Euro Jnl Phil Sci (2013) 3:275–308
DOI 10.1007/s13194-013-0068-z

ORIGINAL PAPER IN PHILOSOPHY OF SCIENCE

Should causal models always be Markovian? The case of multi-causal forks in medicine

Donald Gillies · Aidan Sudbury

Received: 5 July 2012 / Accepted: 15 April 2013 / Published online: 4 June 2013

Abstract The development of causal modelling since the 1950s has been accompanied by a number of controversies, the most striking of which concerns the Markov condition. Reichenbach's conjunctive forks did satisfy the Markov condition, while Salmon's interactive forks did not. Subsequently some experts in the field have argued that adequate causal models should always satisfy the Markov condition, while others have claimed that non-Markovian causal models are needed in some cases. This paper argues for the second position by considering the multi-causal forks, which are widespread in contemporary medicine (Section 2). A non-Markovian causal model for such forks is introduced and shown to be mathematically tractable (Sections 6, 7, and 8). The paper also gives a general discussion of the controversy about the Markov condition (Section 1), and of the related controversy about probabilistic causality (Sections 3, 4, and 5).

Keywords Probabilistic causality · Conjunctive forks · Interactive forks · Multi-causal forks · Markov condition · Bayesian networks · Causal factors · Heart disease

Although the two authors collaborated on all aspects of the paper, the philosophical parts were mainly the work of Donald Gillies, and the mathematical parts of Aidan Sudbury. In particular the proofs of Theorems 1 and 2 were due to Aidan Sudbury. Both authors would like to acknowledge the very important contribution to the paper of Christian Hennig. In early discussions, it was Christian Hennig who made the key points which led to the mathematical formulation of the problem in Section 6, and he also made the interesting comment on the conditions of Theorem 1, which we have included as Section 8. However, Christian Hennig's philosophy of probability and statistics is rather different from the one presupposed in this paper. A general view of his philosophy is to be found in Hennig (2010), and he is at the moment developing it in more detail.

D. Gillies (✉)
University College London, London, UK
e-mail: donald.gillies@ucl.ac.uk

A. Sudbury
Monash University, Melbourne, Australia
e-mail: aidan.sudbury@monash.edu

1 Introduction. Debates about the Markov condition

The theory of causal modelling has developed in a striking fashion since the 1950s, but this development, like most research developments, has been accompanied by some controversies. Perhaps the most significant of these controversies has been that which concerns the role of the Markov condition in causal modelling.[1] Many researchers in the field hold that a causal model cannot be satisfactory unless it is Markovian, that is to say that every node of the model should satisfy the Markov condition. Others, however, hold that the Markov condition is not always satisfied in reality, so that the use of Markovian models can be misleading and that non-Markovian models should be considered. The present paper supports the second of these two positions. It proposes a non-Markovian model for what are called *multi-causal forks*. The use of multi-causal forks is widespread in modern medicine, so that this model should be a useful one, and it is shown that, although it is non-Markovian, it is perfectly tractable mathematically.

Before we introduce multi-causal forks, and the non-Markovian causal model, which we propose for handling them, it will be useful to give a brief account of the debates, which have occurred so far, concerning the Markov condition and its role in causal modelling. This we will do in the present section of the paper.

The first causal model in the modern sense can perhaps be attributed to Reichenbach. In his 1956, p. 159, he introduced what he called a *conjunctive fork*. This is illustrated in Fig. 1.

Here A and B are correlated, and this is explained by the fact that they have a common cause C. C *screens off* A from B, that is to say that A and B are independent given C. It is clear that Reichenbach's screening off condition implies the Markov condition, and that his conjunctive forks are simple Bayesian networks. Indeed they were the first Bayesian networks to be introduced. Reichenbach went further and formulated what he called (1956, p. 157f.), *the principle of the common cause*. This states that, if A and B are correlated, then either A causes B, or B causes A, or A and B have a common cause C which screens off one variable from the other.

Reichenbach's ideas on causality were developed by Salmon, but Salmon found it necessary to introduce, in addition to Reichenbach's conjunctive fork, a second kind of causal fork, which he called an *interactive fork*. As he says (1978, p. 134):

> "It thus appears that there are two kinds of causal forks: (1) Reichenbach's *conjunctive fork*, in which the common cause screens off the one effect from the other, …, and (2) *interactive forks*, exemplified by the Compton scattering of a photon and an electron."

Interactive forks can be illustrated by the same diagram as conjunctive forks (Fig. 1). The difference is that in a conjunctive fork the common cause C screens off A from B, but this is not the case in interactive forks. Salmon shows the need for interactive forks by his interesting example of Compton scattering.

In a Compton scattering experiment, an energetic photon collides with an electron which can be regarded as more or less stationary. The collision is represented by the node C, where the variable C has the energy E as its value. As the result of the

[1] For definitions of the Markov condition, and the other technical terms used in this paper, see appendix.

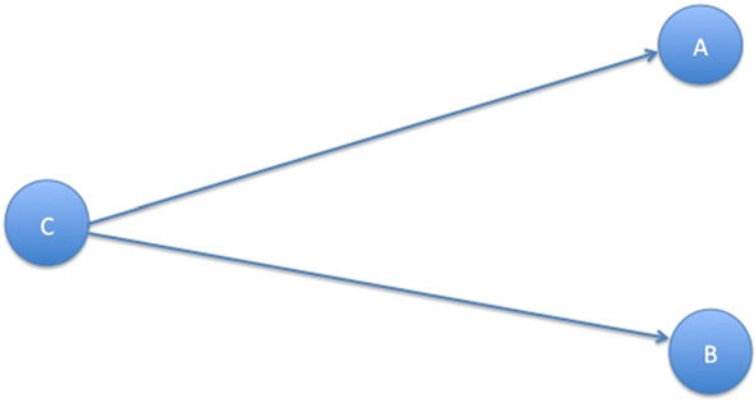

Fig. 1 Conjunctive or Interactive Fork

collision we get an electron with energy E_1 represented by node A, and a photon with energy E_2 represented by node B. Now because of the conservation of energy, we have $E=E_1+E_2$, and so A and B are highly correlated given C. C is a common cause of A and B, but it does not screen off A from B. So this is a causal fork, which is not a conjunctive fork. It is an interactive fork.

Conjunctive forks are Markovian causal models, while interactive forms are non-Markovian causal models. So Salmon clearly supports the view that at least some causal models should be non-Markovian.

Suppes in his 1986 also supports the idea that there are non-Markovian causal models, but in a different way from Salmon. Salmon's non-Markovian causal model is drawn from physics. However, Suppes holds that the Markov assumption has on the whole been successful in physics. He then develops a general argument to show that Markovian conditions are unlikely to be as applicable in the social sciences as in physics. He illustrates this thesis with examples of non-Markovian models in psychology and economics. Suppes concludes his paper as follows (1986, p. 140):

> "Philosophical views of causality—at least if it is intended for them to be relevant to theoretical and empirical work in the social sciences—must not be restricted to the dominant Markovian conceptions of causality that have played such a central role in physics."

Causal modelling received a tremendous impulse from the development of the theory of Bayesian networks in the 1980s. The principal figure was Pearl who introduced and developed the concept of Bayesian network in a series of papers: Pearl (1982; 1985a, b; 1986), Kim and Pearl (1983), and a book: Pearl (1988). An important extension of the theory was carried out by Lauritzen and Spiegelhalter (1988), while Neapolitan's 1990 book gave a clear account of these new ideas and helped to promote the use of Bayesian networks in the AI community. There is one point, however, on which Neapolitan's approach differs from that of the others just mentioned. Pearl, Kim, Lauritzen and Spiegelhalter all interpret the probabilities in Bayesian networks subjectively as degrees of belief. Neapolitan, however, suggests that they might be interpreted objectively. In this paper,

we will always interpret the probabilities in Bayesian networks and causal probability models objectively.[2]

The theory of Bayesian networks was well received by AI researchers, and many successful AI systems were constructed using Bayesian networks in the 1990s. Here we will mention, as an example, a system for partially automating colon endoscopy (Sucar et al. 1993). Such a system has to be able to recognise the lumen of the colon, and give an operator correct advice on how to proceed. The first attempt to construct such a system was purely logic based, and did not use probability. It gave the correct interpretation and advice regarding the lumen in only 39 % of cases. However, when an appropriate Bayesian network was developed for the problem, the percentage of correct results rose to over 90 %. This is a good illustration of the power of the new theory.

Despite such successes, the theory of Bayesian networks was not well received in all quarters. Bayesian networks, by definition, satisfy the Markov condition. Critics, following in the line of Salmon's criticism of Reichenbach, objected that there might be many cases in which the Markov condition is not satisfied. For example, Cartwright in her 1995, discusses a version of the Markov condition, which she describes as (p. 340): "the screening-off condition familiar to philosophers from the work of Reichenbach, Suppes and Salmon". Of this version of the Markov condition she says (p. 341):

> "… it is not universally true for genuine probabilistic causation. Far from it. It is a very special case that holds in unusual circumstances."

She then goes on in her paper to give some examples in which the Markov condition fails.

In 2000, Pearl published a book on causality. In this work, his focus had shifted somewhat away from AI towards discussion of causal models in econometrics and epidemiology. The main emphasis is on what he called: 'functional causal models' and later, e.g. in his 2011, 'structural causal models'. Structural causal models are so-called because the involve features of the structural equation models (SEM) used in econometrics by figures such as Haavelmo (1943). Despite these changes, however, Pearl's structural causal models still satisfy the Markov condition, and so he needed to reply to those who criticized the use of this condition. He does so as follows (2000, p. 62):

> "… criticisms of the Markov assumption, most notably those of Cartwright …, have two characteristics in common:
>
> 1. they present macroscopic non-Markovian counterexamples that are … of the type considered by Salmon …, that is, interactive forks; and
> 2. they propose no alternative, non-Markovian models from which one could predict the effects of actions and action combinations."

