, \ldots, C_k\}$ is available. The presence of missing data is possible.

A2 The analysis aims at making causal inference on the effect of T on Y in order to evaluate the effect on the $(n+1)-th$ case. This assumption essentially rephrases the causal problem in terms of Dawid's *effects of causes* (EoC) (Dawid 2000). In other words, T is assumed to cause Y, and the goal is to quantify the causal effect, discarding all the spurious (confounding) impact of other variables.

A3 The relationships among the variables are represented in terms of an ADAG. The experimenter is able to encode all the prior knowledge into it, but is not willing to work in a completely Expert System framework (i.e. does want to learn from the data some parts of the graphical structure).

A4 The main objective of the learning procedure is to discover *probabilistic* relationships, rather than *causal* ones. In fact, while causation is encoded only in the relationship $T \rightarrow Y$, which is assumed to hold and cannot be removed or modified, the rest of the graph will be concerned with a set of probabilistic features, that are used to discern the direct impact of T on Y.

A5 No further constraints will be included in the learning procedure, so that each covariate is able, but not forced, to: *a)* have a direct impact on either the treatment, or the response; *b)* have a direct impact on an other covariate; and *c)* be subjected to a direct influence from either the treatment or the response.

While A1, A2 and A3 just describe the general framework of our approach, A4 and A5 have relevant practical implications.

By A4, we assume that the experimenter is able to assert the absence of *active agents*, using the terminology of Lauritzen (2000). In other words, the causal effect of the treatment on the response may be biased by the influence of some covariates (see A5). Yet, it is not possible that a variable C exists, which entirely accounts for that effect, since we assume that a direct link between T and Y certainly exists.

As a consequence of this assumption, the *back door* criterion (Pearl 1993) is used to determine the conditions for estimability of the causal effect, as expressed in terms of (2.3) and (2.4). The use of other criteria, such as that of *front door* (Pearl 1993, 1995), is prevented from the presence of the direct connection between the treatment and the response (recall that, in order to apply the front door criterion, a variable C must intercept all direct paths from T to Y, which is impossible if the link $T \rightarrow Y$ exists).

Although this may prove not to hold for some specific causal inference problems, we believe that it is crucial to a great number of them, and hence we focus on this approach.

Finally, assumption A5 suggests that, since no other constraint is considered, apart from the one described in A4, the structural learning can be accomplished, modifying standard algorithms.

2.2.2 The learning algorithm

Two relevant approaches to structural learning are those based on *greedy search* algorithms and those based on a *full Bayesian* procedure (see for example Jordan 2001).

The greedy search algorithm is a local search procedure, which allows to retrieve the (local) maximum on the space of models. A single structure is provided as the result of the algorithm, which is assumed to be the most supported from the observed data.

Conversely, the full Bayesian procedure works in a model averaging framework: at the end of the learning procedure, one ends up with k models, accounting for a high percentage of the posterior probability, over the space of models. Averaging over the most plausible structures, would allow to produce a robust estimation.

While constraint-based learning is not a new topic in Computer Science and Statistics, our case presents some possible problems. In particular, the major issue concerns the fact that models with the same probabilistic features are essentially different for causal purposes.

Suppose for instance that a domain problem consists of the set of variables $\mathbf{X} = \{C_1, C_2, T, Y\}$. The objective of the analysis is to estimate the causal effect of T on Y, under the assumptions A1 - A5, i.e. the link $T \rightarrow Y$ is the causal constraint.

Suppose that a learning procedure is performed, and that the graphical structure of Figure 2.2 is accepted as the most likely to have generated the observed data.

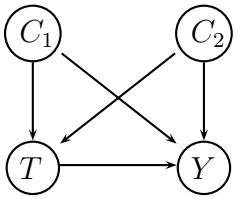


Figure 2.2: Hypothetical result of the causal learning procedure

However, from the probabilistic point of view, the graph of Figure 2.2 is equivalent to that of Figure 2.3.

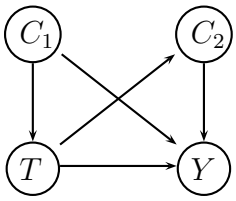


Figure 2.3: A DAG in the same equivalence class of that of Figure 2.2

While in the situation of Figure 2.2 both C_1 and C_2 are needed to estimate the causal effect of T on Y (i.e. they are both sufficient covariates), in the case of Figure 2.3 only C_1 is a sufficient covariate, since C_2 clearly does not satisfy the back door criterion, which we use to assess causality. In this situation, the knowledge of C_2 would better off the precision of the estimation of the causal effect, but would not be required to calculate it.

For this reasons, we reckon that this feature deserves further research, which we will devote in the next future.

2.3 Finding a set of sufficient covariates

Once that the most supported structure is identified, in order to calculate causal effects, based on the assumptions A1-A5, one needs to individuate a subset $\mathbf{S} \subseteq \mathbf{C}$ such that (2.3) and (2.4) holds. Moreover, for the sake

of parsimony, it would be a valuable property should this set be the least demanding for the experimenter.

This problem is equivalent to finding a set of (minimal) separators between Y and $\text{pa}(T)$.

Tian et al. (1998) suggest several algorithms to accomplish this search, based on the d-separation criterion (Verma & Pearl 1990). The basic intuition underlying these algorithms is that a general set for which two nodes A and B are separated in a graph $\mathcal{D}$ is easily found as the set $\mathbf{\Pi} = \text{pa}(A \cup B)$, that is:

$$A \perp\!\!\!\perp B \,|\, \mathbf{\Pi}.$$

Moreover, it is possible to refine this set, using the fact that any separator which cannot be reduced by a single node is then minimal.

In other words, the algorithm works as follows - see Tian et al. (1998) for some comments on the computational effort required:

1. *Define* $\mathbf{K} = Y \cup \text{pa}(T)$*;*

2. *Construct the set* $\mathcal{D}_{An(\mathbf{K})}$*, where:*

$$An(\mathbf{K}) = Y \cup \text{pa}(T) \cup \left(\cup_{\nu \in \text{pa}(T)} \{\text{an}(\nu)\}\right)$$

   is the smallest ancestral set of $\mathbf{K}$*, and* $\mathcal{D}_{An(\mathbf{K})}$ *indicates the part of the original graph* $\mathcal{D}$ *that includes only the nodes in* $An(\mathbf{K})$*;*

3. *Set* $\mathbf{S} = \text{pa}(\mathbf{K})$*;*

4. *Choose one node* ν *from* $\mathbf{S}$*;*

5. *Test if* $\mathbf{S} \setminus \{\nu\}$ *is a separator in* $\mathcal{D}_{An(\mathbf{K})}$*, e.g. using the algorithm of Geiger et al. (1990);*

6. *If* $\mathbf{S} \setminus \{\nu\}$ *is a separator, then set* $\mathbf{S} = \mathbf{S} \setminus \{\nu\}$*; choose a different node from* $\mathbf{S}$*, denote it by* $\{\nu\}$ *and go to step 5. If all nodes in* $\mathbf{S}$ *have been checked, stop;*

7. *return* $\mathbf{S}$.

A problem that may arise concerns the fact that more than one set of covariates turn out to be sufficient and minimal. We address this question in the following.

2.3.1 Choosing among minimal sets of covariates

Suppose for example that the treatment T is the prescription of a given cardiovascular drug, Y is the occurrence of a myocardial infarction (assumed to be a relevant response) and the set $\mathbf{C}$ includes some individual information, such as age A, gender G, and the result of two diagnostic tests, T_1 and T_2.

Suppose that after a learning procedure is performed, the doctor is provided with the most plausible causal mechanism, as depicted in Figure 2.4.

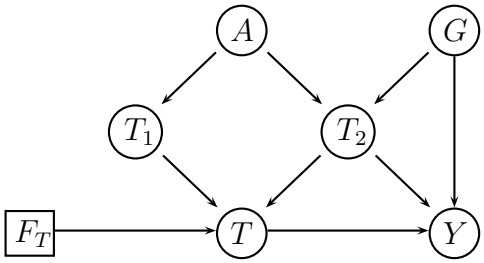


Figure 2.4: Hypothetical results for the cardiovascular example: the most supported causal structure after the learning procedure

In this case, the set $\mathbf{S}$ includes the variable T_2 and either of the variables T_1 or G. Considering the moralised version of the graph, there are in fact four paths connecting Y to F_T:

1. $Y - T - F_T$;
2. $Y - T_2 - F_T$;
3. $Y - G - A - T_1 - F_T$;
4. $Y - T_2 - T_1 - F_T$.

As a consequence, using the algorithm of section 2.3, the minimal sets of sufficient covariates are $\mathbf{S}_1 = \{T_2, G\}$ and $\mathbf{S}_3 = \{T_2, T_1\}$, i.e.:

$$Y \perp\!\!\!\perp F_T \,|\, T_2, G, T \qquad \text{or} \qquad Y \perp\!\!\!\perp F_T \,|\, T_1, T_2, T.$$

Let test T_1 be a highly expensive one, such as those performed using a computer assisted device, whereas test T_2 is just a (low cost) simple measurement that the doctor can perform during the visit. Assume also that the measurement of A and G is at no cost.

Suppose further that, on average, knowing test T_1 produces a lower uncertainty on the outcome conditional distribution, as compared to that deriving from observing G.

Now, what should the practitioner do?

The answer to this question is that the decision should typically balance the level of knowledge that is required, and the (ethical and/or economic) sustainability of the choice.

Hence, an optimal strategy is one that weighs the total cost (opportunely defined) of the sufficient covariates set and the result in terms of the variability of the conditional distribution of the response.

Back to the cardiovascular example, should the difference in terms of reducing the uncertainty deriving by the use of T_1 be non significant, then the doctor could be satisfied with the investigation only of the test T_2 and the new patient's gender.

Knowing this set of covariates, the net effect of the prescribed drug on the occurrence of infarction could be determined easily, applying the *back door formula*, cfr. Dawid (2002, p. 174), or even more straightforwardly, using the semantics of the ADAG. The expensive test T_1 would not be required to unbias the analysis of the effect of the drug on the response.

Conversely, if the difference in uncertainty is significant (i.e. knowing test T_1 generates a dramatically more precise estimation of the effect of T on Y), then the decision maker could justify the additional costs, based on that scientific evidence.

An other important feature of the model is that, according to its very definitions (Dawid 2000, 2002), the evaluation can be performed at the individual level.

For example, suppose that a GP visits two new patients, p_1 and p_2, in the case where, yet being more expensive, the knowledge of the test T_1 slightly reduces the variability of the estimation of the effect of T on Y.

Suppose also that patient p_1 has been already tested with T_1, while p_2 has not yet. In this situation, the GP could choose to provide a more accurate estimation for patient p_1, as this improved accuracy is at no cost.

As for patient p_2, given that the improvement is not massive, the GP can just measure G and T_2, and decide upon the treatment consequently.

A specified *probabilistic sensitivity analysis* can also be performed, in order to derive a break even point; in other words, one could estimate the minimum increase in the precision of the estimation that is worth the deriving additional costs. The decisions would then be taken accordingly.

Moreover, a further interesting feature of the search strategy described above is that it is possible to assign a different weight to each node in $\mathbf{S}$. Then minimality would not be tested with respect to the cardinality of $\mathbf{S}$,

but rather on the actual possibility of measuring the selected nodes on the $(n+1)-th$ case, as described by the associated weight.

2.4 Sufficient covariates and propensity scores

The method of *propensity scores* was introduced by Rosenbaum & Rubin (1983) as a means of examining causal effects in observational studies.

The essential feature of their approach is that causal inference is based on counterfactuals and the potential outcomes model (Rubin 1974). Assuming for the sake of simplicity that the treatment is binary (t, c), the causal effect is typically evaluated in terms of the difference between the response $Y_t(i)$ that *has actually been observed* on the $i-th$ individual for the assigned treatment t, and the response $Y_c(i)$ that *would have been observed*, should the treatment selected be c.

Moreover, particular attention is devoted to treatment assignment, which is an issue for non-randomised studies, where units may have been assigned to a treatment in some deliberate way, which biases naïve comparisons.

A similar reasoning also applies to the ADAG representation of causal problems. Suppose for example that a covariate C is sufficient to identify the causal effect of a treatment T over a suitable response Y, i.e. C verifies (2.1) and (2.2), as depicted in Figure 2.5.

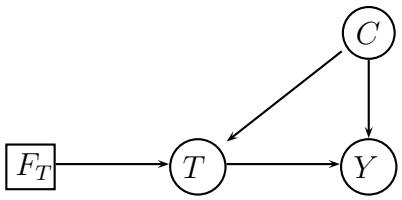


Figure 2.5: The covariate C is sufficient to identify the causal effect of the treatment T over the response Y

The original definition of the propensity score provided by Rosenbaum & Rubin (1983) can be re-arranged in the ADAG framework as:

$$S := \Pr(T = t \,|\, C, F_T = \emptyset). \tag{2.5}$$

As a consequence of (2.5), $\Pr(T = t \,|\, C, S, F_T = \emptyset)$ depends exclusively on S. In fact, S is by definition a function of C, whence $T \perp\!\!\!\perp C \,|\, S, F_T = \emptyset$. Moreover, since T is degenerate when $F_T \neq \emptyset$, then:

$$T \perp\!\!\!\perp C \,|\, S, F_T.$$

The original ID representation of Figure 2.5 can now be extended by inserting S on the path from C to T, as in Figure 2.6. From this, it is possible to easily read off the counterparts of (2.1) and (2.2) with S as sufficient covariate.

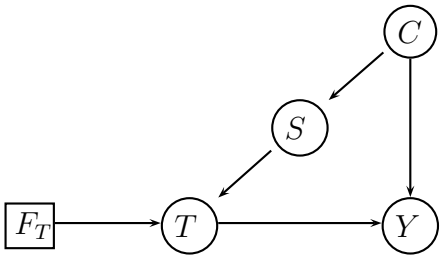


Figure 2.6: The propensity score in terms of the ADAG representation

The main implication of this analysis is that for certain purposes, it is possible to simplify the problem by using S instead of C. The graphical representation helps clarify the role of propensity scores.

However, this result is only possible when it is plausible to assume that C is a sufficient covariate, again a condition that is clearly detectable from the graphical structure of the problem.

Moreover, let us consider an Analysis of Covariance model in terms of C:

$$\mathrm{E}(Y \,|\, T, C, [F_T]) = \alpha_T + \beta' C \tag{2.6}$$

(variables in square brackets are not essential in the conditioning set). From (2.6) it follows that:

$$\mathrm{E}(Y \,|\, T, S, [F_T]) = \alpha_T + \beta' \mathrm{E}(C \,|\, S, [T]). \tag{2.7}$$

For simple analysis, such as testing the irrelevance of treatment, or estimating $\alpha_t - \alpha_c$, it is possible to use the simpler model (2.7).

However, replacing (2.6) with (2.7) for purposes of prognosis of an individual might be neglecting potentially valuable prognostic information in C that is not in S.

All this ignores the real problem, that S is not known a priori, and hence needs be estimated from observed data. Consequently, using a propensity score is not obviously a simplification of the analysis, and may not lead to more precision.

Rephrasing the propensity scores framework in terms of the ADAG representation can make this feature clearer.

2.5 Conclusions

This paper represents a first attempt that we make to find a procedure that allows an experimenter to combine some prior knowledge, some structural constraints and observed data, in order to exploit causality. Working incrementally on Dawid's approach, we built a primal framework to accomplish this task.

However, the complete feasibility of the model needs further investigation, as we identified at least two major problems. First, since structural learning exploits probabilistic relationships, there is the possibility that two or more different models encoding different causal structures, are in fact in the same probabilistic equivalence class.

Moreover, an other important issue is that of the possibility of finding more than one set of sufficient covariate, which would allow the estimation of causal effects.

This second topic seems less troublesome, as we identified a possible strategy that would allow the analysis of such situations. In addition, the possibility of taking into account different variables, upon varying the availability on the $(n+1)-th$ case is highly valuable, and fits the philosophy of EoC problem.

Finally, an other research area that is worth of attention is in our opinion the link that this modelling has with the *propensity scores*, as defined and used in counterfactual analysis. Also in this case, further analysis is required in order to exploit more thoroughly the common elements among the two approaches.

Bibliography

Dawid, A. (1979), 'Conditional independence in statistical theory (with Discussion)', *Journal of the Royal Statistical Society,* B **41**, 1–31.

Dawid, A. (2000), 'Causal Inference without Counterfactuals (with Discussion)', *Journal of the American Statistical Association* **95**, 407–448.

Dawid, A. (2002), 'Influence Diagrams for causal modelling and inference', *Statistical Review* **70**, 161–189.

Geiger, D., Verma, T. & Pearl, J. (1990), d-Separation: From Theorems to Algorithms, *in* 'Proceedings of the 5th Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)', Elsevier Science, New York, NY.

Heckerman, D. & Shachter, R. (1994), Decision-Theoretic Foundations for Causal Reasoning, Technical Report MSR-TR-94-11, Microsoft Research Advanced Technology Division, Redmond, WA.

Jordan, M. (2001), *Learning in Graphical Models*, MIT Press, Cambridge, MA.

Lauritzen, S. (2000), Causal Inference from Graphical Models, *in* O. Barndorff-Nielsen, R. Cox & C. Klüppelberg, eds, 'Complex Stochastic Systems', Chapman and Hall, London, UK.

Lauritzen, S. & Richardson, T. (2002), 'Chain graph models and their causal interpretations', *Journal of the Royal Statistical Society* B **64**, *Part 3*, 321–361.

Pearl, J. (1993), 'Comment: Graphical models, causality, and intervention', *Statistical Science* **8**, 266–269.

Pearl, J. (1995), 'Causal Diagrams for Empirical Research', *Biometrika* **82**, 669–710.

Pearl, J. (2000), *Causality*, Cambridge University Press, Cambridge, UK.

Rosenbaum, P. & Rubin, D. (1983), 'The Central Role of the Propensity Score in Observational Studies for Causal Effect', *Biometrika* **70**, 41–55.

Rubin, D. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **68**, 688–701.

Rubin, D. (1978), 'Bayesian inference for causal effects: the role of randomization', *Annals of Statistics* **6**, 34–68.

Senn, S. & Graf, E. (2004), 'Some Comments on the Propensity Score and Analysis of Covariance', *Unpublished manuscript*.

Spirtes, P., Glymour, C. & Scheines, R. (1993), *Causality, Prediction and Search*, Springer-Verlag, New York, NY.

Tian, J., Paz, A. & Pearl, J. (1998), Finding minimal D-separators, Technical Report 980007, Cognitive System Laboratory, Computer Science Department. University of California, Los Angeles, CA.

Verma, T. & Pearl, J. (1990), Causal networks: Semantics and expressiveness, *in* R. Shachter, T. Levitt, K. L. & J. Lemmer, eds, 'Uncertainty in Artificial Intelligence 4', North Holland, Amsterdam, Netherlands, pp. 69–76.

CHAPTER 3

HANDLING MANIPULATED EVIDENCE*

3.1 Introduction

Bayesian Networks (BNs) have recently been advocated to establish the overall dependence between hypotheses and observable random variables in an investigation case (Dawid & Evett 1997, Garbolino & Taroni 2002).

In fact, the ability to highlight the essential relations among the variables makes a BN particularly useful in order to take trace and evaluate the probabilistic effects of the evidence on the unobservable hypotheses under debate. Moreover, a BN is highly modular in nature, so that it easily allows to increase the model whenever it is required, including relations with previously not considered variables.

Since the investigation is performed in a more formal way as compared to the usual practice, a possible subtle drawback in the use of BN-assisted investigations consists in overconfidence in the results obtained. The most treacherous possibility occurs if manipulated evidence is introduced, i.e. if observations not genuinely arisen from the context are produced by someone to mislead the investigator. Examples of cases where police is induced to focus towards a person different from the culprit include false testimonies, blood traces left intentionally by someone, and many others.

The aim of this work is to build a model that can help the investigator handle some possibly manipulated variables, in order to produce an updating of the probability that the evidence under suspicion is in fact genuine or manipulated, as well as the posterior distribution of the relevant hypotheses.

The structure of the paper is the following: first in section 2, we show how BNs can be used to formalize an investigation case, following its development. We presume that an expert, the investigator, guides the construction of the genuine models presented.

*This paper was published as a Working Paper in the Department of Statistics 'G. Parenti' series: Baio, G. & Corradi, F. (2004), Handling Manipulated Evidence. *University of Florence Press, Florence, Italy, Working Paper* no. 2004/13.

In section 3 we describe a methodology, derived from the original BN structures provided by the expert, that takes into account the possibility of manipulated evidence. This new representation is built modifying the original graphical structure, according to the intervention model originally introduced into the statistical literature by Pearl (1993) and Spirtes et al. (1993), in the causal inference framework.

Finally, we consider the situation where more than one pieces of evidence whose origin is unknown are possibly manipulated. Comparing all the models derived from different configurations, we show in section 4 how to detect a criminal plan aimed at misleading the investigation. In section 5, we discuss the most relevant implications of this work.

3.2 Modelling genuine evidence

In this section we show how an investigation can be translated into a BN framework. We propose an example of increasing complexity, according to the information that successively becomes available to the investigator.

Unlike other works, such as those of Dawid & Evett (1997) and Garbolino & Taroni (2002), our focus is not in defining a collection of formulæ to be used in the calculation of the posterior probabilities of the relevant hypotheses and/or the associated weight of evidence.

In fact, despite the practice to highlight the role of some epistemic and population probabilities is quite common and formally attractive in Forensic Science, in our opinion this approach proves of limited help, when the practitioner has to face the solution of his/her own case, which in general is slightly different from the examples provided.

On the contrary, following the suggestions of Lindley (2000), we rather aim at providing some indications to translate a real investigative case into a BN, and give less importance to the computational aspects, since efficient algorithms are freely available.

Table 3.1 presents a synopsis of the relevant types of variables for a case and their possible relations, which we will be using through all the paper.

In this work, we focus on a single hypothesis, such as '*is the suspect guilty?*', which takes on the values yes or no.

In general, a hypothesis represents a state of nature, which is not observable, but influences probabilistically some of the other relevant variables, and is usually the main object of the inference.

For this reason, in a BN a hypothesis node H is typically represented as a *root* of the graph, i.e. $\text{pa}(H) = \emptyset$. Nevertheless, it is possible that a hypothesis H_1 is specified as a function of a more general conjecture

Unobservable	*Type of variables*	*Symbol*	*Characteristics*
	• Working hypotheses	$\mathbf{H}$	$\text{pa}(H_i) \subseteq \emptyset \cup \{\mathbf{H} \setminus H_i\}$, for each $H_i \in \mathbf{H}$
	• Data generating model	$\mathbf{M}$	$\text{pa}(M_i) = \emptyset$ for each $M_i \in \mathbf{M}$
Observable	*Type of evidence*	*Symbol*	*Characteristics*
Clear origin	• Specification	$\mathbf{S}$	$\text{pa}(S_i) \nsubseteq \{\mathbf{M} \cup \mathbf{H} \cup \mathbf{T}\}$ for each $S_i \in \mathbf{S}$
	• Control	$\mathbf{K}$	$\text{pa}(K_i) \subseteq \{\mathbf{H} \cap \mathbf{T}\} \cup \{\mathbf{H} \cap \mathbf{T} \cap \mathbf{S}\}$ for each $K_i \in \mathbf{K}$
	• Natural	$\mathbf{N}$	$\text{pa}(N_i) \subseteq \{\mathbf{H}\} \cup \{\mathbf{H} \cap \mathbf{S}\}$ for each $N_i \in \mathbf{N}$
Unclear origin	• Possibly manipulated (treatments)	$\mathbf{T}$	$\text{pa}(T_i) \subseteq \{\mathbf{H} \cap \mathbf{M}\} \cup \{\mathbf{H} \cap \mathbf{M} \cap \mathbf{S}\}$, for each $T_i \in \mathbf{T}$

Table 3.1: Summary of the relevant variables in a case and their characteristics

H_0, and hence $H_0 \in \text{pa}(H_1)$. Garbolino & Taroni (2002) describe a set of archetypical situations in which some hypotheses are related to one another and evaluated in light of the observation of a single piece of evidence.

3.2.1 One single piece of evidence

Suppose that a crime is committed. A witness testifies to have seen an individual shooting a policeman during a robbery. Next, a suspect is individuated. A possible BN representation of this problem is that depicted in Figure 3.1 (the figures reported herein are absolutely fictional and merely serve as an explicative tool – all the calculation were performed using and modifying the `Matlab` package `BNets`, by Kevin Murphy).

Figure 3.1: An example of DAG

The variable H expresses the working hypothesis. The variable T represents the witness testimony and its possible values are $T = t_1$ in case the witness declares to recognise the suspect as the individual he has seen, and

$T = t_2$ in case he does not. At this stage, the testimony is not under suspicion, so that T is classified as a natural evidence, whence, according to Table 3.1, the relevant variables are grouped in the sets $\mathbf{H} = H$ and $\mathbf{N} = T$.

The investigator may define the Conditional Probability Table (CPT) for the variable H, depending on the circumstances that led to the identification of the suspect. Alternatively, $\Pr(H = yes)$ can be instrumentally set to 0.5, in case that rather than on the posterior distribution of H, the focus is on the weight of evidence $\mathbf{E}$.

In fact, using the Bayes theorem in terms of odds ratio we have:

$$\frac{\Pr(H = yes\,|\,\mathbf{E})}{\Pr(H = no\,|\,\mathbf{E})} = \frac{\Pr(\mathbf{E}\,|\,H = yes)}{\Pr(\mathbf{E}\,|\,H = no)} \times \frac{\Pr(H = yes)}{\Pr(H = no)},$$

where the left hand side represents the posterior odds, and the right hand side is the product of the weight of evidence (likelihood ratio) and the prior odds.

By the assumption of uniform prior distribution for H, the prior odds equal 1, so that in this case the weight of evidence is simply given by the posterior odds, which are directly available as a result from the propagation algorithm.

As for the testimony, suppose that the investigator can assess the CPT of Table 3.2. In case that the person under investigation is actually guilty, the investigator assigns a high probability, say 0.9, to the fact that the witness testifies to recognise him.

Conversely, when the hypothesis of guilt is not true, the probability that $T = t_1$ is low, for example 0.3. This assumption makes possible that the witness recognises the suspect, although he is innocent.

	$H = yes$	$H = no$
$T = t_1$	0.9	0.3
$T = t_2$	0.1	0.7

Table 3.2: The CPT for the testimony T, given the hypothesis H

Given the evidence $\mathbf{E}_1 = \{T = t_1\}$, i.e. that the witness claims he recognises the suspect, it is straightforward to update the hypothesis of guilt as $\Pr(H = yes\,|\,\mathbf{E}_1) = 0.75$, by means of the Bayes theorem.

3.2.2 More pieces of conditionally independent evidence

Usually, the investigator cannot be satisfied with just one evidence, and is likely to look for other observable variables that can confirm (or disprove) its suggestions.

The most natural choice is to look for other variables directly influenced by H, but conditionally independent on the other variables. A classical choice could be to check on the suspect alibi.

Suppose, for instance, that the suspect declares that he was home watching TV with his wife, who is then interrogated.

The variable W in Figure 3.2 represents the woman testimony, and takes on the values w_1 in case she declares that her husband was watching TV with her by the time that the crime was committed, and w_2 in case she does not provide him with a plausible alibi.

The graphical structure of Figure 3.2, known as *diverging connection* (Cowell et al. 1999), encodes the assumption that $W \perp\!\!\!\perp T \mid H$, i.e. that the distribution of the alibi is independent on the testimony, given that the value of the hypothesis H is actually known.

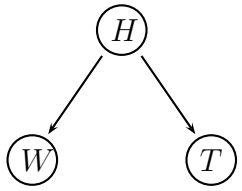


Figure 3.2: A working hypothesis H on a suspect's guilt, a witness testimony T, and the statements of the suspect's wife, W

According to the representation of Table 3.1, the variables are then groped in the sets $\mathbf{H} = H$ and $\mathbf{N} = \{T, W\}$. The CPT provided for the variable W from the investigator is shown in Table 3.3. Since the woman does not recall the exact time that her husband got home that night, the investigator assigns a positive (though small) probability to the event that yet being actually guilty, the suspect got back to his home after committing the crime.