[2] There are actually two senses of 'objective' as applied to probability. 'Objective' can mean objective in the scientific sense, or objective in the logical, or epistemic, sense. In this paper, we will confine ourselves to probabilities, which are objective in the scientific sense. For probabilities, which are objective in the logical, or epistemic, sense, see Williamson 2005, Ch. 5, pp. 65–106, and Williamson 2010.

These comments can be considered as posing a challenge to anyone who wants to defend non-Markovian causal models. The non-Markovians should show that there are non-Markovian counterexamples not reducible to Salmon's interactive forks, and they should produce non-Markovian models which are mathematically tractable and from which action-guiding results can be deduced. Later in the paper we will demonstrate both these points. We will produce (see Section 2) a non-Markovian counterexample, which is quite different from Salmon's interactive forks, and we will produce (in Sections 6, 7 and 8) a non-Markovian causal model of a well-known situation in preventative medicine, and show that action-guiding results can be deduced from it.

Pearl argues that apparent non-Markovian counterexamples, such as Salmon's interactive forks, can be dealt with by introducing into the network latent, or unobservable, variables by means of which the Markov condition can be restored. He admits that exceptions to this claim might occur in quantum mechanics. Perhaps this is a reference to Salmon's example of Compton scattering. As he says (Pearl 2000, p. 62):

> "Only quantum-mechanical phenomena exhibit associations that cannot be attributed to latent variables, and it would be considered a scientific miracle if anyone were to discover such peculiar associations in the macroscopic world."

Interestingly, Cartwright gives an example in her 1989, which is formally identical to Salmon's Compton scattering example, but deals with events in the everyday macroscopic world. She writes (1989, p. 114):

> "For example, an individual has \$10 to spend on groceries, to be divided between meat and vegetables. The amount that he spends on meat may be a purely probabilistic consequence of his state on entering the supermarket; so too may be the amount spent on vegetables. But the two effects are not produced independently. The cause operates to produce an expenditure of n dollars on meat if and only if it operates to produce an expenditure of $10-n$ dollars on vegetables."

A simpler everyday example along the same lines would be that of a mother who repeatedly has to divide pieces of cake between her two children—the child who has behaved better recently getting the larger slice.

Pearl would no doubt say that Cartwright's macroscopic interactive fork could be turned into a Markovian causal model by adding some latent variables. However he does not discuss this example, or say what latent variables would be required to deal with it. As we shall see in Section 5, adding latent variables to a causal model is by no means an unproblematic process.

Pearl's general conclusion is that (2000, p. 63):

> "… counterexamples to the Markov condition are relatively rare and can be explained away through latent variables."

This is in rather striking contrast to Cartwright's view that the Markov condition is (1995, p. 341): "a very special case that holds in unusual circumstances." Before we discuss this interesting and important controversy further, however, it will necessary to distinguish between two different questions concerning the Markov condition, which are often treated together.

In the past few decades, a good deal of research has been carried out on trying to produce machine learning programs which are capable of obtaining causal relations from statistical data. The success of this research remains somewhat controversial. Its advocates claim that many striking successes have been achieved, while its detractors are sceptical as to whether any result of significance has been produced. Now these machine learning programs often make use of Bayesian networks, and hence of the Markov condition. Some of Cartwright's criticisms of the Markov assumption are directed against its use in such machine learning programs. For example, in her 2001 paper: 'What is Wrong with Bayes Nets?', she mainly criticizes the use of Bayes Nets in the work on machine learning by Spirtes et al. (1993), and by other authors working on related research programmes. She characterises the aim of her paper as that of discussing (Cartwright 2001, p. 242): "… a variety of algorithms for inferring causal relations from independencies. These I will loosely call 'Bayes-nets methods'."

In the present paper our aim is to discuss the use of causal Bayesian networks, and other causal networks in modelling situations, which arise in the natural sciences and medicine. We do not want to discuss the problematic question of whether Bayesian networks can be used to obtain causal relations from statistical data. In order to separate sharply the question of interest here from the machine learning question, we propose the following strategy.

In the 450 years from 1500 to 1949, during which no computers or machine learning existed, scientists nevertheless discovered a great number of hypotheses or models to explain data. They did so by a process which could be called: 'human learning'. Popper (1963) analyses human learning as a sequence of conjectures and refutations. A first conjecture C_1 is put forward to explain some data. It is then tested out rigorously against this data, and, if it is refuted, a new conjecture C_2 is put forward. This in turn is tested out against the data, and so on, until, hopefully, a conjecture C_n is reached, which is not refuted, but on the contrary well corroborated by the data. This was the method by which Kepler made his discovery that the planet Mars moves in an ellipse with the Sun at one focus. Kepler had access to the very accurate observations about positions of Mars, which had been made by Tycho Brahe. His first conjecture about the orbit of Mars round the Sun was that it was circular. However, this was refuted by Tycho Brahe's data. Kepler then tried, as his second conjecture, the idea that the orbit was generated by a circle and an epicycle. This again could not be made to fit Tycho Brahe's data. Kepler introduced the hypothesis of an elliptical orbit as his fourth conjecture, but this time his conjecture was very well corroborated by the data. The hypothesis that Mars (and the other planets) move in ellipses with the Sun at one focus is not of course a causal model. However, causal models can be learned by humans, through exactly the same process of conjectures and refutations without using computers and machine learning.

As we have already remarked, the success of attempts to obtain causal relations from statistical data remains somewhat controversial. However, one thing would, we think, be agreed by everyone, namely that the machine learning of causal relations has not completed superseded the human learning of such relations. The majority of causal claims are formulated and tested out by humans using the familiar method of conjectures and refutations. Perhaps some day machine learning will completely supersede human learning, but, for the moment, human scientists are still needed for much of the time.

Given this situation, we propose, in this paper, to limit ourselves to considering causal models, which are the products of human learning rather than machine learning. This is in order to put aside the question of the efficacy of what Cartwright calls: 'Bayes-nets methods' in machine learning, and to focus instead on the question, which we want to tackle in this paper. This is the question of the value of different types of causal network for modelling situations in the natural sciences and medicine. Should we confine ourselves to the use of Markovian causal models, or is there a role for non-Markovian causal models as well?

Having formulated our research area with some precision, let us now return to the question of whether "counterexamples to the Markov condition are relatively rare" (Pearl 2000, p. 63), or whether the Markov condition is "a very special case that holds in unusual circumstances" (Cartwright 1995, p. 341). Williamson discusses counter-examples to the Markov condition[3] in his 2005, pp. 51–57. He begins by showing that the Markov condition implies Reichenbach's principle of the common cause. Now many exceptions to Reichenbach's principle of the common cause have been discovered, and these constitute counter-examples to the Markov condition. In addition further counter-examples have been discovered in the context of trying to develop Bayesian networks in various contexts. One of the interesting features of Williamson's treatment is that he tries to classify the reasons why the conditional independence of variables in a network can fail. He says (2005, p.52):

> "… probabilistic dependencies arise … because the variables are related through meaning, through logical connections, through mathematical connections, because they are related by (non-causal) physical laws, or because they are constrained by local laws or boundary conditions."

All this would seem to favour Cartwright against Pearl, but perhaps the question should be formulated in a somewhat different fashion. To do so, we will consider in a little more detail the example of Sucar et al. (1993), which was mentioned earlier. The system was designed to recognise the lumen of the colon, and give advice to the operator about how to proceed. Part of the network, which was initially tried, is shown in Fig. 2.

Here L stands for the lumen, which causes a large dark region (or LDR) to appear on the screen. This in turn produces values for the variables S, which measures the size of the region in pixels, M, which measures its mean intensity, and V, which measures the variance of that intensity. There was an abundant supply of videotapes of colon endoscopies, in each frame of which, the lumen could be indicated by an expert. These gave a mass of frequency data from which probabilities in an objective sense could be estimated. In this case, the Markov condition implies that S, M and V should be probabilistically independent given LDR. Since the probabilities involved could be estimated from data, it was possible to test these consequences of the Markov condition using frequency data. These statistical tests showed that, given

[3] On pp. 51–57 of his 2005, Williamson is actually discussing exceptions to what he calls the *Causal Markov Condition*, which is defined on p. 50, and is distinguished from the *Markov Condition*, which is defined on p. 15. In this paper, we are using the term 'Markov condition' to cover both what Williamson calls the 'Markov condition' and what he calls the 'Causal Markov Condition'. So on pp. 51–57 of his 2005, Williamson is indeed discussing what are exceptions to the Markov condition *in our sense of the term*. The differences here are purely terminological.

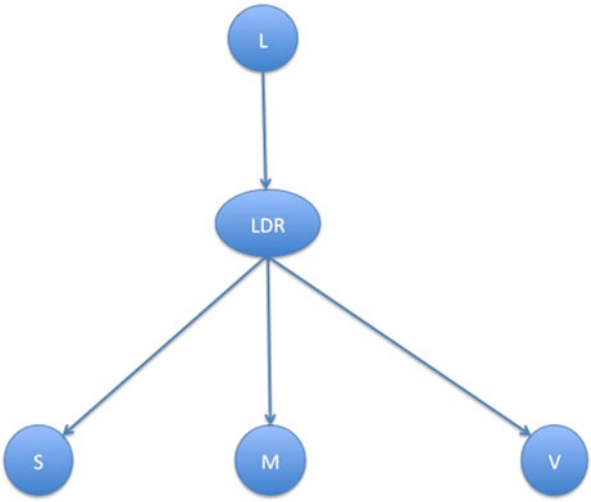

Fig. 2 Part of a Network for Colonoscopy

LDR, S&M and S&V could reasonably be regarded as independent, but that M & V were strongly correlated and so could not be regarded as independent. This failure of the Markov condition is one, which Williamson classifies as due to a mathematical connection, since well-known mathematical relations exist between the variables M and V (see Williamson 2005, p. 54).