	$H = yes$	$H = no$
$W = w_1$	0.02	0.80
$W = w_2$	0.98	0.20

Table 3.3: The CPT for the suspect's alibi W, given the hypothesis H

The Bayes Theorem can be directly invoked to solve the inferential issue, but even in this simple case, a specialised BN algorithm, such as the *Junction Tree* (Cowell et al. 1999) can be used to accomplish the calculations more straightforwardly.

The BN proposed in Figure 3.2 is well known in the statistical literature as *Bayesian Naïve Classifier* (Friedman et al. 1997), and is now established as a successful competitor of the commonly used (multi)logit model.

Suppose that the the woman provides his husband with an alibi, i.e. $W = w_1$: the new available evidence is $\mathbf{E}_2 = \{T = t_1, W = w_1\}$, so that $\Pr(H = yes \,|\, \mathbf{E}_2) = 0.0698$.

3.2.3 Adding a control evidence

Since the conflict between the testimony and the alibi, the investigator needs to find other variables in order to check on them. In other words, given the testimony, the investigator seeks for a *control evidence*. Table 3.1 shows formally the definition of such variables.

The investigator notices a surveillance camera set at a cash dispenser just in front of the crime scene, and finds out that the CCTV video is available.

In this case, the random variable representing the video is different from the alibi variable, since it is not plausible to assume conditional independence with the testimony T. In fact, both the witness and the camera look at the same scene, although from different perspectives, whence it is necessary to establish a dependence structure between them.

The original BN can be modified accordingly, to take into account this new variable. A suitable graphical representation is that of Figure 3.3, where the variable A is the observation of the ATM surveillance video. Notice that in this case, the presence of the direct link between T and A is such that these two nodes are not independent, even in case H was known. This representation is an extension of the Bayesian Naïve Classifier model, which allows for correlation among observable variables.

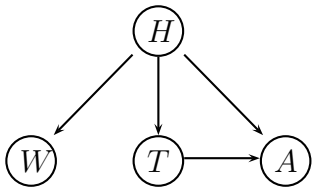


Figure 3.3: A working hypothesis H on a suspect's guilt, a witness testimony T, the wife statement W, and the video recorded by a CCTV of an ATM nearby the crime scene, A

Let the possible values for A be:

- a_1, if the suspect appears in the video less than half an hour before the crime was committed. In this case, it would be unlikely (although still possible) that he reached his home before the crime was committed;

- a_2, if the suspect is shown in the video more then half an hour before the crime. This time would easily allow him to reach his home before the time that the crime was committed.

The investigator assigns the probabilities of Table 3.4 to these events.

	$H = yes$		$H = no$	
	$T = t_1$	$T = t_2$	$T = t_1$	$T = t_2$
$A = a_1$	0.9	0.7	0.2	0.1
$A = a_2$	0.1	0.3	0.8	0.9

Table 3.4: The CPT for the variable A, given the testimony T and the hypothesis of guilt H

Suppose that the person under investigation appears in the video just 10 minutes before the estimated time of the crime. The BN updates the probability of guilt, given the evidence $\mathbf{E}_3 = \{T = t_1, W = w_1, A = a_1\}$ gathered by the investigator as $\Pr(H = yes \mid \mathbf{E}_3) = 0.2523$. This new evidence increases the posterior probability of guilt, although uncertainty remains on whether the suspect is actually the culprit of the crime. The two testimonies are in conflict, and the control evidence is not enough to explain away this situation.

3.2.4 Adding a specification evidence

The investigator is willing to gain more understanding of the problem. In order to do that, one possibility is to look for some *covariates*, or specification variables, in the terminology of Table 3.1.

Like a control evidence, these are variables that may be directly connected to a node in the set $\mathbf{N}$ (and more specifically in the set $\mathbf{T}$, as will be shown in section 3.3). However, for some reasons, it is easier or more natural to express this dependence such that a covariate is a parent of the node T, rather than a child as happens for the control evidence.

The role of a specification variable is typically that of reducing the uncertainty on a given node, in order to try to explain away possible conflicts generated by other non consistent evidence.

Suppose for instance that the investigator gets the sight of the witness tested, in order to check on his ability to recognise the suspect. The variable

V takes on the values v_1 if the witness has no visual defects, and v_2 in case he has. The new graphical representation is depicted in Figure 3.4.

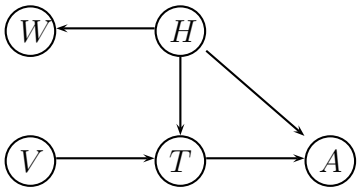


Figure 3.4: A more complicated situation. The new DAG describes the fact that the investigator checks also the visual capacity of the witness, as represented by the variable V

According to the classification of Table 3.1, the variables involved in the problem are now grouped as $\mathbf{H} = H$, $\mathbf{N} = \{T, W\}$, $\mathbf{K} = A$ and $\mathbf{S} = V$.

The investigator could define a prior distribution for the variable V, for instance using some population based statistics, i.e. $\Pr(V = v_1) = 0.7$ and $\Pr(V = v_2) = 0.3$. However, this is not essential: in fact, the value assumed by V is always known to the investigator, so that this random node becomes essentially deterministic (degenerate) and no inference is required about it.

On the contrary, it is necessary that the CPT of the variable T is modified in order to take into account the new situation, as in Table 3.5.

	$H = yes$		$H = no$	
	$V = v_1$	$V = v_2$	$V = v_1$	$V = v_2$
$T = t_1$	0.99	0.65	0.25	0.05
$T = t_2$	0.01	0.35	0.75	0.95

Table 3.5: The CPT for the variable T, given its parents V and H

Notice that the CPT for the testimony is now slightly different from that depicted in Table 3.2, because of the new variable that is made available. Hence, the information status of the investigator is changed and so are his conjectures.

Suppose that the result of the visual test is v_1: in this case, the evidence would be $\mathbf{E}_4 = \{T = t_1, W = w_1, A = a_1, V = v_1\}$ and $\Pr(H = yes \mid \mathbf{E}_4) = 0.3082$. As compared to the previous section, the probability of guilt is increased by the knowledge of the positive visual test. However, the investigator is uncomfortable with the results obtained, as two pieces of evidence tend to incriminate the suspect, whereas another one tends to acquit him.

Moreover, the suspect claims to have been framed, and that in fact, yet having cashed some money at the ATM only a few minutes before the crime was committed, he reached his home quite quickly that night, and the witness declared to recognise him only in order to make him considered guilty. How should the investigator handle this situation?

3.3 Handling manipulated evidence

3.3.1 Modelling external interventions on the observed evidence

The case of external intervention on a variable within a stochastic system is one of the paradigms of causal inference (Holland 1986). Despite many scholars are still working with different approaches, a point of agreement is that causality mechanisms are mimicked by the presence of external interventions, which modify the natural dynamics of the stochastic system under study.

Two major contributions to the literature are those of Spirtes et al. (1993) and Pearl (1993), among the first to apply BNs to the study of causality. In order to do so, a new semantic is defined that takes into account the fact that one or more variables are subjected to intervention.

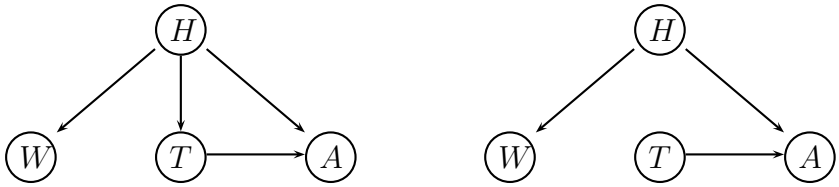
The central idea is that any direct link between the intervened node and its parents has to be removed. If the link $H \rightarrow T$ is suggestive of a causal mechanism, there is no point in modifying H after that T is set to a given value, since the observation of $T = t$ is not attributable to that causal mechanism, but rather to the intervention.

If we make reference to the forensic case, this feature is quite relevant. In fact, the intervention model is such that the knowledge that the evidence is not genuine is critically taken into account, leading to a more appropriate inference on the unobservable hypothesis.

Conversely, the descendants remain dependent on T, either it arose naturally or by intervention. This circumstance has a special relevance when a descendant of the possibly manipulated node is also in the set de(H), and its origin is not under suspicion (see Figure 3.5).

Under the natural model, the observation of $T = t$ modifies the distribution of A both directly and through updating the distribution of the unobservable node H. Therefore, the most likely value of A is the one that is most consistent with *a)* the observed value of T, and *b)* the value of H induced by $T = t$.

However, if T did not arise genuinely, the distribution of A is only modified by T itself, as the distribution of H is not updated by the available



a) Natural model b) Intervention model

Figure 3.5: The DAG representation of the external intervention. In case the evidence arose by means of external manipulations, the direct connections between the intervened node T and its parent is removed. The rest of the graph is unchanged

evidence, since T and H are not directly connected in the intervention model.

Consequently, using the natural model when the evidence is not genuine assigns a higher probability to values of A that in fact are not as likely to occur. Hence, the value of A that becomes available after observing T can be in conflict with the previous evidence, suggesting the possibility of manipulation.

For instance, when the genuine model of Figure 3.5a holds, the value $A = a_2$ becomes unlikely after observing $T = t_1$: $\Pr(A = a_2 \mid T = t_1) = 0.175$. On the contrary, using the intervention model of Figure 3.5b, the same value becomes much more plausible: $\Pr(A = a_2 \mid T = t_1) = 0.45$.

Dawid (2002) proposed a unified representation of the problem, using a decision theoretic approach based on the Augmented DAG (ADAG). This is a graphical model in which a possibly manipulated variable T is explicitly associated to an external *intervention* variable, F_T, which is used to rule its demeanour. Such a model is depicted in Figure 3.6.

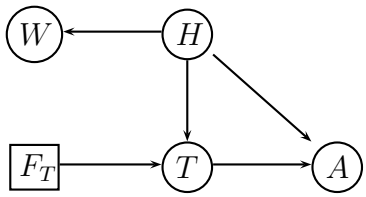


Figure 3.6: The ADAG representation of the intervention model

The possible external intervention is modelled as a *decision variable*, represented as a square. The variable F_T takes on the elements of the set $\{\mathcal{T} \cup \emptyset\}$, where $\mathcal{T}$ is the set of values that the possibly intervened node T may assume.

Unlike a random node, F_T is not associated to a CPT, as its state is always decided (known) by the experimenter. Therefore, it serves as a switch and it is used to allow the experimenter to activate a given scenario. When $F_T = \emptyset$, then the intervention is void, and hence T is a random variable governed by its conditional probability distribution.

Conversely, when $F_T = t, t \in \mathcal{T}$, then an intervention occurred. As a result, T becomes a degenerate variable, whence $\Pr(T = t \mid \text{pa}(T)) = 1$, for every configurations of the variables in pa(T). As required, in case of external intervention, the parents are not updated by T.

Back to the simple example of § 3.2.3, if the observed evidence was genuine, then F_T would be set to $\emptyset$, and the DAG representation implicit in the ADAG of Figure 3.6 would be the same as that depicted in Figure 3.5a.

On the contrary, should the investigator believe that the testimony is not genuine, then F_T would be set to the value t_1, and so would T. However, in this case, the knowledge of T should not update the CPT of its parent H. In other words, in case that $F_T \neq \emptyset$, the correspondent DAG is modified as in Figure 3.5b.

The use of the ADAG translates into a more compact representation of the problem, since both the situations are handled by the intervention node. The CPT of the variable T is then built as in Table 3.6 and comprises both the natural and the intervention cases.

	$F_T = \emptyset$		$F_T = t_1$		$F_T = t_2$	
	$H = yes$	$H = no$	$H = yes$	$H = no$	$H = yes$	$H = no$
$T = t_1$	0.9	0.3	1	1	0	0
$T = t_2$	0.1	0.7	0	0	1	1

Table 3.6: The CPT of the possibly manipulated variable T. When the evidence is genuine, the CPT is that specified by the expert; in case of manipulation, the distribution of the variable becomes independent on the other parent, H, and degenerate to the value specified by F_T

Dawid's model has been originally used to deal with standard causal inference problems, where the objective is to estimate the effect of a 'treatment' T over a 'response' A, discarding all the factors, defined as 'potential confounders', which can generate spurious relations between them.

In the situation of Figure 3.6, a standard causal model would use the observations of the treatment T and of the confounder H to infer the desired causal effect on the unobserved response A.

The graph of Figure 3.6 encodes the constraints suggested by Dawid

(2002) that allow the identifiability of such causal effect:

$$F_T \perp\!\!\!\perp H, \tag{3.1}$$

$$A \perp\!\!\!\perp F_T \,|\, T, H. \tag{3.2}$$

By condition (3.1), Dawid's model assumes that the probability distribution of the observed confounder H must not depend on how T arose. In other words, if it is known that an intervention occurred on T, the distribution of H must not change with the value set for F_T. Moreover, this distribution must remain the same as the case in which the evidence is certainly genuine ($F_T = \emptyset$), but T has not been observed yet.

The graphical representation of the relations among F_T, H and T is called a *v-structure* (Cowell et al. 1999), and implies that yet being marginally independent, F_T and H become conditionally dependent, after the observation of their common child T. However, because of the particular CPT assigned to the node T, in the intervention model where $F_T \neq \emptyset$, although T is implicitly set to the value of F_T, the link between H and T is removed, whence H and F_T are both marginally and conditionally independent.

Assumption (3.2) instead indicates that the knowledge of T and H is *all* that is needed for A to be independent on F_T, in which case the response is not modified by the way that the treatment arose. This situation basically amounts to the fact that the causal mechanism that relates T to A is conveniently explained by the variables in the set $\{T, H\}$, so that the differences in the response A can be directly attributable to T, once H is made available. For this reason, a variable H that verifies (3.1) and (3.2) is called a *sufficient covariate* (Dawid 2002).

As compared to the standard case, the objective of our analysis is reversed, being to evaluate how the unobservable variable H is modified whether T is genuine or not, after observing the available evidence, including A.

Assumption (3.1) is straightforward, as it makes sense to assume that given that the testimony is manipulated, no matter what the witness declares, the investigator's uncertainty over the hypothesis of guilt will remain the same.

Assumption (3.2) simply means that the knowledge of the actual value of H and of the testimony is sufficient to guarantee that the control evidence A has a clear origin with respect to the testimony, being independent on F_T without the need of any further information.

Obviously, the actual value of H is hidden, and its estimation is the objective of our analysis. However, assuming the validity of conditions (3.1)

and (3.2), the investigator takes the responsibility to ensure the absence of other unmeasured factors that can be connected to both A and T, which could confound the inference on H. This feature entitles the investigator to use A in order to check on T.

If the evidence is genuine ($F_T = \emptyset$), the observation of both T and A updates the distribution of H, whereas in the intervention case only the control evidence can modify the distribution of the hypothesis of guilt.

For instance, if after the suspect claim of having been framed, we consider the possibility of manipulation of T, it is possible to regroup the nodes in the following way: $\mathbf{H} = H$, $\mathbf{N} = W$, $\mathbf{T} = T$, $\mathbf{K} = A$. The origin of the variable T is now unclear to the investigator; therefore, it is included in the set $\mathbf{T}$, rather than in the set $\mathbf{N}$, as in the § 3.2.3.

In the intervention case, the observed evidence is $\mathbf{E}_5 = \{F_T = t_1, W = w_1, A = a_1\}$. Using a propagation algorithm on the ADAG of Figure 3.6, we obtain that $\Pr(H = yes \mid \mathbf{E}_5) = 0.1011$, whereas, by definition, using the natural model for which the evidence is $\{F_T = \emptyset, T = t_1, W = w_1, A = a_1\}$ the posterior probability of guilt would be 0.2523, the same inference described in § 3.2.3.

3.3.2 Model assessment: the probabilistic evaluation of the intervention node

In the previous section, we explored the features of the ADAG representation, which allows the investigator to regard properly the two cases where it is known that the evidence is genuine or manipulated.

All the same, the investigator would be even more interested in the possibility of evaluating probabilistically the two competing models:

- m_1: the unclear origin evidence T is in fact genuine;
- m_2: the unclear origin evidence T is manipulated,

conditionally on all the observed variables.

To this aim, it is necessary to define a further specialised version of the ADAG representation, as the one depicted in Figure 3.7. We term this graph *Model Assessment DAG* (MADAG), and we characterise the model node as a dashed circle. In this case, we define a new random variable M_T, which takes into account the two possibilities described above.

As depicted in Table 3.1, the model nodes are considered as roots of the graphical representation. This assumption is useful to characterise them as unobservable conjectures about the data generating process, whose uncertainty is updated by the evidence.

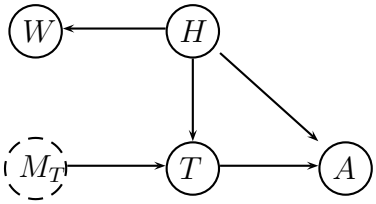


Figure 3.7: The Model Assessment DAG (MADAG) representation of the problem. The intervention node is now modelled as a random variable, rather than a decision node, in order to take into account the fact that a probabilistic evaluation is needed

Just like the intervention node F_T in the ADAG, the model node M_T acts on the possibly intervened node so that the update of its parents is avoided, in case of manipulated evidence (i.e. when model m_2 holds). Yet, since the investigator is in doubt whether the observed pieces of evidence are genuine or invalidated by some intervention, T must remain a random variable in either case.

With respect to the standard graphical representation, the use of the MADAG requires the following operations. First, the investigator must provide a prior CPT for the model nodes. Again, if the interest is in the evaluation of the weight of evidence with respect to the competing models, it is effective to choose $\Pr(M_T = m_1) = \Pr(M_T = m_2) = 0.5$. Obviously, in case that the investigator has different prior knowledge about the two models, it is easy to modify the CPTs accordingly.

Second, while the distribution under the natural model m_1 is provided by the expert, the one for the intervention model is to be defined so that consistency between the two regimes is maintained. If the origin of a piece of evidence is under suspicion, the investigator cannot update his uncertainty on which is the model that generated the data, by means of the observation of that node only. This condition can be re-expressed as:

$$\begin{aligned}\Pr(T = t \,|\, M_T = m_2) &= \mathrm{E}_{\mathbf{P}}\left[\Pr(T = t \,|\, M_T = m_1)\right] \\ &= \sum_{p \in \mathcal{P}} \Pr(T = t \,|\, \mathbf{P} = p, M_T = m_1) \Pr(\mathbf{P} = p) \quad (3.3)\end{aligned}$$

for any $t \in \mathcal{T}$. The set $\mathbf{P} = \{\mathrm{pa}(T) \setminus M_T\}$ includes all the parents of T except the model node, and the average $\mathrm{E}_{\mathbf{P}}$ is taken over all the possible configurations of the variables in $\mathbf{P}$, indicated by the set $\mathcal{P}$.

The right hand side of equation (3.3) is by definition the marginal probability distribution of the variable T, under the model m_1. Therefore, condition (3.3) ensures that $\Pr(T \mid M_T = m_1) = \Pr(T \mid M_T = m_2)$, for each configuration of the observed parents.

This result is consistent with the representation of the problem: in case the origin of the variable T is unclear to the investigator, in absence of additional evidence, its observation cannot modify his prior belief on which of the two models generated the data. In fact, the Bayes Theorem expressed in terms of odds ratio states that:

$$\frac{\Pr(M_T = m_1 \,|\, T = t)}{\Pr(M_T = m_2 \,|\, T = t)} = \frac{\Pr(T = t \,|\, M_T = m_1)}{\Pr(T = t \,|\, M_T = m_2)} \times \frac{\Pr(M_T = m_1)}{\Pr(M_T = m_2)},$$

which by (3.3), implies that:

$$\frac{\Pr(M_T = m_1 \,|\, T = t)}{\Pr(M_T = m_2 \,|\, T = t)} = \frac{\Pr(M_T = m_1)}{\Pr(M_T = m_2)}.$$

The model m_2 can be seen as *nested* within m_1 (cfr. O'Hagan 1994): the latter includes the former, and they differ only in the fact that m_2 does not depend on the variables in the set $\mathbf{P}$, whereas m_1 does. Consequently, they are marginally equivalent, marginalisation being over that set.

The definition of the CPT for the variable T essentially renders the testimony independent on the model node M_T, even if this property cannot be read off by the inspection of the graph. Moreover, despite condition (3.2) is assumed to hold, since H is unobservable $A \not\perp M_T \mid T$, which allows to update also the uncertainty over the data generating process, when the control evidence is made available.

In the example of § 3.2.3, $\mathbf{P} = \{H\}$ whence $\mathcal{P} = \{yes, no\}$, and the distribution of T under the natural regime m_1 is that of Table 3.2. Therefore, applying (3.3), the coherent CPT of the variable T is that shown in Table 3.7.

	$M_T = m_1$		$M_T = m_2$	
	$H = yes$	$H = no$	$H = yes$	$H = no$
$T = t_1$	0.9	0.3	0.6	0.6
$T = t_2$	0.1	0.7	0.4	0.4

Table 3.7: The CPT of the possibly manipulated variable T. When the evidence is genuine, the CPT is that specified by the expert; in case of manipulation, the distribution of the variable becomes independent on the other parent, H, and on average the two distributions are equivalent, by definition

Given the observed evidence $\mathbf{E}_3 = \{T = t_1, W = w_1, A = a_1\}$, the probabilities for the unobservable variables are updated as $\Pr(H = yes \mid \mathbf{E}_3) = 0.1579$ and $\Pr(M_T = m_1 \,|\, \mathbf{E}_3) = 0.3754$.

As compared to the inference obtained using the ADAG, the results derived here are subjected to an additional source of variability, i.e. that related to the model itself. The probability of guilt is a mixture of the natural and of the intervention case, with weights given by the posterior probability of the model node.

In general, it would be appropriate to check on the possibly manipulated node by means of several pieces of evidence. This situation could be easily handled by extending the MADAG of Figure 3.7 with other nodes K of clear origin.

3.3.3 Model assessment for one manipulated evidence with covariates

Let us consider the situation of § 3.2.4, now with the variable T supposed to have an unclear origin to the investigator. The new graphical representation is depicted in Figure 3.8.

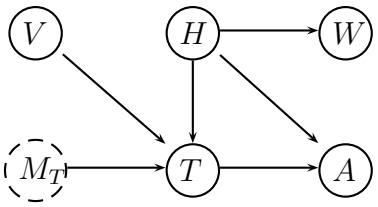


Figure 3.8: The MADAG describes the fact that the investigator checks also the visual capacity of the witness, as represented by the variable V

According to the classification of Table 3.1, the variables involved in the problem are now grouped as $\mathbf{H} = H$, $\mathbf{N} = W$, $\mathbf{T} = T$, $\mathbf{K} = A$, $\mathbf{S} = V$ and $\mathbf{M} = M_T$.

	$M_T = m_1$				$M_T = m_2$			
	$H = yes$		$H = no$		$H = yes$		$H = no$	
	$V = v_1$	$V = v_2$	$V = v_1$	$V = v_2$	$V = v_1$	$V = v_2$	$V = v_1$	$V = v_2$
$T = t_1$	0.99	0.65	0.25	0.05	0.62	0.35	0.62	0.35
$T = t_2$	0.01	0.35	0.75	0.95	0.38	0.65	0.38	0.65

Table 3.8: The CPT for the variable T, given its parents V, H and M

The natural distribution for the node T is that depicted in the left half

of Table 3.8. Moreover, applying condition (3.3), we obtain that:

$$\Pr(T = t \,|\, M_T = m_2) = \sum_{h \in \mathcal{H}} \sum_{v \in \mathcal{V}} \Pr(T = t \,|\, H = h, V = v, M_T = m_1) \Pr(V = v \,|\, H = h) \Pr(H = h),$$

as V and H are marginally independent on M_T, from the graphical structure of the problem.

Since V is known to the investigator, $\Pr(V = v) = 1$, regardless on the value assumed by any other variables. Hence, we have that:

$$\Pr(T = t \,|\, M_T = m_2) \;=\; \sum_{h \in \mathcal{H}} \Pr(T = t \,|\, H = h, V = v, M_T = m_1) \Pr(H = h).$$

Back to the numerical example, in case $V = v_1$, then $\Pr(T = t_1 \,|\, M_T = m_2) = 0.62$, whereas in case $V = v_2$, then $\Pr(T = t_1 \,|\, M_T = m_2) = 0.35$, as reported in the right half of Table 3.8 – it is straightforward to calculate $\Pr(t = t_2 \,|\, V, M_T = m_2) = 1 - \Pr(t = t_1 \,|\, V, M_T = m_2)$.

In this calculation we treated the variable V as a *datum*. Although the investigator is able to provide a probability assessment about T, for all the possible values that V may take on, the specification evidence is always treated as known, and for this reason it is not marginalised off by the coherence procedure of condition (3.3).

Having observed the evidence $\mathbf{E}_4 = \{T = t_1, W = w_1, A = a_1, V = v_1\}$, the probabilities are updated as $\Pr(H = yes \,|\, \mathbf{E}_4) = 0.1723$ and $\Pr(M_T = m_1 \,|\, \mathbf{E}_4) = 0.3438$. Again, the results are changed, because of the new information status.

3.4 More complicated situations

3.4.1 More than one manipulated pieces of evidence

Let us now concentrate on the case where the investigator is not certain about the origin of more than one pieces of evidence. Given the increasing complexity of the case, and the growing conflict among the pieces of evidence, the investigator decides to regard the alibi represented by W as possibly manipulated as well.

In order to assess this testimony, the investigator also questions the suspect about the programme that his wife could refer to. In Figure 3.9, this is represented by the variable D, assuming the possible values d_1 in case he is able to describe the programme, and d_2 in case he is not.

According to the definitions of Table 3.1, in this case we have that $\mathbf{H} = H$, $\mathbf{T} = \{T, W\}$, $\mathbf{S} = V$ and $\mathbf{K} = \{A, D\}$. Moreover, since there are

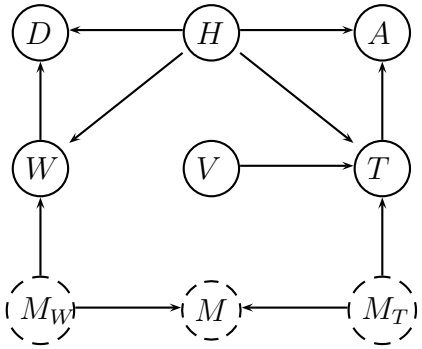


Figure 3.9: The case goes on: the investigator gathers more pieces of evidence, some of which have uncertain evidence. For each possibly manipulated variable, the investigator defines a model node, and seeks for a suitable control evidence

two variables in $\mathbf{T}$, the set $\mathbf{M}$ comprises the two model nodes M_T and M_W, which respectively rule the behaviour of the nodes T and W, just as described in the previous sections.

While the CPT for the variable T as a function of its own model node M_T is that of Table 3.5, the distribution of the variable W, in the natural and in the intervention case, as derived by the application of condition (3.3) is shown in Table 3.9. As for the variable D, suppose that the CPT provided by the investigator is that of Table 3.10.

	$M_W = m_1$		$M_W = m_2$	
	$H = yes$	$H = no$	$H = yes$	$H = no$
$W = w_1$	0.02	0.80	0.41	0.41
$W = w_2$	0.98	0.20	0.59	0.59

Table 3.9: The CPT of the possibly manipulated wife's testimony, under the genuine and the intervention model

	$H = yes$		$H = no$	
	$W = w_1$	$W = w_2$	$W = w_1$	$W = w_2$
$D = d_1$	0.15	0.01	0.90	0.10
$D = d_2$	0.85	0.99	0.10	0.90

Table 3.10: The CPT for the variable D, representing whether the suspect is able to describe the TV programme referred to by his wife or not

Finally, in order to analyse the whole case, it is possible to define a further model node, M, which takes into account the combinations of the

two data generating models for the possibly manipulated variables, taking on the following four values:

- m_1: both T and W are genuine;
- m_2: W has been manipulated, whereas T is genuine;
- m_3: W is genuine and T has been intervened;
- m_4: both T and W have been manipulated.

The CPT of the node M is easily identified as that depicted in Table 3.11. Since the nature of M is essentially deterministic, we do not include it in the set **M**, and it is only used as an instrumental node to allow easier calculations.

	$M_T = m_1$		$M_T = m_2$	
	$M_W = m_1$	$M_W = m_2$	$M_W = m_1$	$M_W = m_2$
m_1	1	0	0	0
m_2	0	1	0	0
m_3	0	0	1	0
m_4	0	0	0	1

Table 3.11: The CPT for the model node M, upon varying the model nodes for the two possibly manipulated variables T and W

The structure of Figure 3.9 encodes the assumption that the observation of only one possibly manipulated evidence is not able to update the prior knowledge on the data generating model.