In this case, however, it was possible to restore the Markov condition in a simple fashion. It was argued that, since M and V were correlated, it might be possible to get just as good results by eliminating one of them. In fact, it turned out that eliminating V improved the performance of the system, as related to L, on all measures (for details, see Sucar, Gillies and Gillies, p. 206). Eliminating V also reduced the network size making object recognition faster. So this change produced a system, which was not only more reliable, but also more efficient.

Generalising from this example, we can produce the following improved formulation of the problem. The question as initially posed was whether violations of the Markov condition were relatively rare or very common. However, any proposed Markovian causal model should be regarded as a conjecture, which ought to be tested using data. Sometimes statistical tests will confirm that the Markov condition holds, [4]but, in other cases, they will show that it is violated. The important question, however, is not that of how frequently such statistical violations of the Markov conditions occur, but whether, if the Markov condition is refuted, it can be restored by a suitable modification of the network. In fact, a number of techniques are available for modifying a network in order to restore the Markov condition. (1)The first, and simplest, is that used in the colonoscopy example, and consists in eliminating one of the variables between which a conditional probabilistic dependency exists. (2) A second technique is to add an arrow joining the two dependent variables. We

[4] In speaking of statistical tests confirming that the Markov condition holds, we do not mean to imply that they establish the Markov condition with certainty, but only that they provide some evidence in its favour. A scientific conjecture can never be established with certainty by a number of tests with favourable outcomes, since future tests may always show that the conjecture is false in some respects.

will give an example of this modification technique later in the paper. (3) A third technique is to add new variables to the network. These can either (3i) be observable variables, or (3ii) be unobservable or latent variables. In fact Pearl refers to this last approach when he says (2000, p. 63) that "… counterexamples to the Markov condition … can be explained away through latent variables." Pearl mentions only one of the techniques, i.e. (3ii), for modifying a network to restore the Markov condition, but, in what follows, we will allow all three, and indeed others, if they can be discovered.

In terms of this formulation of the problem, we can now state the aim of the paper as follows. Let us suppose that we are attempting to devise a causal model to explain some phenomenon in the natural sciences or medicine. It may be sensible to begin by assuming the Markov condition, since models satisfying the Markov condition are easier to handle mathematically. However, and this is the key point, *we should not assume that what is mathematically convenient necessarily holds in the real world*. On the contrary, it is a basic principle of scientific method that we should test out the assumptions of our model against the data. In some cases, these statistical tests will show that the Markov condition does hold, and there is no problem. Even if statistical tests show that the Markov condition does not hold, it might be possible to modify the network in order to restore the Markov condition. There are several ways in which this can be done, and, using some of them, we might well produce a modified network, which is well-confirmed by the data, and satisfactory for our purposes. However, yet another situation is possible. It could turn out that modifying the network to restore the Markov condition produces a more complicated network which is not well-confirmed empirically, nor satisfactory for our purposes; while the non-Markovian network is well-confirmed empirically and much simpler. If, in such a case, we can make the non-Markovian network mathematically tractable in the sense of showing how to deduce from it the results we need to guide our actions, then such a non-Markovian network is surely to be preferred to a Markovian one. It would seem, in such a case, merely dogmatic to insist that the non-Markovian model is unsatisfactory simply because it does not satisfy the Markov condition. The aim of the rest of this paper is to produce such a non-Markovian model, designed to explain some important situations which arise in modern medicine.

2 Indeterministic causality and multi-causal forks

Before introducing our example, it is necessary to make a couple of distinctions regarding causality. The first of these is between *deterministic* and *indeterministic* causality. 'A causes B' involves a deterministic notion of causality if, *ceteris paribus*, the instantiation of A is always followed by B. A simple example of deterministic causality is: 'The sprinkler causes the grass to get wet.' Here 'The sprinkler' is short for 'The sprinkler being properly connected to a working water supply and turned on'. When that happens, *ceteris paribus*, the grass always gets wet.

Deterministic causality is the traditional concept of causality, which is analysed by 18th and 19th century philosophers such as Hume and Kant. In the 20th century, however, a new concept of causality appeared largely in connection with medical epidemiology. An example of this new, or indeterministic, type of causality is:

'Smoking causes lung cancer'. This is now a generally accepted causal law, and yet smoking is not always followed by lung cancer. To show this, we will quote statistics to be found in Doll and Peto (1976). These are concerned with a sample of 34,440 male doctors in the UK. The 1976 paper deals with the mortality rates of the doctors over the 20 years from 1 November 1951 to 31 October 1971. During that time, 10,072 of those who had originally agreed to participate in the survey had died, and 441 of these had died of lung cancer. As about 83 % of the doctors sampled were smokers, this means that only about 5 % of these smokers died of lung cancer. So, although smoking causes lung cancer, smoking is not always followed by lung cancer.

But although smoking is not invariably followed by lung cancer, smoking definitely increases the probability of getting lung cancer. Doll and Peto calculated the annual death rate from lung cancer per 100,000 men standardised for age. The results in various categories were as follows (1976, p. 1527):

Non-smokers	10
Smokers	104
1–14 g tobacco per day	52
14–24 g tobacco per day	106
25 g tobacco per day or more	224
(A cigarette is roughly equivalent to 1 g of tobacco)	

These results do indeed show a striking correlation between smoking and lung cancer. Smokers are on average more than 10 times more likely to die of lung cancer than non-smokers, and this figure rises to more than 22 times for heavy smokers who consume 25 g or more of tobacco per day. These results are highly significant statistically.

Since its introduction in the 1950s in the investigation of smoking and lung cancer, the notion of indeterministic causality has become ubiquitous in medicine. Consider, for example, the claims that fast food causes heart disease, that infection with the papilloma virus causes cervical cancer, or that some particular genes cause Alzheimer's. In all these important recent claims in medicine, the notion of causality is that of indeterministic causality. There are, however, a lot of problems connected with indeterministic causality, which are far from having been resolved. Galavotti, who refers to indeterministic causality as probabilistic causality, gives a good account of these problems in her 2010 where she writes (p. 140):

> "The first problem that arises as soon as causality is taken as a probable rather than constant conjunction is that of identifying causal as opposed to spurious relations, without getting muddled with problems of the Simpson's paradox kind. … . Moreover, the virtuous circle linking causality, explanation and prediction within classical determinism (of the Laplacean kind) breaks down in the case of probabilistic causality."

Galavotti is quite correct to draw attention to "problems of the Simpson's paradox kind", and we will encounter some of these problems in Section 3.

In addition to the deterministic/indeterministic distinction, there is one other distinction concerned with causality, which we shall find of use. This is the

distinction between *generic* and *single-case* causality. A causal claim such as A causes B is said to be generic, if it can be instantiated on different occasions. Our two preceding examples: 'The sprinkler causes the grass to get wet' and 'Smoking causes lung cancer' are both examples of generic causality. By contrast, a causal claim is single-case if it applies to only one instance. An example is: 'A heart attack caused Mr Smith's death'.[5]

In this paper we will confine ourselves exclusively to generic causality, and will use the term causality only in this sense from now on. This is partly because we regard generic causality as more fundamental than single-case, and hence think it is better to begin by analysing generic causality. It is also because we agree with Campaner and Galavotti, when they write (2007, p. 181): "… the relationship between type and token causality is highly problematic." They go on to point out that Suppes adopts the strategy of dealing with type causality first, and leaving a theory of token causality to be developed later. This is the strategy, which we will adopt here.

19th and early 20th century scientific medicine, as it was developed by Pasteur, Koch, and others, used a deterministic notion of causality. An attempt was made to show that each disease had a single cause, which was both necessary and sufficient for the occurrence of that disease.[6] So, for example, tuberculosis was caused by a sufficiently large number of tubercle bacilli in the patient. As we have seen, however, from the 1950s on, medicine has had to introduce an indeterministic notion of causality. This goes hand in hand with explaining diseases as caused by the conjunction of several causes acting together. This is multi-causality as opposed to the earlier mono-causality. The various different causes, which act together to produce the disease, could be called *causal factors*. The term 'risk factor' is also used, but will not be adopted here since it seems more suitable for a purely probabilistic concept rather than a causal concept. The use of multi-causality, or several causal factors, leads to the introduction of *multi-causal forks*, as shown in Fig. 3.

Here Z, a disease, has a finite number n of causal factors X_1, X_2, …, X_n. Perhaps the most important case of a multi-causal fork in contemporary medicine is the case of heart disease (see Levy and Brink 2005). In the last 60 years, quite a number of causal factors for heart disease have been discovered. These include: smoking, eating fast food, high blood pressure, diabetes, and obesity. In the last few years, investigations have begun into possible genetic causal factors. As heart disease is still the number one killer in the developed world, this case is obviously an important one. A fully developed causal model for heart disease would have to include all the factors just mentioned and perhaps others as well. However, it seems sensible to begin an investigation of multi-causal forks with a rather simpler situation. Accordingly, we will initially confine ourselves to multi-causal forks with two prongs, as shown in Fig. 4.