However, when both the nodes in **T** are made available, the BN establishes an undirect connection between them, via the nodes H, D and A. Consequently, observing T and W does modify the probabilities of the model nodes, according to how consistent the two pieces of evidence are. Besides, the observation of the control evidence allows to produce a sharper update of the probabilities of the unobservable variables.

Suppose that the investigation leads to the following observed evidence $\mathbf{E}_6 = \{T = t_1, W = w_1, A = a_1, V = v_1, D = d_2\}$, i.e.:

- The witness testifies that he recognises the suspect as the man he saw on the crime scene;
- The suspect wife testifies that she and her husband were watching TV together;

- The suspect is shown on the video recorded at the ATM CCTV, just in front of the crime scene, 10 minutes before that the crime was committed;

- The visual capacity of the witness is tested positively;

- The suspect fails to describe the TV programme that his wife claimed they were watching together.

Using the BN of Figure 3.9, the investigator obtains that the posterior probability of guilt is $\Pr(H = yes \mid \mathbf{E}_6) = 0.9277$. As for the model that generated the observed evidence, the results are the following: $\Pr(M = m_1 \mid \mathbf{E}_6) = 0.0403$, $\Pr(M = m_2 \mid \mathbf{E}_6) = 0.5509$, $\Pr(M = m_3 \mid \mathbf{E}_6) = 0.0507$ and $\Pr(M = m_4 \mid \mathbf{E}_6) = 0.3581$.

The most likely model, given $\mathbf{E}_6$ is $M = m_2$ indicating that the wife testimony is not genuine. Should the investigator not take into account the possibility of observing manipulated pieces of evidence, starting from the same CPTs as for the genuine model, his inference would be that $\Pr(H = yes \mid \mathbf{E}_6) = 0.6588$. In fact, in this case, the testimony of the wife would be treated as genuine and, even if the suspect fails to recall the TV programme, the alibi provided by the woman would decrease the posterior probability of guilt.

Conversely, if given these findings the investigator accepts the model $M = m_2$ as the one that generated the observed data, the estimated probability of guilt would be 0.9872, leading the investigator to incriminate the suspect with even more strength.

3.4.2 Evaluating two pieces of evidence by comparison

Finally, we consider a very relevant situation that can occur during an investigation: the case where two different pieces of evidence need be compared, before the investigator can assess their actual relevance. A typical example is that of a blood trace left on a crime scene. A sample of the suspect DNA is analysed, but neither the crime scene trace, the 'crime sample', nor the suspect sample are relevant *per se*. Only if the two samples match (i.e. the DNA is of the same type), an evidence against the suspect is provided.

Even in this situation, the investigator may reckon that one of the two samples has been manipulated, and can easily specialise the BN representation of the investigations, in order to take into account properly of this feature.

Considering again the previous example, suppose further that a gun is found in the suspect's house, which is compatible with the bullets found on the crime scene. However, given the complexity of the case, the investigator reckons that someone else could possibly have put the gun in the suspect's house. For this reason, he gets the suspect tested with the paraffin glove method. A BN representation of this problem could be that depicted in Figure 3.10.

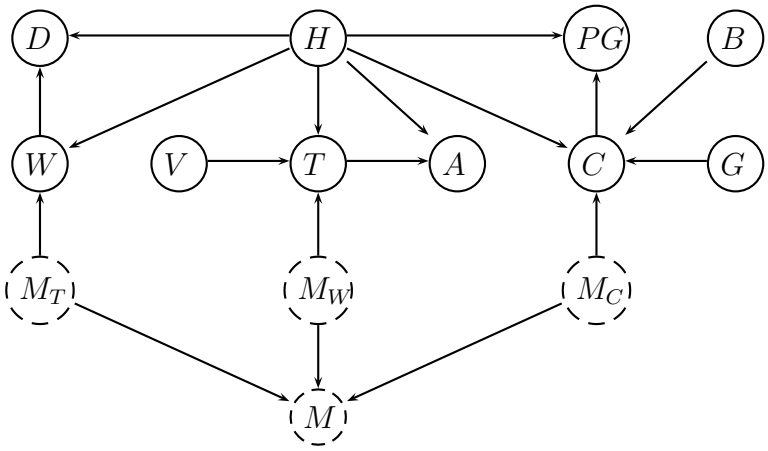


Figure 3.10: The MADAG for the comparison of two pieces of evidence. The nodes B and G are evaluated jointly, into the node C

The nodes B and G stand respectively for the bullets found on the crime scene and the gun found at the suspect's house. If the two are compatible, the variable C, representing the 'compatibility match', takes on the value c_1, whereas when they are not, it takes on the value c_2, with a probability defined by the investigator, may be according to some population data.

The variable PG indicates the result of the paraffin glove test, taking on the values pg_1, in case the test is positive, and pg_2 in case the test is negative.

According to Table 3.1, the variables are grouped as follows: $\mathbf{H} = H$, $\mathbf{M} = \{M_T, M_W, M_C\}$, $\mathbf{T} = \{T, W, C\}$, $\mathbf{K} = \{A, D, PG\}$ and $\mathbf{S} = \{V, B, G\}$.

Yet being more complicated, the situation depicted in Figure 3.10 is not too different from that described in the previous sections.

In this case, it is possible to identify 8 different data generating models, from the case where all the variables in $\mathbf{T}$ arose genuinely, to that where all have been manipulated, through the possible combinations (one manipulated and two natural; one natural and two manipulated variables), as depicted in Table 3.12.

M	T	W	C
m_1	natural	natural	natural
m_2	natural	manipulated	natural
m_3	natural	natural	manipulated
m_4	natural	manipulated	manipulated
m_5	manipulated	natural	natural
m_6	manipulated	manipulated	natural
m_7	manipulated	natural	manipulated
m_8	manipulated	manipulated	manipulated

Table 3.12: The possible models generating the observed evidence

Given the prior CPTs for the variables involved in the BN of Figure 3.10, it would be possible to make inference on **H** and **M** simply applying the calculation strategies described in the previous sections.

Suppose, for example, that the investigator is able to define the natural distribution of the node C, as in the upper half of Tables 3.13. Applying condition (3.3) it is straightforward to derive the intervention distribution, as in the lower half of Table 3.13.

	$M_C = m_1$							
	$H = yes$				$H = no$			
	$G = g_1$		$G = g_2$		$G = g_1$		$G = g_2$	
	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	0.9	0.3	0.3	0.9	0.8	0.5	0.5	0.8
$C = c_2$	0.1	0.7	0.7	0.1	0.2	0.5	0.5	0.2
	$M_C = m_2$							
	$H = yes$				$H = no$			
	$G = g_1$		$G = g_2$		$G = g_1$		$G = g_2$	
	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	0.85	0.40	0.40	0.85	0.85	0.40	0.40	0.85
$C = c_2$	0.15	0.60	0.60	0.15	0.15	0.60	0.60	0.15

Table 3.13: The CPT for the variable C, given its parents B, G and H

The distribution of the control variable PG is that depicted in Table 3.14.

If the evidence was $\mathbf{E}_7 = \{T = t_1, W = w_1, A = a_1, V = v_1, D = d_2, Pg = pg_1, G = g_1, B = b_1\}$, then the posterior probability of guilt would be $\Pr(H = yes \mid \mathbf{E}_7)$= 0.9693. As for the models, the result is shown in Figure 3.11. If the possibility of manipulation was not considered, the posterior probability of guilt would be 0.8306.

	$H = yes$		$H = no$	
	$C = c_1$	$C = c_2$	$C = c_1$	$C = c_2$
$Pg = pg_1$	0.95	0.85	0.45	0.05
$Pg = pg_2$	0.05	0.15	0.55	0.95

Table 3.14: The CPT for the variable Pg, representing the result of the paraffin glove test

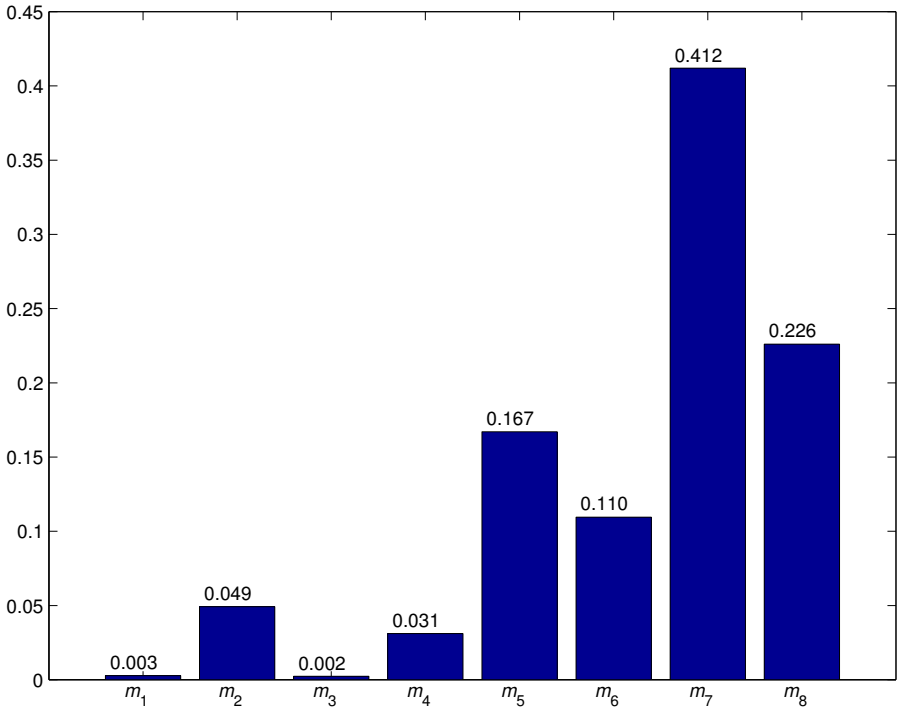


Figure 3.11: The posterior probability distribution for the model node, given the evidence $\mathbf{E}_7$

Given the prior probabilities asserted by the investigator, and the pieces of evidence gathered, the most likely models are m_2 (T and C genuine, and W manipulated), and m_2 (T genuine, and W and C manipulated).

While the wife testimony can be regarded as not genuine, given this findings, the evidence C is less straightforward to interpret. In fact, for its very nature, it is not directly observed, but only 'induced' by the comparison of B and G. For this reason, its power on the model node M_C is lower, as compared to the other pieces of evidence in the set $\mathbf{T}$. Nevertheless, since B and G are compatible, the posterior probability that $C = c_1$ (given *all* the other available nodes) is 0.7742. This probably leads the investigator to regard the node C as genuine.

3.4.3 Synthesis of the case investigation

Table 3.15 shows an overall summary of the case evaluation, upon varying the different information status. We consider the situation when the investigator is not aware of the possibility that one or more evidence is manipulated, and compare it to the models based on the suitable MADAG representations.

In general, one can appreciate how the inference is changed when the possibility of manipulation is taken into account. In case number 5, we consider the situation of § 3.3.3, but we suppose that also W is subjected to manipulation. As compared to the results provided for case 4 (and in § 3.3.3), the MADAG model proves to provide a more consistent inference (W will turn out to be the most likely manipulated node). The weights of evidence are even more impressive.

Pieces of evidence gathered	*Posterior probability (weight of evidence) for* $H = yes$		
	Manipulable nodes	All genuine	Allow for manipulation
1. t_1	T	0.7500 (3.000)	0.6250 (1.667)
2. t_1, w_1	T	0.0698 (0.075)	0.0400 (0.041)
3. t_1, w_1, a_1	T	0.2523 (0.338)	0.1579 (0.187)
4. t_1, w_1, a_1, v_1	T	0.3082 (0.445)	0.1723 (0.208)
5. t_1, w_1, a_1, v_1	T, W	0.3082 (0.445)	0.7474 (2.958)
6. t_1, w_1, a_1, v_1, d_2	T, W	0.6588 (1.931)	0.9277 (12.83)
7. $t_1, w_1, a_1, v_1, d_2, pg_1, b_1, g_1$	T, W, C	0.8306 (4.903)	0.9693 (31.57)

Table 3.15: The posterior probability of guilt (in parentheses the associated weight of evidence) calculated using the natural models (without the possibility of accounting for not genuine evidence), and the MADAG models, upon varying the manipulable nodes

Finally, Table 3.16 depicts the weight of evidence for the hypothesis that each single unclear variable is in fact genuine, given different status of knowledge.

As appears clear, the evidence of W is likely to be not reliable (the highest posterior probability that it is genuine is only about 0.20, leading to a weight of evidence of 0.25 at most). In addition, it is interesting to study the progress of the posterior probability that $M_T = m_1$. As one can see, unless the node W becomes suspect (starting from case number 5), the evidence tends to suggest that T has been manipulated, discrediting the witness.

Pieces of evidence gathered	Manipulable nodes	*Weight of evidence for* $M_T{=}m_1$	$M_W{=}m_1$	$M_C{=}m_1$
1. t_1	T	1.0000	–	–
2. t_1, w_1	T	0.5244	–	–
3. t_1, w_1, a_1	T	0.6010	–	–
4. t_1, w_1, a_1, v_1	T	0.5239	–	–
5. t_1, w_1, a_1, v_1	T, W	1.1372	0.2527	–
6. t_1, w_1, a_1, v_1, d_2	T, W	1.4462	0.1001	–
7. $t_1, w_1, a_1, v_1, d_2, pg_1, b_1, g_1$	T, W, C	1.5310	0.0699	1.0404

Table 3.16: The weight of evidence for the model nodes, upon varying the informative status and starting from a uniform prior

However, once the investigator realises that W is possibly not genuine, then T is less and less in conflict with the other pieces of evidence, and the weight of evidence in favour of $M_T = m_1$ increases up to 1.53 (for better reading, Table 3.17 presents a synopsis of the variables used in the paper, with their possible values).

The sensitivity to the choice of the prior value for $\Pr(H = yes)$ has been investigated in Figure 3.12, with respect to case 7 of Table 3.15.

The posterior distribution $\Pr(H = yes \mid \mathbf{E})$ for case 7 of Table 3.15 has been calculated upon varying the prior value for $\Pr(H = yes)$, over the range $[0; 1]$. The differences in the posterior probability of guilt is noticeable.

The highest difference is of 0.43, and is reached when $\Pr(H = yes)$ is set a priori to 0.08, as shown in Figure 3.12.

3.5 Discussion

In this paper we showed a methodology to deal with unclear origin variables, within an investigation case. This possibility is ensured by the BN structure that we associated to the problem.

The very first advantage in using a BN is the fact that the overall judgement on the working hypothesis is articulated into each single relation among the variables.

At first sight, this could be perceived as a drawback, as the investigator may reckon that he is not able to provide a comprehensive assessment. However, in our opinion this only makes more straightforward the evaluation of the whole pieces of evidence available. Besides, it can become clear which variables need be investigated more thoroughly, before a sharpest

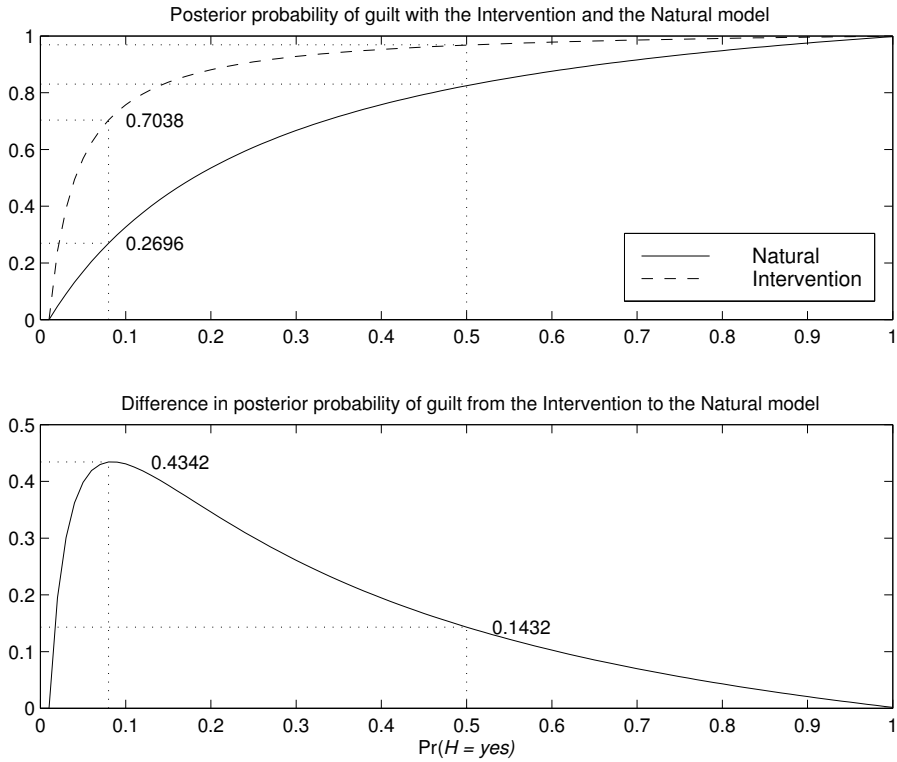


Figure 3.12: Probabilistic sensitivity analysis to the choice of prior distribution for the hypothesis $H = yes$, for case 7 of Table 3.15

opinion can be reached.

A second important feature of this method is that it explicitly models the presence of conflicting evidence. Some works in the statistical literature have focused on this matter (Jensen et al. 1991, Jensen 1995), defining a diagnostic statistics, which is able to detect possible conflicts between different pieces of evidence.

The model proposed in this paper provides an alternative way of quantifying inconsistencies in the evidence, as it directly calculates the posterior distributions for both the working hypotheses and the data generating models.

Third, unlike most standard Bayesian analysis of forensic data, the use of the weight of evidence is not particularly useful in the framework we presented here. In fact, the most important feature of this measure is its invariance to the choice of the prior distribution for the hypothesis of interest.

Since in our case, the working hypothesis is evaluated jointly with the model variable, the weight of evidence changes with the choice of the prior

Variable	Description	Values
H	Is the suspect guilty?	yes or no
T	The witness testimony	t_1 = the witness recognises the suspect as the man on the crime scene t_2 = the witness does not recognise the suspect as the man on the crime scene
W	The wife testimony	w_1 = the wife declares that she and her husband were watching TV but does not recall the time precisely w_2 = the wife declares that she was not with the suspect
A	The ATM video	a_1 = the suspect is shown in the video less than half an hour before the crime a_2 = the suspect is shown in the video more than half an hour before the crime
V	Witness's visual test	v_1 = the witness has no visual problem v_2 = the witness has visual problems
D	The TV program	d_1 = the suspect recalls the TV show d_2 = the suspect fails to recall the TV show mentioned by his wife
B	The bullets type	b_1 = the bullets found on the crime scene are of 'type 1' b_2 = the bullets found on the crime scene are of 'type 2'
G	The gun type	g_1 = the gun found in the suspect's house is of 'type 1' g_2 = the gun found in the suspect's house is of 'type 2'
C	Compatibility match	c_1 = the bullets and the gun are compatible c_2 = the bullets and the gun do not match
PG	Paraffin glove test	pg_1 = the suspect tests positive pg_2 = the suspect tests negative

Table 3.17: A synopsis of the variables introduced in the paper, along with their possible values

distributions for the model node. For this reason, the evaluation of the posterior distribution of the unobservable variables can be more relevant.

Finally, we reckon that this kind of modelling can also be applied to other research areas, such as Economics, in order to take into account the

possibility that some of the pieces of evidence, upon which decisions are taken, have in fact been manipulated. As for Statistics, this model could be applied in the detection and analysis of outliers.

Bibliography

Cowell, R., Dawid, P., Lauritzen, S. & Spiegelhalter, D. (1999), *Probabilistic Networks and Expert Systems*, Springer-Verlag, New York, NY.

Dawid, A. (2000), 'Causal Inference without Counterfactuals (with Discussion)', *Journal of the American Statistical Association* **95**, 407–448.

Dawid, A. (2002), 'Influence Diagrams for causal modelling and inference', *Statistical Review* **70**, 161–189.

Dawid, P. & Evett, I. (1997), 'Using a graphical method to assist the evaluation of complicated patterns of evidence', *Journal of Forensic Sciences* **42**, 226–231.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997), 'Bayesian Network Classifiers', *Machine Learning* **29**, 131–163.

Garbolino, P. & Taroni, F. (2002), 'Evaluation of scientific evidence using Bayesian networks', *Forensic Science International* **125**, 149–155.

Holland, P. (1986), 'Statistics and Causal Inference', *Journal of the American Statistical Association* **81**, 945–968.

Jensen, F. (1995), Cautious Propagation in Bayesian Networks, *in* P. Besnard & S. Hanks, eds, 'Proceeding of the Eleventh Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann Publishers, San Francisco, CA, pp. 519–528.

Jensen, F., Chamberlain, B., Nordahl, T. & Jensen, F. (1991), Analysis in Hugin of data conflict, *in* P. Bonissone, M. Henrion, L. Kanal & J. Lemmer, eds, 'Uncertainty in Artificial Intelligence 6', Elsevier Science Publishers, Amsterdam, Netherlands, pp. 323–329.

Lindley, D. (2000), 'The philosophy of Statistics', *The Statistician* **49**, 293–337.

O'Hagan, A. (1994), *Bayesian Inference, volume 2B Kendall's Advanced Theory of Statistics*, Arnold, London, UK.

Pearl, J. (1993), 'Comment: Graphical models, causality, and intervention', *Statistical Science* **8**, 266–269.

Pearl, J. (2000), *Causality*, Cambridge University Press, Cambridge, UK.

Spirtes, P., Glymour, C. & Scheines, R. (1993), *Causality, Prediction and Search*, Springer-Verlag, New York, NY.

CHAPTER 4

THE ECONOMIC EVALUATION OF INFLUENZA VACCINATION IN THE ELDERLY POPULATION: A MODEL BASED ON BAYESIAN NETWORKS AND INFLUENCE DIAGRAMS*

4.1 Introduction

Influenza infection is a major cause of illness, morbidity and mortality through out the world; the World Health Organisation estimates that influenza affects 5-15% of global population each year. The high-risk groups of influenza complication include mainly elderly and patients with cardiovascular or pulmonary disorder, and metabolic disease (diabetes). Institutionalised population is also considered at risk, because of the ease of viral transmission. The influenza vaccination is effective in reducing acute complications among high-risk patients, particularly in influenza-like illness (ILI), hospitalisation and mortality from all causes (Vu et al. 2002).

In terms of effectiveness, there is a clear and proved link between immunogenicity and protective effect (Potter 2001), with an inverse relation to higher titre of antibody and the rate of infection. It is widely accepted from clinical study that haemagglutination inhibition antibody levels could be used as a surrogate of protective activity.

In order to improve the immune response, a subunit vaccine adjuvated with MF59 was developed, and several studies demonstrated that this vaccine is more immunogenic than the other types of vaccine (subunit vaccine, slit virus and virosomal vaccine).

The burden of influenza on health care systems becomes highly relevant, since the increasing proportion of population aged over 65. As reported in Nichols (2001) influenza causes up to 300,000 excess hospitalisations and up to 40,000 excess deaths, among the high-risk population. Therefore, substantial public health implications arise from the decision maker perspective, when programming an annual vaccination strategy. In particular, the public decision maker faces a strategic problem when de-

*This paper has been presented to the International ISPOR Meeting, Arlington, VA, May 2004 and to the Italian Statistical Society Meeting, Bari, Italy, June 2004, and has been submitted for publication to the Journal of Health Economics in October 2004

ciding: *a)* whether to implement a vaccination campaign, and *b)* which vaccine(s) to prescribe. On the one hand, a vaccination program can result in an effective strategy, both from the financial and the clinical point of view, in light of the reduction of hospitalisations and deaths among the elderly. On the other hand, the cost effectiveness of this choice is highly dependent on some exogenous parameters, such as influenza attack rate (which is likely to vary from year to year), and coverage rate (i.e. the proportion of patients that are actually vaccinated). The decision should then be based on a mixture of pre-constituted opinions and empirical data, in order to take into account pros and cons, over the entire population.

The aim of this paper is then to build a decision model which allows the decision makers to evaluate the possible results under different scenarios, and to choose the decision associated to the highest expected utility, expressed in terms of incremental cost effectiveness ratio (*ICER*). In other words, first we calculate the total costs associated to different scenarios (do not vaccinate the reference population; vaccinate the reference population with a standard vaccine; vaccinate the reference population with the MF59 vaccine). Then, we combine these disutility measures with some effectiveness indicators, such as reduction in death, hospitalisation and access to other health resources (GP visits, pharmacological treatments), in order to obtain an economic evaluation of the different options.

The paper is then structured as followed: section 4.2 presents the statistical approach used in the definition of the model; section 4.3 describes the dataset, while section 4.4 presents the main results obtained. Finally, in section 4.5 the main conclusions are discussed.

4.2 Bayesian Networks, Influence Diagrams and decision problems

The decision model studied in this paper is based on the Bayesian Networks methodology (for example, see Jordan 2001, for a thorough description). A Bayesian Network (BN) is a graphical model that provides an alternative representation of the probabilistic relationships among a set of relevant variables. It proves to be very effective in presence of a complex system of variables, where the main goal of the statistical analysis is to make inference on the joint probability distribution.

Formally, a BN is represented by $\mathcal{B} = \{\mathcal{G}, \mathbf{\Theta}\}$, where $\mathcal{G}$ is a particular graphical model, encoding the probabilistic relationships of a set of variables, $\mathbf{X} = \{X_1, \ldots, X_n\}$, and $\mathbf{\Theta}$ is a vector, whose elements are the parameters of interest (Heckerman 1996, Jensen 1998). The vector $\mathbf{\Theta}$ is related to a set of local probability distributions, $\mathbf{P}$, which is used to de-

scribe the conditional independence among the variables in $\mathbf{X}$. For each of these variables, the local probability measure is exclusively a function of the parents (i.e. the nodes which the variable directly depends on), indicated by pa(X_i), and is independent on the other nodes. Hence, combining all the local probability distribution, a BN can actually characterise the joint probability distribution of the variables in $\mathbf{X}$. Moreover, this feature allows to factorise the complex joint distribution into a set of simpler components, as described by the so called Markov property:

$$\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \,|\, \text{pa}(X_i))$$

In order to represent graphically and estimate a decision problem, a BN can be augmented with decision and utility nodes, associated to the realisation of the random variables. Such a model is named Influence Diagram (ID) and its first application dates back to the work of Howard & Matheson (1981). Among the others, Szolovits (1995) and Owens et al. (1997) have discussed the advantages of structuring a medical decision-making by means of an ID. In synthesis, these advantages could be summarised by the following:

- An ID is formally equivalent to a Decision Tree. This latter is a representation of a decision problem where the dimension grows exponentially with the number of decisions. The results associated to any branch are considered, in terms of the probability of occurrence and the total expected utility is then calculated. The decision associated with the highest value is then chosen as the 'optimal';

- However, an ID has the advantage of allowing a more compact representation. This is mainly due to the fact that the decision nodes need not be exploded in order to analyse the results associated to each possible choice. In fact, an ID is based on the conditional probability tables (CPTs). Each variable is associated to a CPT, which describes the probability distribution of that node, given all the possible configurations of the parents;

- The Bayesian nature of BNs and IDs permits to take into account any prior information that is available; this can be expressed in terms of the CPTs for the variables in the model and assumes a noticeable value for a decision maker who is responsible for the final choice. On the one hand, this feature is a source of criticism, in that the prior distributions can strongly affect the results. On the other hand,

by using this tool, the decision maker can combine his (her) expert prior opinion / information with empirical evidence in a very direct way; moreover, the model would easily manage any modification of the structural prior conditions (i.e. the attack rate could dramatically vary from one year to another, leading to very different assumptions);

- The temporal ordering of a set of decisions is directly taken into account in an ID. Moreover, it is assumed that choice nodes (random variables) that precede a decision node represent in fact observed quantities, by the time that the decision is taken. On the other way around, decision nodes that precede choice nodes indicate that the decision directly affect the realisation of those variables. This characteristic properly mimics the actual behaviour of decision making, which is based on whatever relevant happened before, and is likely to affect what comes after. Hence, the formulation of the model becomes more straightforward to the decision maker;

- The evaluation of different scenarios is also quite direct: it is possible to instantiate (set) either a chance or a decision node to a given value and measure how the probability distribution of the variables that are related to the instantiated node are consequently modified. Moreover, the expected utilities associated to the decision node(s) can be calculated under different scenarios, allowing for a direct form of probabilistic sensitivity analysis (Spiegelhalter et al. 2004);

- A BN is a model for the entire joint probability distribution of a set of variables; consequently, there is no need for the definition of 'dependent' (response) and 'independent' (covariates) variables. Each node is evaluated in terms of the dependence relationships with the others, overcoming the limitation of some regression models, where a given form of causality is given as an assumption, without the possibility of learning over it. Moreover, the presence of missing values is easily handled by the network structure;

- Finally, a procedure of update of the prior CPT can be performed, starting from raw data for the variables of the BN. Moreover, the whole structure of the network (i.e. the asserted set of relationships among the variables, in terms of cause – effect) can be evaluated from empirical evidence. This feature is not easy to implement in general, and it is particularly cumbersome in presence of decision and utility nodes, because of computational limitations (Jordan 2001). Nevertheless, using the duality between BNs and IDs, one can express a

model in terms of a BN, allow for empirical data-driven update of both the structure and the parameters of interest, and, finally, work back in a decision framework, in order to calculate expected utilities and discern among alternatives decisions.