[5] The terminology 'type/token' is often used for the distinction, which we have described as 'generic/single-case'. Russo and Williamson argue that generic/single-case is a better terminology than type/token, since the terms 'type' and 'token' normally refer to objects, whereas causal claims characteristically relate events and variables (see Russo (2009) and Russo and Williamson (2011)). We are sympathetic to their point of view and will use generic/single-case, except when quoting from authors who use type/token. Hitchcock (2010, 1.3) uses 'general/singular' for the same distinction.

[6] For details, see Codell Carter (2003).

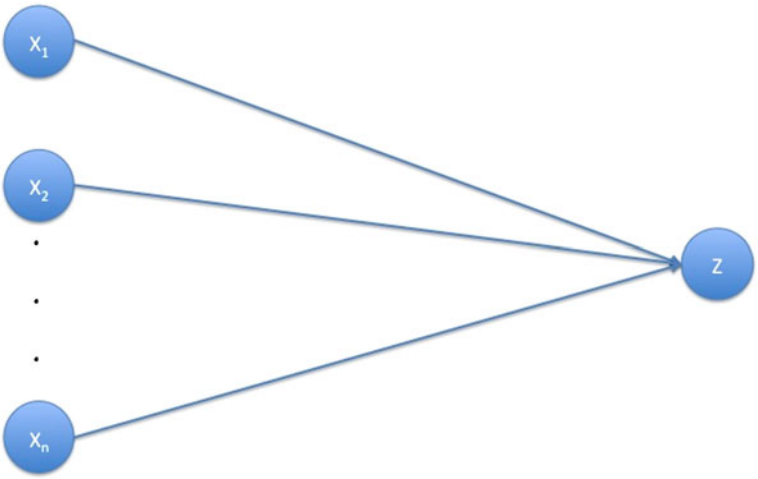

Fig. 3 n-Pronged Multi-Causal Fork

In our example of heart disease, we will take X=smoking, Y=eating fast food, and Z=heart disease. This is shown in Fig. 5.

The multi-causal forks of Figs. 4 and 5 are obviously different from the conjunctive or interactive forks illustrated by Fig. 1, for the arrows in Fig. 4 run in the opposite direction from those in Fig. 1. Applied to the multi-causal fork of Fig. 4, the Markov condition states that X should be independent of Y. However, this is not satisfied in our heart disease example, since smoking is not independent of eating fast food. The two are correlated. This justifies the claim made earlier that that there are non-Markovian counterexamples not reducible to Salmon's interactive forks. As a matter of fact, multi-causal forks are mentioned in Reichenbach's classic 1956. He calls them 'forks open towards the past, constituted by a common effect' (see p. 159). However, he does not discuss them in detail.

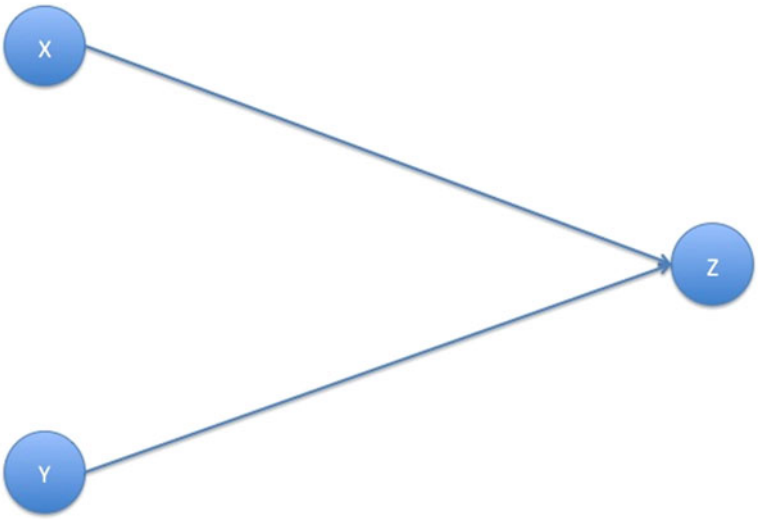

Fig. 4 2-Pronged Multi-Causal Fork

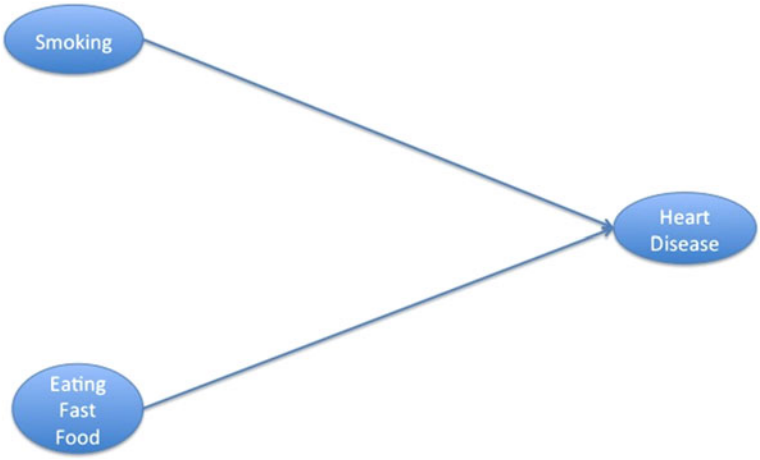

Fig. 5 Multi-Causal Fork for Heart Disease

Our proposed non-Markovian causal model is a multi-causal fork designed for medical examples such as the causal factors in heart disease. To make good our claim that such a model is worthy of being taken seriously, we have to show that it is (i) mathematically tractable and (ii) that we can deduce from it action-guiding results. Before proceeding with this task, it is worth considering briefly what kind of action-guiding results we need to deduce from a multi-causal fork such as the heart disease example of Fig. 5.

Multi-causal forks are routinely used by cardiologists to recommend to patients, strategies for preventing heart disease. A cardiologist would typically advise a patient to give up smoking, and to replace a fast food diet by healthy eating (perhaps some combination of the traditional Mediterranean and Japanese diets). Moreover, if a patient is not strong-willed enough to give up both smoking and fast food, then the cardiologist would advise him or her to give up at least one. All this advice seems intuitively correct, and we have to show that the cardiologist's strategies can all be justified by deductions from our heart disease model of Fig. 5.

To carry out these deductions, we need to provide some link between the causal influences shown by arrows in Fig. 5, and the related probabilities and statistics. To do so, we will make use of some ideas taken from an intellectual development known as 'probabilistic causality'. A consideration of probabilistic causality, however, lands us once again in controversial territory. There is a major problem connected with probabilistic causality, and it is by no means clear that it has been solved. Cartwright was, for a time, one of the advocates of probabilistic causality, but, once again, her work, and that of others of the same school, has been criticized by Pearl both in his 2000 and in a recent (2011) paper. We will discuss probabilistic causality in the next Section (3), and Pearl's alternative approach in Section 4. Then in Section 5, we will make a comparison of the two approaches.

3 Probabilistic causality and its main problem

The appearance of indeterministic causality in science in the 1950s was soon followed by the emergence of probabilistic causality among philosophers of science.

Reichenbach in his 1956 can, once again, be seen as a pioneer. However, the theory of probabilistic causality really got underway with the works of Good (1961, 1962), and Suppes (1970). Subsequently important developments of the approach were made by, among others, Cartwright (1979), Salmon (1980), Eells (1991), and, more recently, Twardy and Korb (2004), Galavotti (2010). Russo (2009) gives an excellent overview of the debates in this area. Pearl, as we shall see in the next Section (4), has been the principal critic of probabilistic causality.

The original hope of the programme was to define causality in terms of probability, but this hope was definitely abandoned by researchers in probabilistic causality from Cartwright (1979) onwards (see Twardy and Korb 2004, p. 241). These later writers had the more modest aim of establishing a link between indeterministic causality and probability, and to do so they made use of a principle which had been formulated in the early days of the programme. This could the called the Causality Probability Connection Principle (CPCP or CP^2). This can be stated as follows:

$$\text{If A causes B, then } P\left(B\middle|A\right) > P\left(B\middle|\neg A\right) \quad \text{(CPCP)}$$

CP^2 seems intuitively reasonable, and certainly holds in the paradigm case of 'smoking causes lung cancer' as we saw in Section 2. However, counter-examples to CPCP were discovered.

The most famous such example is due to Hesslow (1976). Suppose that we have a reference class of young women, all with male partners, and for whom the only contraceptive method available is the pill. We can still use Fig. 4 as an illustration, but this time we set X=Taking the Contraceptive Pill, Y=Pregnancy, and Z=Thrombosis.[7] This is shown in Fig. 6.

We will suppose that both pregnancy and taking the contraceptive pill cause thrombosis, but that the probability of getting thrombosis is higher for those who are pregnant than for those who take the pill. Here preventing the occurrence of one of the causal factors (i.e. taking the pill) does not help to avoid thrombosis. Stopping taking the pill in this population makes pregnancy very likely, and that in turn gives a higher probability of thrombosis. So it may well be the case that stopping taking the pill increases the probability of thrombosis. This is in sharp contrast to the heart disease case, where either stopping smoking, or stopping eating fast food reduces the probability of getting heart disease. This shows that it is by no means such a simple matter to deduce results from our multi-causal model of Fig. 4, as applied to heart disease, which will justify the preventative strategies normally recommended by cardiologists. In the Hesslow example, which is also a multi-causal model illustrated by Fig. 4, a similar preventative strategy, i.e. giving up taking the contraceptive pill, far from helping to prevent thrombosis would increase the chance of thrombosis occurring. There is thus a significant mathematical problem here, which we will state precisely and solve in Sections 6, 7 and 8.

Returning to the discussion of probabilistic causality, let us confine ourselves to two pronged multi-causal forks and, as a final simplification, assume that X, Y and Z

[7] It might be objected that Fig. 4 is not appropriate as a model for the Hesslow example, since in this case X, i.e. taking the pill has a causal influence on Y, i.e. pregnancy. So an arrow should be added, joining X to Y, as in Fig. 7. We will discuss this possibility in Section 5, but leave it aside for the moment.