4.3 Material and Methods

The model used in this study is based on several different components. According to the work of Shuffman & West (2002), the real decision problem has been reconstructed and represented in terms of an Influence Diagram, using a set of random nodes (variables of interest, each associated to a local probability distribution), decision nodes (the decision of whether to vaccine the reference population or not), and utility nodes, expressed in terms of costs associated to each possible choice. Among the random nodes, we can distinguish three different kinds of variables. First, we have the exogenous variables (i.e. the coverage rate, which is dependent on the reference population, and the attack rate, which the decision maker cannot directly control, as it is flu-season dependent). Second, we can consider the vaccine-dependent variables (i.e. variables taking into account the reduction in the utilisation of health resources generated by the vaccination). Finally, there are the risk measures (i.e. variables used to estimate the probability of death, hospitalisation, GP visits and of incurring in influenza related drug prescription).

In order to evaluate the vaccination problem, we need to define the set of CPTs and to associate some utility measure to the possible choices.

4.3.1 Conditional Probability Tables (CPTs)

The definition of the local probability distributions for the random nodes is a basic part of the model. In this case, we had not access to a comprehensive database, which could allow us to observe directly all the quantities of interest. For this reason, we followed a two-fold strategy: when possible, we tried to use sample data from an observational study carried out on the elderly population of Pianiga (Venice, Italy) during three flu seasons (2000-2001, 2001-2002 and 2002-2003). The average size of the Pianiga population in the study period was 9,307 residents, and the subjects with at least 65 years of age were 1,641 (17.6%).

Co-operation of four General Practitioners organised in Family Medicine Group (FMG) in the same area was obtained. The FMG has an active population of approximately 5,750 patients (with an average of 5,721 in-

dividuals in the study period); among them, 1,114 subjects are at least 65 years of age (67.9% of Pianiga elderly population). Patient of FMG who were at least 65 yeas old on specific flu season (on October, 1st for each flu season) were included in the study, if they were not institutionalised. Influenza seasons were defined by the dates that the first and last influenza isolates were obtained, according to surveillance data reported to Italian Ministry of Health.

The FMG used a microcomputer-based patient record system. This system includes data on patients (demographic characteristics), vaccination status and type of vaccine used, and a prescriptions-writing function that allows prescriptions to be written using the Anatomical Therapeutic Chemical (ATC) classification[1].

The drug prescriptions directly related to influenza syndrome were analysed, divided by: symptomatic treatment (antipyretic, antitussive, etc); respiratory drugs (ATC: R) and antibiotic (ATC: J01); antivirals (ATC: J05).

To estimate influenza related hospitalisation, we analysed data collected from the specific flu season periods from the hospital records of the Veneto Region. Cases of influenza related hospitalisation were identified from all hospital discharge records, using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM).

When we needed information about variables that are not directly measurable in the Pianiga database, or in order to provide more statistically robust estimations, we performed a meta-analysis of literature data, to obtain the distributions of interest. The estimations of exogenous variables were based on Italian data, while the effectiveness of vaccination (either for the standard vaccine or the MF59) was derived by a set of multi national studies. Table 4.1 summarises the data and the sources used in this study.

For all the variables presented in Table 4.1, we built the CPTs using a probability distribution that gives a high mass to the point estimation calculated from the meta-analysis. This was done by performing a standard learning, based on a non informative prior $Beta(1,1)$ distribution. To preserve the computational facilities associated to a Bayesian Network, the

[1]The Anatomical Therapeutic Chemical classification of pharmaceutical products has been developed and maintained by the European pharmaceutical Marketing Research Association (EphMRA), starting from 1971. A Classification Committee has been constituted to take care for new entries, changes and improvements. The 1st level of the anatomic therapeutic classification indicates the anatomical main group (C – Cardiovascular System). The 2nd level identifies the main therapeutic groups (C1 – Cardiac Therapy). Finally, the 3rd level separate out the pharmacological/therapeutic subgroups (C1B – Anti-Arrhythmics). The 3rd ATC level is a widely accepted standard (typically, Antitrust Authorities around the world apply it) to classify products for purposes of identifying the manufacturing market in pharmaceuticals.

	Data sources	Point estimation	
Exogenous variables			
Coverage rate	Gasparini et al. (2002), Friuli (1999), Crocetti et al. (2001), Gasparini et al. (2001), Montomoli et al. (2002)	0.5253	
Attack rate	Gasparini et al. (2002), Gasparini et al. (2001), Montomoli et al. (2002)	0.1658	
Vaccine-dependent variables		*Standard vaccine*	*MF59*
Reduction in GP visit generated by vaccination	Shuffman & West (2002)	0.4490	0.5056
Reduction of influenza related treatments generated by vaccination	Shuffman & West (2002)	0.5300	0.6264
Reduction of hospitalisation for Influenza & Pneumonia generated by vaccination	Shuffman & West (2002)	0.3679	0.4331
Reduction of hospitalisation for Chronic Heart Failure generated by vaccination	Shuffman & West (2002)	0.2413	0.3195
Reduction of hospitalisation for Respiratory Disease generated by vaccination	Shuffman & West (2002)	0.2913	0.3644
Reduction of mortality generated by vaccination	Shuffman & West (2002)	0.4978	0.5493

Table 4.1: Summary of the data used to build the probability distributions of the random nodes

results were summarised on a grid of 10 values in the interval $[0; 1]$. This strategy also allows a sensitivity analysis, since different scenarios can be considered, for the exogenous and the vaccine-related variables, simply by instantiating them to a given value.

As for the risk measures, the CPTs were directly derived from the probability distributions of the relative parents. In particular, each risk measure is assigned a $Bernoulli$ distribution, i.e. each takes value 1 in case the event described (GP visit, death, hospitalisation, drug prescription) occurs, or value 0 otherwise. The probability distributions are weighed by the occurrence rates that have been observed in the Pianiga database. This was done in order to insure a more robust representation of the reality under study.

Figure 4.1 depicts the network structure of the decision problem. The variables 'which vaccine?' and 'vaccinated?' in Figure 4.1 are sort of instrumental nodes. In other words, they are used to tune the effects of the

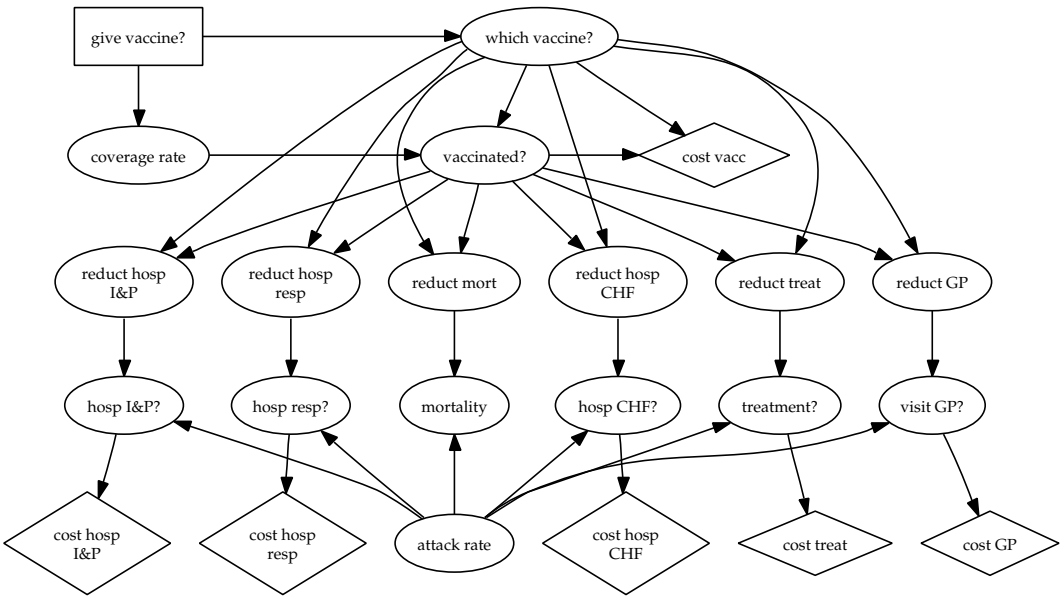


Figure 4.1: The ID representation of the influenza vaccination problem

decision and allow us to model the impact of the decision on the vaccine-dependent variables. Hence, their CPTs are degenerate, in that they only serve as activator/deactivator of a particular scenario.

A similar approach is that of Dawid (2002), developed for the analysis of causal structures using the decision theoretic approach.

4.3.2 Utility measures

The diamond nodes shown in Figure 4.1 represent the utility measures attached to the decision (described by the rectangular node). While associating a proper numerical utility measure to the risk of death (the node labelled as 'mortality' in Figure 4.1) can be cumbersome, it proves to be relatively easy to define a (dis)utility measure for all the other risk measures. In fact, the decision maker should weigh the choice of whether vaccine the elderly population based *a)* on the global reduction of the risks; and *b)* on the total costs associated.

Hence, we calculated the relevant direct costs shown in Table 4.2, weighing the unit cost of each health care resource by the occurrence probability that we observed in the Pianiga database, in order to provide an estimation of costs that is strictly related to the clinical practice reality that we are monitoring.

We used the software `Hugin` to perform the evaluation of the posterior

Costs	Sources	Estimation (€)
GP visit	Official Italian Health Fee Listing	12.42
Symptomatic treatments	Official Italian Drug Listing	18.76
Antibiotic treatments	Official Italian Drug Listing	31.48
Antiviral treatments	Official Italian Drug Listing	54.74
Hospitalisation for Influenza & Pneumonia	Elaboration on DRG fees	4,057.93
Hospitalisation for Chronic Heart Failure	Elaboration on DRG fees	8,305.82
Hospitalisation for Respiratory Disease	Elaboration on DRG fees	5,375.85
Unit cost of standard vaccine + dispensation	Official data on Public Bids	12.87
Unit cost of MF59 vaccine + dispensation	Official data on Public Bids	14.12

Table 4.2: Summary of the cost data used to define (dis)utility measures

probability distribution for all the variables of the model described in Figure 4.1.

4.3.3 Cost effectiveness analysis

The cost effectiveness analysis was then carried out combining the results obtained from the evidence propagation in the ID. This terminology means that the whole system is evaluated after one on more nodes are set to a given value.

We compared three different health programs: *a)* the decision maker chooses not to vaccine the population; *b)* the decision maker chooses to vaccine the population with a standard vaccine; *c)* the decision maker chooses to vaccine the population with the MF59 vaccine. Notice that the model described in Figure 4.1 could also allow a combined strategy, where the decision maker chooses to vaccine the population making available (possibly in different proportions) both the standard and the MF59 vaccine. The costs and the results in terms of effectiveness could then be directly evaluated from the ID resolution.

The cost effectiveness analysis is based on the Incremental Cost Effectiveness Ratio, *ICER* (Gold et al. 1996). For each pairwise comparison, it is defined as the ratio of the difference in total costs, Δ_c, to the difference in any specific effectiveness measure, Δ_e. This value represents the differ-

ence between the probabilities of occurrence of a given risk under the two alternative programs compared (i.e. 'do not vaccine' vs 'give standard vaccine', or 'give standard vaccine' vs 'give MF59 vaccine', or 'do not vaccine' vs 'give MF59 vaccine').

In other words, comparing the health programme a to the programme b (a and b being either of the three described above), will lead to the following formula:

$$\begin{aligned} ICER &= \frac{\Delta_c}{\Delta_e} \\ &= \frac{\overline{c}_a - \overline{c}_b}{\overline{e}_a - \overline{e}_b} \end{aligned}$$

where $\overline{c}_a$ and $\overline{c}_b$ are the average costs and $\overline{e}_a$ and $\overline{e}_b$ are the average effectiveness measures chosen, respectively for programme a and b.

The main effectiveness measure is mortality, although one can evaluate the cost effectiveness of a program with respect to the consumption of any other health care resource considered in the model.

4.4 Results

The algorithm of evidence propagation (Jensen & Dittmer 1994) provides a solution to the ID, i.e. a way to determine the choice associated to the highest expected utility. In fact, all the probability distributions are exploited starting from the prior structure assigned to the CPTs. The total expected utility is then calculated and associated to each possible choice. A first result from the ID of Figure 4.1 is that, given the premises described above, the expected (dis)utility, i.e. the total annual cost, equals €53.04 per patient, should the decision maker decide not to promote a vaccination campaign, and equals €51.18 per patient, in case the decision maker decides to go for a vaccination campaign. This latter result is an average of the two specific situations where the vaccine can be either the standard or the MF59.

Consequently, the vaccination proves to be a cost saving program.

Nevertheless, it is important to evaluate this result in the light of the benefits that are gained from the population, in terms of risks reduction. Table 4.3 summarises some interest findings.

As one can notice, the alternative based on MF59 vaccine always produces the highest risk reduction, in that the probabilities of occurrence of the events are always lower than in the other programs. The standard vaccination also proves to be effective in reducing the risk of experimenting the events, as compared to the null option (do not vaccine the population).