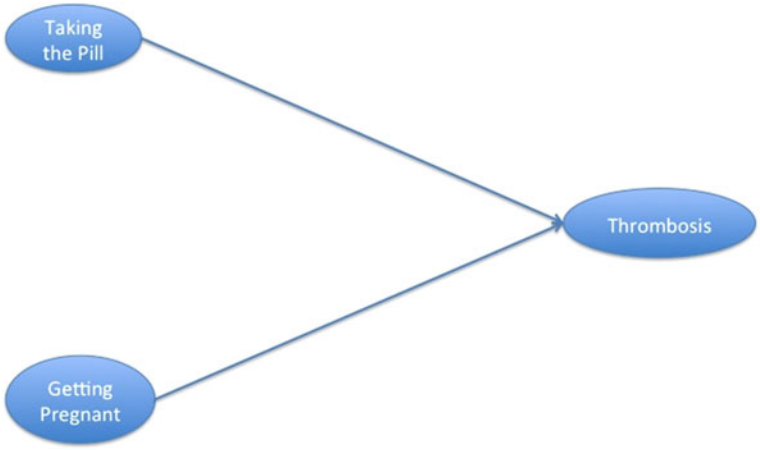

Fig. 6 Multi-Causal Fork for Hesslow Example

are bivariate variables taking the values 0 or 1. So people are classified into smokers (X=1) or non-smokers (X=0), and similarly in the other cases. This is obviously a simplification, since in the earlier table of results regarding smoking and lung cancer, the doctors were classified into various groups according to the amount that they smoked. However, as our project is to study the relations between causality and probability, it seems best to take a simple, though quite realistic, case first, and then to consider the effects of adding more complexities later. We will refer to the two pronged multi-causal fork with binary variables as our simple model.

In the context of this simple model, CPCP takes the form:

$$\text{If X causes Z, then } P\left(Z=1\middle|X=1\right) > P\left(Z=1\middle|X=0\right) \qquad (1)$$

Informally this states that if X is an indeterministic cause of Z, then the occurrence of X raises the probability of Z occurring. If we take X as smoking and Z as lung cancer, then estimating probabilities from the observed frequencies in the table given earlier, we have that P(Z=1 | X=1) is about ten times P(Z=1 | X=0).

Let us now consider the Hesslow counter-example in terms of our simple model. As before, we set X=Taking the Pill, Y=Getting Pregnant, and Z=Thrombosis. Now consider (Eq. 1). P(Z=1 | X=1) is the probability of getting a thrombosis for those taking the pill. By assumption, this is a positive but quite small probability. Next consider P(Z=1 | X=0). If X=0, then it is very probable that Y=1, that is to say that those who don't take the pill have a high probability of getting pregnant, since there is no other method of contraception. However, P(Z=1 | Y=1) is by assumption much greater than P(Z=1 | X=1). So we are likely to have P(Z=1 | X=0)>P(Z=1 | X=1), which contradicts (Eq. 1).

The existence of Hesslow's and related counter examples to CP^2 constitutes the main problem of probabilistic causality. This problem has not led the proponents of probabilistic causality to abandon their approach. They have rather sought for ways of modifying CP^2 so that the counter examples are excluded and the principle still holds. We will next try to formulate our own solution to the problem and then compare it to suggestions, which have already been made in the literature.

We are interpreting the probabilities in causal models objectively, and this means that all the probabilities are defined within some reference class. We can therefore ask the question: 'For which reference class or reference classes does (Eq. 1) hold?' The variables X, Y, Z are all associated with some underlying reference class **S**. However, Hesslow's counter-example shows that (Eq. 1) does not in general hold for this reference class **S**. This is because the causal effects of X on Z are disturbed by the causal effects of Y on Z, given that X and Y are themselves causally linked. This suggests that, in order to make manifest the causal effect of X on Z, we have to hold Y fixed. Since we are dealing with the simple case of binary variables, we have just two cases Y=0 and Y=1. Our proposal then is to divide **S** into two disjoint reference classes **S** & (Y=0) and **S** & (Y=1), and to claim that (Eq. 1) holds for each of these two reference classes, but not necessarily for **S** itself.

The division of a reference class into two disjoint reference classes can be illustrated by the example of the reference class consisting of a sequence of throws of a standard fair die. This reference class can be divided into the reference class consisting of those throws, which have an odd result, and the reference class consisting of those throws which have an even result. In Hesslow's example, the underlying reference class **S** say consists of a set of young women living with male partners in a situation in which taking the pill is the only contraceptive method available. This can be divided into the following two disjoint reference classes: **S** & (Y=0), i.e. the set of those who do not become pregnant in the time period under consideration, and **S** & (Y=1), i.e. the set of those who do become pregnant. Now, in both these reference classes (Eq. 1) holds. Consider **S** & (Y=0), which is the set of young women who do not become pregnant. Those who take the pill (X=1) have a higher probability of getting a thrombosis than those who do not (X=0) because of the side effects of the pill. Consider next **S** & (Y=1), which is the set of young women who do become pregnant. In this case it might be objected that there are no members of this set who take the pill (X=1). However, it could be replied that the pill is unlikely to be 100 % effective, and that, because of the side effects of the pill, someone who both took the pill and became pregnant is likely to have a higher probability of getting a thrombosis than someone who became pregnant without having taken the pill. So (Eq. 1) again holds.

So our suggestion is the following. Suppose we are dealing with a multi-causal fork, as in Fig. 3, in which a number of possibly interrelated indeterministic causes (X_1, X_2, … X_n) are combining to produce an effect Z. We can only draw conclusions about probabilities from the claim that X_1 causes Z, if we assume that the values of the other indeterministic causes (X_2, … X_n) are fixed. So these probabilistic conclusions hold in each member of a partition of the underlying reference class. From this it does not follow in general that they hold in the reference class as a whole, though this may be true in some cases. This shows that the Hesslow counter-example is closely related to the Simpson paradox.[8]

This proposed solution to the problem has many points in common with those put forward in the probabilistic causality research programme. Let us now briefly compare our solution with those of Cartwright (1979) and Eells (1991). Cartwright formulates what could be considered a version of CPCP in 1979 (p. 26) and Eells

[8] A good recent discussion of Simpson's Paradox is to be found in Bandyoapdhyay et al. (2011).

in 1991 (p. 8). The general idea of these formulations is the same as that presented here, but there are the following two differences. First of all both Cartwright and Eells formulate CPCP as a necessary and sufficient condition of the form 'X causes Z iff …, whereas we formulate it as a sufficient condition of the form 'If X causes Z, then …'. An 'iff' formulation would be necessary if we were trying to define causality in terms of probability, but, as already stressed, we are not attempting to give a definition of causality in terms of probability which we regard as impossible, but rather to establish a connection between indeterministic causality and probability. Secondly both Cartwright and Eells make use of the notion of a complete set of causal factors. Speaking of the effect E, Cartwright speaks of (p. 26): "A complete set of causal factors for E …", while Eells writes (p. 86):

> "In assessing X's causal relevance to Y we have to hold fixed all the other factors F_1, …, F_n, that are causally relevant to Y, independently of X, and then observe the probabilistic impact of X on Y…."

By contrast our formulation is model relative. We only consider the causal factors, which are included in the model. There may be further causal factors, which are not in the model. A similar point of view is to be found in Twardy and Korb (2004), who, using a rather different terminology, write (p. 242):

> "… the concept of objective homogeneity does the real work in helping us make sense of probabilistic causality, *if only in combination with a known causal structure*." (Our italics)

The objection to the formulations of Cartwright and Eells is that we can never know whether we have discovered all the causal factors operating, and so their versions of CPCP can never be applied in practice. To this it might be replied that, equally, if our model is incorrect because it omits some important factors, then the model relative application of CPCP would likewise be incorrect. Now, of course, any proposed causal model, like any scientific hypothesis, might be wrong. What we should do about this is to test every conjectured model as severely as we can, and change those models, which are refuted, until, hopefully, we reach a well-corroborated model which is adequate for our purposes.

The point of CPCP is to enable models involving indeterministic causality to be tested. In the deterministic case there is no problem. Suppose we want to test the deterministic claim that A causes B. All we have to do is to check that the *ceteris paribus* conditions are satisfied, and instantiate A. If B follows, the claim is confirmed, while if B does not follow, then the claim is refuted. In the case of indeterministic causality, things are similar but a bit more complicated. We have first to derive some probabilistic conclusions from our causal model using an appropriate version of CP^2. We can then test these probabilistic conclusions against the data by using statistical tests. This shows why the 'if … then' formulation of CPCP is all that is required. An 'iff' formulation is not needed.

This then is our approach, which is in the tradition of probabilistic causality and is based on a version of the Causality Probability Connection Principle. In the next section we will compare it to the quite different approach advocated by Pearl.

4 Pearl's alternative approach

In his 2011, Pearl gives an exposition of what he calls 'the structural theory of causation'. Section 33.5 of the paper deals with the question of structural versus probabilistic causality. Abbreviating probabilistic causality to PC, Pearl writes (p. 714):

> "… the PC program is known mainly for the difficulties it has encountered, rather than its achievements. This section explains the main obstacle that has kept PC at bay for over half a century, and demonstrates how the structural theory of causation clarifies relationships between probabilities and causes."