Posterior probabilities for risk measures	MF59 vaccine	Standard vaccine	Do not vaccine
GP visit	0.0604	0.0629	0.0828
Influenza related treatments	0.1102	0.1188	0.1658
Hospitalisation for Influenza & Pneumonia	0.0150	0.0160	0.0200
Hospitalisation for Chronic Heart Failure	0.0550	0.0580	0.0660
Hospitalisation for Respiratory Diseases	0.0006	0.0006	0.0008
Mortality	0.0023	0.0024	0.0033
Average Annual Cost	€50.44	€51.92	€53.04

Table 4.3: Summary of the posterior probabilities for the risk measures, under alternative programs

The total direct costs per year and per patient are of €50.44 in case the decision maker chooses to vaccine population with the MF59 vaccine, and of €51.92 in case the program chosen is the vaccination with the standard vaccine.

Table 4.4 presents the *ICER* for each risk measure and for the three pairwise comparisons.

$ICER = \frac{\Delta_c}{\Delta_e}$	'MF59' vs 'Standard'	'MF59' vs 'Do not vaccine'	'Standard' vs 'Do not vaccine'
GP visit	–592.00	–116.52	–56.78
Influenza related treatments	–172.09	–46.94	–24.04
Hospitalisation for Influenza & Pneumonia	–1,480.00	–522.00	–282.50
Hospitalisation for Chronic Heart Failure	–493.33	–237.27	–141.25
Hospitalisation for Respiratory Diseases	–	–13,050.00	–5,650.00
Mortality	–16,444.44	–2,718.75	–1,298.85

Table 4.4: Summary of the posterior probabilities for the risk measures, under alternative programs

Given that the values Δ_c are always negative for the pairwise comparisons defined above (cfr. Table 4.3), and that vaccination is effective in reducing the likelihood of the outcome measures, than the *ICER* is always a negative value, and can be interpreted as saving per event averted.

As appears clear, the vaccination proves to be a highly cost effective strategy, as compared to the null option. Both MF59 and standard vaccine produce a better health status for the population and savings for the health provider. This is mainly due to the fact that the high-cost events

(hospitalisations) are reduced (see Table 4.3), leading to relevant savings for the public payer. The extra cost generated by the vaccine acquisition and dispensation is more than balanced from this advantages.

From the clinical point of view, as widely accepted, the vaccination proves to provide great benefits, especially in terms of reducing mortality rates, which in fact are considered as the most important effectiveness measure.

In particular, the best advantages from the decision maker perspective seem to occur when comparing the 'MF59' alternative to the 'standard vaccine' one. In fact, in this case the low difference in the 'entry' cost (vaccine price and cost of dispensation) leads to a higher *ICER*. In terms of mortality, the implementation of the vaccination campaign based on MF59 would save €16,444 per each death averted, as compared to the vaccination campaign based on standard vaccine.

4.4.1 Probabilistic Sensitivity Analysis

As reported above, the ID approach allows a very direct way of performing a sensitivity analysis for the results. It only needs to instantiate a variable to one of its values to evaluate how the probabilities and costs are modified. Moreover, one can perform directly a multivariate sensitivity analysis, as the number of variables that are instantiated at the same time can be greater than one.

As an example, since we work with the decision maker perspective, we reckoned that a relevant scenario would concern variations in the influenza attack rate. Hence, we evaluated what would be if the attack rate increased (pandemic season).

The results are shown in Figure 4.2. As appears plausible, in case the attack rate is significantly low (i.e. lower than 0.05), than vaccination does not represent a good value for money. Nevertheless, increasing the attack rate produces higher and higher advantages in the vaccination scenario, either with the standard or the MF59 vaccine.

Finally, we performed a *Break Even Point Analysis* with respect to unit cost of the reference programme (Figure 4.3), evaluating the total cost for the population of elderly people in Italy – about 10 million people, as reported by the Italian Statistical Institute (ISTAT 2003).

In other words, assuming that a change in MF59 unit cost would not affect changes in the other programmes utilisation, we compared the total costs of the three options.

Given the estimated value of €5.7, as described above, vaccination with MF59 is a cost-saving option. Besides, even if this unit cost would be in-

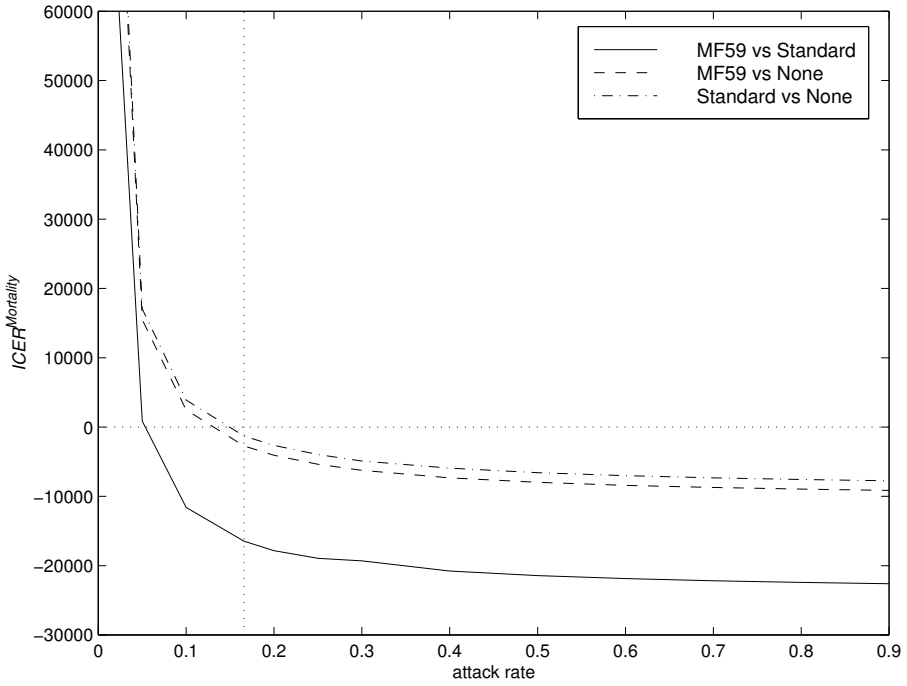


Figure 4.2: Sensitivity analysis upon varying the attack rate value

creased up to €8.5, the public decision maker would face saving, as compared to the standard vaccination option. The unit cost could be increased up to €10.5, in order to equal the total cost of treating influenza, as compared to the null option.

4.5 Discussion

This paper focuses on the analysis of the decision aspects of the influenza vaccination. The scenario that was evaluated was based on the real experience in Italy, but the approach followed is highly transferable to any different reality.

The model combines established findings from the medical literature and sampling data from an observational study into an Influence Diagram structure. The main objective of the model is to provide an estimation of the whole probabilistic relationships among the variables considered, as well as the evaluation of costs associated to the different health programs available to the decision maker.

The main limitations of the study concern the fact that, from the statistical point of view, it would be highly relevant to have access to a complete

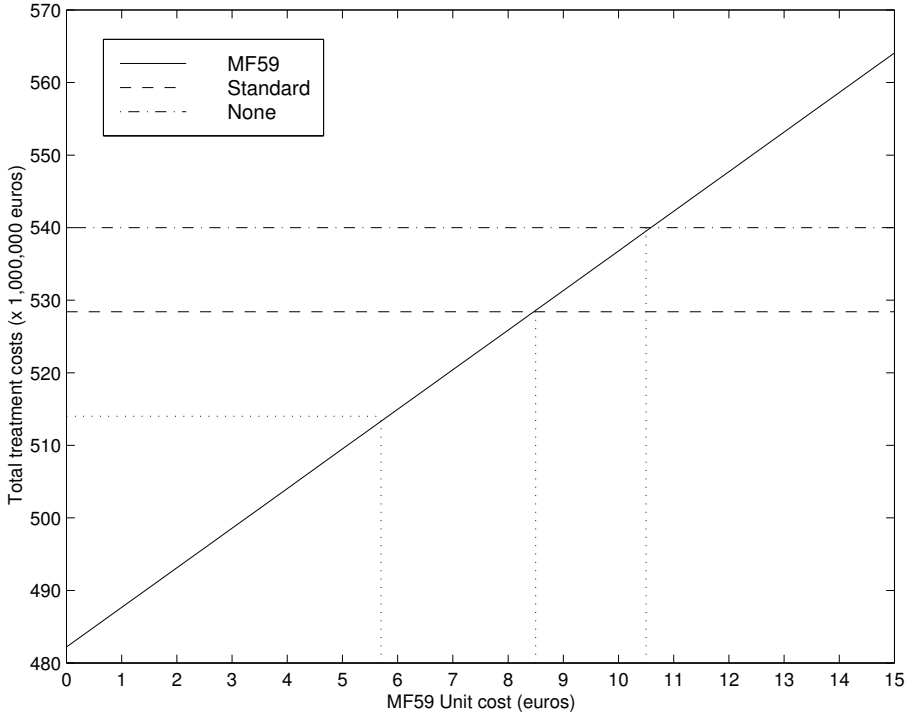


Figure 4.3: Sensitivity analysis upon varying MF59 unit cost

database, in order to estimate both the parameters and the network structure by empirical evidence. This was not possible due to the lack of a database that could cover adequately all the dimension of the analysis. Yet, we combined the prior information provided by clinical studies and expert opinions used to build the network structure to a set of empirical data, in order to obtain a result that mimic as closely as possible the reality. Moreover, from the technical point of view, more efforts are required, in order to program an efficient algorithm that can allow the data-driven probability updating in an ID. Yet, using the BN approach, we obtain a more straightforward representation, which can facilitate the translation of the statistical features into a 'clinical' ('administrative') language.

From our analysis, we provide evidence that the influenza vaccination is a cost effective strategy, as compared to the null option. Moreover, the MF59 vaccine proves to be both more effective and cheaper in the long run, as compared to both the standard vaccination and the null program. Should the attack rate increase (notice that very mild flu seasons were experienced lately in Italy), the vaccination turned to be even more cost effective, with greater benefits coming from the adjuvated vaccine.

The pharmacoeconomics evaluation based on BN and ID has a huge potentiality, in our opinion, for the reasons we described above. We reckon that statistical, economic and clinical research is to be focused on that topic in the next future.

Bibliography

Crocetti, E., Arniani, S., Bordoni, F., Maciocco, G., Zappa, M. & Buiatti, E. (2001), 'Effectiveness of influenza vaccination in the elderly in a community in Italy', *European Journal of Epidemiology* **17**, 163–168.

Dawid, A. (2002), 'Influence Diagrams for causal modelling and inference', *Statistical Review* **70**, 161–189.

Friuli, R. (1999), 'Rapporto sulla vaccinazione contro l'influenza nella popolazione anziana nella stagione 1997-1998'.

Gasparini, R., Lucioni, C., Lai, P., De Luca, S., Durando, P., Sticchi, L., Garbarono, E., Bacilieri, S. & Crovari, P. (2001), 'Influenza surveillance in the Italian region of Liguria in the winter 1999-2000 by general pratictioners and paeditricians: socio-economic implications', *Journal of Preventive Medicine and Hygene* **42**, 83–86.

Gasparini, R., Lucioni, C., Mazzi, S. & Pregliasco, F. (2002), 'Vaccino adiuvato virosomale vs vaccino tradizionale nella strategia antinfluenzale: una valutazione farmacoeconomica', *Pharmacoeconomics Issues in Vaccines* .

Gold, M., Siegel, J., Russel, L. & Weinstein, M. (1996), *Cost effectiveness in health and medicine*, Oxford University Press, New York, NY.

Heckerman, D. (1996), A Tutorial on Learning With Bayesian Networks, Technical Report MSR-TR-95-06, Microsoft Research Advanced Technology Division, Redmond, WA.

Howard, R. & Matheson, J. (1981), *Readings in Decision Analysis*, Strategic Decision Group, Menlo Park, CA.

ISTAT (2003), *Statistiche demografiche*.
*http://demo.istat.it/pop1999/start.html

Jensen, F. (1998), *Bayesian Networks and Decision Graphs*, Springer, New York, NY.

Jensen, F. & Dittmer, S. (1994), From Influence Diagrams to Junction Tree, *in* L. de Mantras & D. Poole, eds, 'Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann Publishers, San Francisco, CA, Edmonton, Canada.

Jordan, M. (2001), *Learning in Graphical Models*, MIT Press, Cambridge, MA.

Montomoli, E., Pozzi, T., Alfonsi, V., Grezzi, B., Pilia, S., Ravasio, R., Lucioni, C. & Gasparini, R. (2002), 'Valutazione benefici/costi della vaccinazione antinfluenzale in una popolazione di soggetti anziani in due stagioni epidemiche a confronto (1999-2000 - 2000/2001) nella Provincia di Siena', *Pharmacoeconomics* .

Nichols, K. (2001), 'Cost Benefit Analysis of a Strategy for Healthy Working Adults against Influenza', *Archives of Internal Medicine* **161**, 749–759.

Nichols, K. & Goodman, M. (2002), 'Cost effectiveness of influenza vaccination for healthy persons between ages 65 and 74 years', *Vaccine* **20**, S21–S24.

Owens, D., Shachter, R. & Nease, R. (1997), 'Representation and Analysis of Medical Decision Problems with Influence Diagrams', *Medical Decision Making* **17**, 241–262.

Potter, C. (2001), 'A history of influenza', *Journal of Applied Microbiology* **91**, 572–579.

Shuffman, P. & West, P. (2002), 'Economic evaluation of strategies for the control and management of influenza in Europe', *Vaccine* **20**, 2562–2578.

Spiegelhalter, D., Abrams, K. & Myles, J. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley and Sons, Chichester, UK.

Szolovits, P. (1995), 'Uncertainty and Decisions in Medical Informatics', *Methods of Information in Medicine* **34**, 111–121.

Vu, T., Farish, S., Jenkins, M. & Kelly, H. (2002), 'A meta-analysis of effectiveness of influenza vaccine in persons aged 65 years and over living in the community', *Vaccine* **20**, 1831–1836.

– *“Causality. There is no escape from it. We are forever slaves to it. Our only hope, our only peace is to understand it, to understand the why. Why is what separates us from them. You from me. Why is the only real social power. Without it you are powerless.”*

The Merovingian in “The Matrix reloaded”