The main obstacle is of course what Pearl calls the 'probability raising' trap, or (p. 714) "the idea that causes raise the probability of their effects". The problem, according to Pearl is that philosophers have tried to express the relationship 'raises the probability of' in the language of probability theory by means of inequalities like $P(E \mid C) > P(E \mid \neg C)$ or equivalently $P(E \mid C) > P(E)$.[9] This is a mistake, however, because (Pearl 2011, p. 715): "the relationship 'raises the probability of' is counterfactual (or manipulative) in nature, and cannot, therefore, be captured in the language of probability theory." In order to express this relationship we need to use the language of the *do*-calculus, introduced by Pearl, which goes beyond probability theory. As Pearl himself says (2011, p. 715):

> "The way philosophers tried to capture this relationship, using inequalities such as
>
> $$P\left(E\middle|C\right) > P(E)$$
>
> was misguided from the start—counterfactual 'raising' cannot be reduced to evidential 'raising' or 'raising by conditioning'. The correct inequality, according to the structural theory …, should read:
>
> $$P\left(E\middle|do(C)\right) > P(E)$$
>
> where $do(C)$ stands for an external intervention that compels the truth of C. The conditional probability $P(E \mid C)$ … represents a probability resulting from a passive observation of C, and rarely coincides with $P(E \mid do(C))$."

So Pearl's main idea seems to be that the problem of probability raising is solved by replacing $P(E \mid C)$ by $P(E \mid do(C))$. One might expect him therefore to go on to show that the counter-examples to probability raising such as Hesslow's counter-example can be eliminated by making this move. However, Pearl does not do this, and does not mention any of the well-known counter-examples to probability raising in his 2011 article. We will continue our exposition of Pearl's own argument in a moment, but first it seems interesting to see how his suggestion applies to the Hesslow counter-example as formulated earlier in the paper. As a word of warning we should say that this investigation would not be accepted as legitimate by Pearl

[9] Our formulation of CP^2 given above is an example of this.

because our formulation uses multi-causal forks, which, because they are non-Markovian, he does not accept as valid. Still the investigation is not without interest, and we will accordingly carry it out.

So let us replace (Eq. 1) by (Eq. 2)

$$\text{If X causes Z, then } \mathrm{P}\Big(\mathrm{Z}=1\Big|do(\mathrm{X}=1)\Big) > \mathrm{P}\Big(\mathrm{Z}=1\Big|do(\mathrm{X}=0)\Big) \qquad (2)$$

As before our underlying reference class consists of a set of young women all with male partners in a situation in which the only method of contraception is the pill. Before, we imagined that we were simply observing which of the women took the pill, so that the ordinary probabilistic conditioning P(Z=1 | X=1) seemed appropriate. However, we could instead imagine a situation, in some very authoritarian country, in which an active intervention was made by the government. Some women, perhaps those who, according to the government, have wrong political opinions or ethnic character, would be forced under police supervision to take the pill [*do* (X=1)], while others, whose children it was thought would be more useful to the state, would be prevented from using the pill [*do*(X=0)]. In this new situation, the counter-example which applied to (Eq. 1) would apply in just the same manner to (Eq. 2). Those who were forced to take the pill would have a lower probability of getting thrombosis than those who were prevented from taking the pill, because the latter would have a much higher probability of becoming pregnant and so having pregnancy induced thrombosis. It follows from this that use of the *do*-calculus on its own does not solve the conundrums of CPCP. Pearl must be making some further assumptions—which of course is the case. Let us now examine what these further assumptions are.

The key further assumption is that, when we are dealing with indeterministic causes, we should use a structural causal model. These models are described in Pearl (2000), and they are of two types. The first type, which could be called *observable*, consists of a set of observable parameters, which are connected to their parents by a functional equation involving an error or disturbance term. These error or disturbance terms are assumed to be independent, and from this it follows that the Markovian assumption is satisfied so that the network is a Bayesian network. But what about cases where the Markovian assumption is known not to be satisfied—for example Salmon's interactive forks? Pearl proposes to deal with these by introducing latent, unobservable variables. Referring to the parents of a variable X_i as PA_i, Pearl writes (2000, p. 44):

> "If a set PA_i in a model is too narrow, there will be disturbance terms that influence several variables simultaneously and the Markov property will be lost. Such disturbances will be treated explicitly as 'latent' variables … . Once we acknowledge the existence of latent variables and represent their existence explicitly as nodes in a graph, the Markov property is restored."

So Pearl's second type of structural causal model could be called *latent*, because it involves latent, unobservable variables as well as observable variables. He thinks that any significant, but apparently non-Markovian, causal model can be reduced to a latent structural causal model. There might indeed be some non-Markovian models, which could not be so reduced, but Pearl does not think they would be of any use. As he says (2000, p. 62):

> "… we confess our preparedness to miss the discovery of non-Markovian causal models that cannot be described as latent structures. I do not consider this loss to be very serious, because such models—even if any exist in the macroscopic world—would have limited utility as guides to decisions."

Now the multi-causal forks, which we described earlier, are not in general structural causal models. They involve only observable variables, but the Markov assumption is not always satisfied. So Pearl would suggest that such models be reduced to structural causal models by, for example, introducing unobservable latent variables. We will consider how this might be done in a moment, but let us now return to Pearl's solution to the problem of whether causes raise the probabilities of their effects. Essentially his approach is that we should formulate the problem within a structural causal model, and we can then calculate the value of $CE = P(y \mid do(x)) - P(y)$. Sometimes it will be greater than zero and sometimes not. However, no additional assumption along the lines of CPCP needs to be made. In the case of observable structural causal models, the calculation of CE can definitely be carried out. As Pearl says (p. 717): "The solution follows immediately from the identification of causal effects in Markovian models …" Pearl is a bit more cautious in the case of latent structural causal models. He writes (p. 717): "The solution is less obvious when P is defined over a proper subset W of V, where {V−W} represents the unmeasured variables." However he thinks that there are results which (p. 717) "reduce this problem to algorithmic routine."

Such then is Pearl's proposed solution to the problem of whether causes raise the probabilities of their effects. We will now present some criticisms. The main difficulty in Pearl's approach seems to us to be his assumption that, whenever we are handling indeterministic causes, we should do so by introducing a structural causal model. In some cases, of course, structural causal models may be quite appropriate, but, in other cases, it might be simpler and easier to use different kinds of causal model, such as an interactive fork, or a multi-causal fork. Multi-causal forks are very simple causal models, which apply in a straight-forward way to well-known examples of the use of indeterministic causality in medicine, such as the causal factors of heart disease. They can be handled quite easily. So why should they be banned? As we know, Pearl would reply that the use of such non-Markovian models is unnecessary, because they can easily be replaced with Markovian models by adding latent variables. In the next section we will consider how this might be done in the case of our simple model of a multi-causal fork giving causal factors for heart disease.

5 Restoring the Markov condition by adjusting the model

In Section 1, we described 3 methods, which could be used to restore the Markov condition in networks for which the Markov condition failed. These were: (1) eliminating variables, (2) adding arrows, and (3) adding variables. Method (1) was illustrated by the colonoscopy example of Fig. 2. Let us now examine how, using such methods, we might try to restore the Markov condition in the case of two-pronged multi-causal forks, as illustrated by Fig. 4.

For multi-causal forks, method (1), i.e. eliminating variables, does not seem appropriate. So let us try method (2). We can than add an arrow joining X to Y, as shown in Fig. 7.

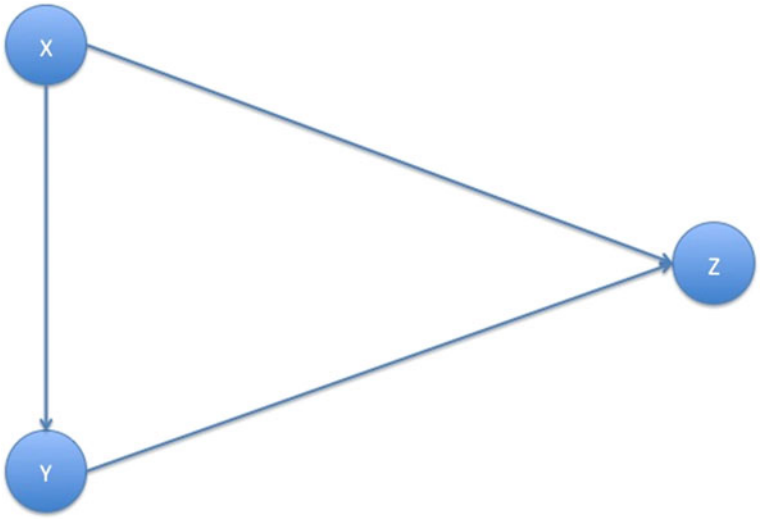

Fig. 7 Adding Arrow to Multi-Causal Fork

Now this does seem appropriate for the Hesslow example. If we apply it, we transform Fig. 6 into Fig. 8, by adding the additional postulate that taking the pill has a causal influence on getting pregnant.

Now this additional postulate is plainly correct, since taking the pill prevents pregnancy, and prevention is of course a form of causal influence. Indeed, if our main aim had been to analyse this example, there would have been no need to introduce a non-Markovian model. The Markovian model of Fig. 8 is clearly satisfactory. Indeed in his 2001, Hitchcock analyses the Hesslow example using a model, which is an elaboration of Fig. 8 (see Hitchcock 2001, p. 364, Fig. 1). His analysis in terms of the distinction between net effect and component effect is very convincing. Reassuringly Hitchcock tells us (2001, p. 366) that "birth control pills are now considerably safer than when Hesslow's example was first presented."

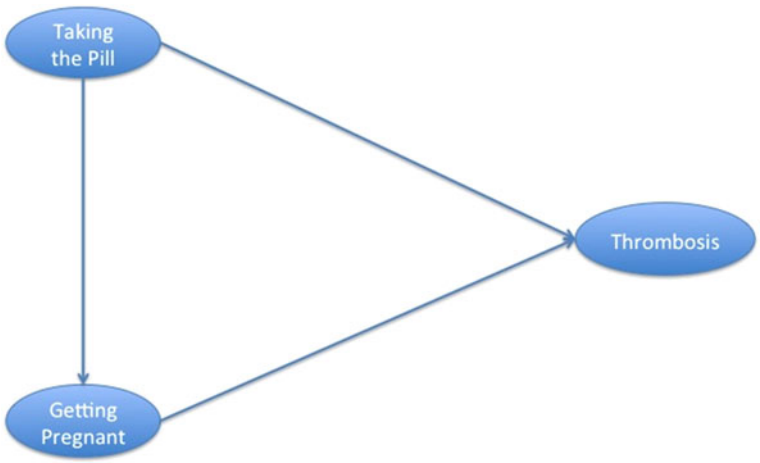

Fig. 8 Adding Arrow in Hesslow Example

However, our main aim was not in fact to analyse the Hesslow example, but rather the heart disease example of Fig. 5. The point of introducing the non-Markovian model of the Hesslow example, as shown in Fig. 6, was to provide a simple and vivid illustration of the pitfalls of non-Markovian models. Obviously anyone advocating non-Markovian models should find some way of bringing their pitfalls to light, so that these pitfalls can be avoided. The main claim of the paper is that a non-Markovian causal model is suitable for the heart disease example of Fig. 5. So the key question is whether the Markov condition can be restored for this model.

Of course, once again, we could try adding an arrow as in Fig. 7. In this case, this would amount to making the claim that smoking causes the eating of fast food. Such a claim, in contrast to the corresponding claim in the Hesslow case, is not at all plausible. smoking and eating fast food are indeed correlated, but does the first cause the second? It is just possible that having a dose of nicotine may cause a craving for the consumption of food high in salt, sugar, and saturated fat, but there is no physiological evidence for such a causal pathway. Nor is there any evidence for other causal pathways which would justify an arrow of causal influence joining smoking to eating fast food (or vice versa). In this case, then, it seems we should try what is anyway Pearl's preferred method, and introduced a latent variable U (= unobservable) between X and Y, as in Fig. 9.

Now there would be no mathematical difficulty involved in such a move. The problems, which arise here, have an empirical, or scientific, character. In general, problems of this kind arise as soon as we start using unobservable variables in modelling some observable phenomenon. We can illustrate these problems by considering a simple case involving ordinary rather than random variables. Suppose we are modelling some observable quantity y, and, to do so, take into account n observable variables x_1, x_2, …, x_n. We then postulate the following model

$$y = f(x_1, \ x_2, \ \ldots \ , \ x_n) \tag{3}$$

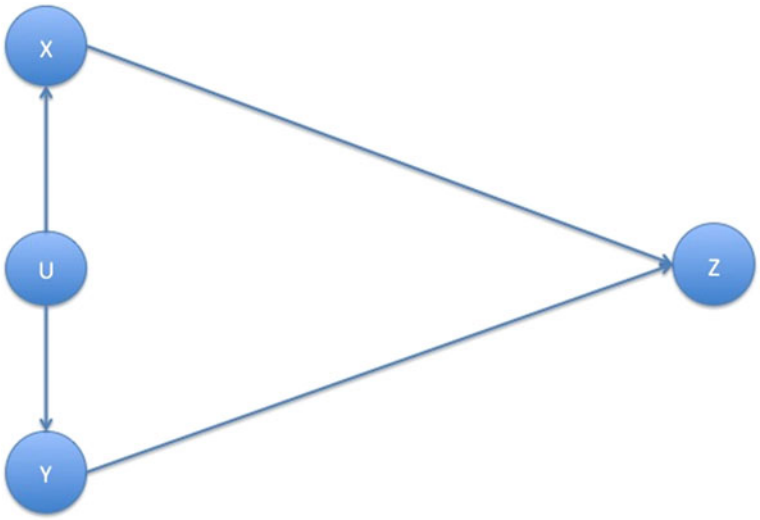

Fig. 9 Adding Latent Variable to Multi-Causal Fork

Unfortunately when this model is tested out, it is found that its predictions differ quite dramatically from observation. However, not at all daunted, we decide to adjust the model by adding a latent or unobservable variable u.

$$y = f(x_1, \ x_2, \ \ldots \ , \ x_n) + u \tag{4}$$

Since u is unobservable, we postulate that its value is given by y - f(x_1, x_2, …, x_n). We then conclude that the adjusted model (Eq. 4) now agrees exactly with observation. Obviously such a procedure would be pseudo-science rather than science.

We give this example not in order to argue that the use of latent or unobservable variables is always wrong. On the contrary, there are many examples in which the use of such variables has proved fruitful. Our aim is rather to warn of dangers associated with the use of such variables. They can all too easily lead to converting empirical testable models into pieces of pure mathematics. In order to avoid such a danger, some steps are needed. The unobservable variable U should be interpreted in some way so that it can be checked whether there is anything, which corresponds to U in the real world. If possible, some independent method of measuring U should be developed. Then, most of important of all, any claims of the form 'U causes X' should be tested against evidence to see whether they really hold or not. Merely writing down such claims without bothering to test them against data is to carry out fiction or fantasy rather than science.[10]

Similar warnings against the dangers of the use of unobservable variables are to be found in Korb et al (2004). In the context of a critique of determinism, they consider the model

$$Z = a_1X + a_2Y + U$$

and comment as follows (p. 324):

> "But, Z is not a strict function of any of X or Y or the combination of the two: there is a residual degree of variation, described by U. U is variously called the residual, the error term, the disturbance factor, etc. Whatever it is called, once we add it into the model, the model is deterministic, for Z certainly is a function – a linear function, of course – of the combination of X, Y and U. Does this make the physical system we are trying to model with the equation (or, Bayesian network) deterministic? Well, only if as a matter of fact U describes a variable of that system. Since as a matter of actual *practice* U is typically identified only in negative terms, as what is 'left over' once the influences of the other parents of Z have been accounted for, and since in that typical practice U is only ever measured by measuring Z and computing what's left over after our best prediction using X and Y, it is simply not plausible to identify this as a variable of the system."

[10] This remark may seem exaggerated. Yet there are many learned and highly mathematical papers published in leading journals which devote themselves to constructing models for fantasy examples such as 'A spell cast by Merlin caused the prince to turn into a frog.' In our view such work is valueless. Clearly in the real world spells do not cause princes to turn into frogs. Why should causality, as imagined in such a fantasy world, have anything to do with causality in the real world? Causal modellers should devote themselves to genuine scientific examples, which arise in the real world. There are a rich variety of these, and there is consequently no need to bring in the consideration of purely imaginary examples.

Bearing these potential dangers in mind, let us look at the result of introducing a latent or unobservable variable U into our simple heart disease model. This is illustrated in Fig. 10.

The first step in dealing with the model of Fig. 10 is to try to find some interpretation of the unobservable variable U. One possibility would be to interpret U as a measure of the psychological disposition to go for immediate gratifications without regard for any long-term negative consequences. Such a disposition might be described as 'improvidence'. It seems plausible that improvident people might enjoy the pleasures of smoking and eating fast food without taking account of the long-term negative consequences on their health. But while such an account sounds reasonable enough at a common sense level, it is by no means easy to establish that there really is such a psychological disposition and to find some way of measuring it. Suppose, however, we manage to overcome these problems, we have still got to establish empirically that U causes smoking, and U causes eating fast food. Now, in general, it is by no means easy to establish relations of indeterministic causality in medicine. Think of the case of 'smoking causes lung cancer'. This is now generally accepted, but for decades it was a highly controversial claim, and it took a great deal of evidence to convince the community as a whole that it was true. The problems of establishing causal relations in medicine have been discussed in an illuminating fashion in Russo and Williamson (2007). These authors propose what has come to be known as the Russo-Williamson Thesis (or RWT). There are various forms of this thesis, but a simple one is the following. In general to establish empirically that a causal relation holds in medicine, one needs the conjunction of two types of evidence, namely (i) *statistical* evidence, coming, for example, from epidemiology or randomised control trials, and (ii) evidence of *mechanisms*, usually coming from laboratory research. In Gillies (2011), the RWT is illustrated by the example: 'smoking causes heart disease'. Here the statistical evidence from epidemiology showed that there was a strong correlation between smoking and heart disease. This statistical evidence was supplemented by many laboratory studies, which elucidated the mechanisms that linked smoking to an increased tendency for atherosclerotic plaques to form in arteries. Consider again the non-Markovian causal model of

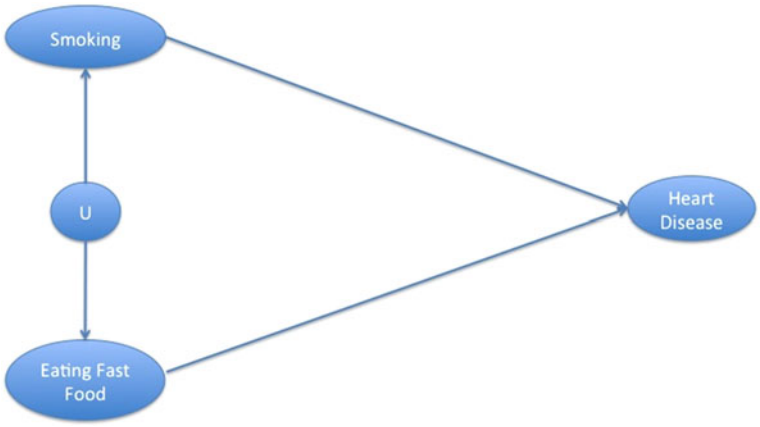

Fig. 10 Adding Latent Variable in Heart Disease Example

Fig. 5. Here the main causal links are well established by a great deal of scientific empirical evidence. Moreover, there exists a mass of data from which the probabilities used in the model can be estimated. Contrast this with the model of Fig. 10. Here a long empirical scientific investigation would be needed to establish the postulated causal links 'U causes smoking' and 'U causes eating fast food'. Moreover, in order to show that the Markov condition was indeed satisfied in this model, we would need to test whether smoking and eating fast food were independent given U. As our earlier discussion showed, this cannot be taken for granted a priori.

Of course it is an easy matter from a purely mathematical point of view to write down any number of latent variables, link these with arrows to observable variables, and postulate that the Markov condition is always satisfied. Once this is done, probabilities within the model can be calculated using standard techniques. However, such a procedure, though mathematically rigorous, may *not* be satisfactory from an empirical-scientific point of view. To guide our actions, we should use causal models, which have been rigorously tested, and are well-corroborated by evidence. It is not satisfactory to use mathematically elegant models, which have little empirical confirmation. In the examples we are considering, the non-Markovian causal model of Fig. 5 has been rigorously tested and is strongly corroborated by evidence. For the Markovian model of Fig. 10 to reach the same level of empirical corroboration, a great deal of empirical-scientific research would be needed. So we see the situation as one of a trade off. Those who favour the simpler mathematics of the Markovian model of Fig. 10, would be forced to carry out a great deal of empirical scientific work. Those, who favour the non-Markovian model, can avoid this empirical scientific work, which has already been done, but they are forced to tackle the more complicated mathematical problem of trying to handle the non-Markovian case. Can this mathematical problem be solved? In the next 3 sections, we will show that it can.[11]

6 Mathematical formulation of the problem

For simplicity, we will confine ourselves to the simple causal network illustrated in Fig. 4. Our problem is the following. We have a disease Z which is known to have two indeterministic causes X and Y. It would seem sensible in such a case to try to avoid Z by eliminating through our actions at least one of X and Y, and preferably both. It would seem to be a good strategy for doctors to advise such a course of action. However, as the pregnancy, contraceptive pill, thrombosis example shows, this advice would not always be correct. Can we then formulate some mathematical condition on X, Y and Z such that the 'common-sense' advice of eliminating at least one of X and Y is in fact correct advice?

We assume that X, Y and Z are random variables defined on some underlying probability space Ω. So the joint distribution $\{X, Y, Z\}$ is defined, and so also are the marginal distributions $\{X,Z\}$, $\{Y,Z\}$, etc. This makes our network a probabilistic network as well as a causal network, but note that it need not be a Bayesian network. For it to be a Bayesian network, the Markov condition would have to be satisfied,

[11] An informal summary of the results of these sections is given at the beginning of Section 9.

and, in this simple case, the Markov condition is that the random variables X and Y are independent. This is not the case in the two examples we considered in Figs. 5 and 6. Smoking and eating fast food are not independent, and neither are taking the pill and getting pregnant.

For the purpose of Theorem 1, we will make the further assumption that X,Y and Z are binary variables taking the values 0 or 1. Z is assumed to be a disease, and Z=1 means 'having the disease', while Z=0 means 'managing to avoid the disease'. We are also assuming that X and Y are indeterministic causes of Z. The question now is: 'what probabilistic assumptions does this allow us to make?' Here we adopt the version of CPCP, which was argued for in Section 3. This amounts to assuming that, if we set one of the variables X, Y to an arbitrary value, then a positive value of the other will increase the probability of getting the illness. We can state it mathematically by defining

$$Z_{ij} =_{def} P\Big(Z = 1 \Big| X = i,\ \ Y = j\Big) \ \text{ for } \ i,\ \ j = 0, 1$$

The assumption made on the basis of X, Y being causal factors is then the following:

$$\begin{aligned} Z_{11} &> Z_{01} \\ Z_{11} &> Z_{10} \\ Z_{01} &> Z_{00} \\ Z_{10} &> Z_{00} \end{aligned}$$

We will call this the causal factor assumption in the binary case.

What we want to prove, to justify the strategy of eliminating at least one of the causal factors in order to reduce the probability of getting the disease, is the following:

$$P\Big(Z = 1 \Big| X = 1\Big) > P\Big(Z = 1 \Big| X = 0\Big) \tag{5}$$

$$P\Big(Z = 1 \Big| Y = 1\Big) > P\Big(Z = 1 \Big| Y = 0\Big) \tag{6}$$

(Eq. 5) and (Eq. 6) both seem to hold in the smoking, fast food example, but (Eq. 5) fails in the contraceptive pill, pregnancy example. So (Eq. 5) and (Eq. 6) do not follow from the causal factor assumption. Can we formulate a mathematical condition, which, if it is added to the causal factor assumption will ensure that (Eq. 5) and (Eq. 6) follow? As will be shown in the next section, if we assume that X, Y are independent (the Bayesian network case), then (Eq. 5) and (Eq. 6) do follow. However, this independence assumption does not hold in the various examples we are considering. Theorem 1 of Section 7 shows that, if we assume P(Y=1 | X=1)> P(Y=1 | X=0), then (Eq. 5) follows, and of course if we simply reverse X, Y in this condition, we get (Eq. 6).

Now the condition P(Y=1 | X=1)>P(Y=1 | X=0) is a very reasonable one. It clearly fails in the contraceptive pill, pregnancy case, since there we obviously have P(Y=1 | X=1)<P(Y=1 | X=0), whereas it plausibly holds in the smoking, fast food case, and this could be checked empirically.

In order to prove Theorem 2, we drop the assumption that X and Y are binary variables. The binary variable assumption amounts to classifying individuals as smokers or non-smokers, or as fast food eaters or non-fast food eaters. However, this is obviously inadequate when trying to relate smoking or fast food eating to heart disease. In such an investigation, it is obviously very important to consider the quantity of tobacco that an individual smokes, or the quantity of fast food, which he or she consumes. Indeed in the statistics about smoking and lung cancer given in Section 1, the effects of smoking different amounts of tobacco were considered. It is generally held to be important to take into account the so-called 'dose relation' in assessing causality. In the case of smoking, the dose relation would be how the probability of getting the disease varies with the quantity smoked. For these reasons, it is better to take X and Y to be continuous random variables taking non-negative values.

The causal factor assumption can be given in a form appropriate to continuous random variables X and Y. The condition P(Y=1 | X=1)>P(Y=1 | X=0) now becomes a strong version of positive correlation between X, Y. Under these assumptions Theorem 2 is proved. We now give the details.

7 Two theorems

Note that if Z is 0,1 then $P(Z=1)=E(Z)$. Assume X, Y are 0,1. Put

$$e_{jk} = P(Z = 1,\ X = j,\ Y = k),\ e_{j.} = P(Z = 1,\ X = j),\ e_{.k} = P(Z = 1,\ Y = k)$$
$$p_{jk} = P(X = j,\ Y = k),\ p_{j.} = P(X = j),\ p_{.k} = P(Y = k),$$

where $e_{j.}=e_{j1}+e_{j0}$ and $p_{j.}=p_{j1}+p_{j0.}$ These imply $P(Z=1|X=j, Y=k)=e_{jk}/p_{jk}$.

(We assume the e_{jk} and p_{jk} are all non-zero.).

We shall need the following lemma.

Lemma If $1 > a > b > 0, \frac{x}{b} > \frac{y}{1-b} > 0$ then $\frac{a}{b}x + \frac{1-a}{1-b}y > x + y$.

Proof The assumptions are equivalent to $(a-b)(1-b)x>(a-b)by$ which is equivalent to what was to be proved.

If inequality (8) below is an equality, then X,Y are independent and (9) follows simply.

Theorem 1 Assume that changing either X or Y from 0 to 1, keeping the other constant, increases the conditional probability of Z. This is equivalent to the 4 inequalities (left-hand expression greater than middle 2 etc.)

$$\frac{e_{11}}{p_{11}} > \frac{e_{01}}{p_{01}},\ \frac{e_{10}}{p_{10}} > \frac{e_{00}}{p_{00}} \tag{7}$$

Further assume P(Y=1|X=1)>P(Y=1|X=0) or

$$\frac{p_{11}}{p_{1.}} > \frac{p_{01}}{p_{0.}} \tag{8}$$

then P(Z=1|X=1)>P(Z=1|X=0) or

$$\frac{e_{1.}}{p_{1.}} > \frac{e_{0.}}{p_{0.}} \tag{9}$$

There is a similar theorem for conditioning on *Y*.

Proof Summing the first and fourth inequalities of (7) gives

$$\frac{e_{11}}{p_{1.}} + \frac{e_{10}}{p_{1.}} > \frac{e_{01}}{p_{0.}} \frac{p_{11}\Big/p_{1.}}{p_{01}\Big/p_{0.}} + \frac{e_{00}}{p_{0.}} \frac{p_{10}\Big/p_{1.}}{p_{00}\Big/p_{0.}} \tag{10}$$

Put $\frac{p_{11}}{p_{1.}} = a, \frac{p_{01}}{p_{0.}} = b$ and $x = \frac{e_{01}}{p_{0.}}$, $y = \frac{e_{00}}{p_{0.}}$, then inequalities (8) and (7) imply $1 > a > b > 0, \frac{x}{b} > \frac{y}{1-b} > 0$. Applying the lemma to the r.h.s. of (10) gives

$$\frac{e_{11}}{p_{1.}} + \frac{e_{10}}{p_{1.}} > \frac{e_{01}}{p_{0.}} + \frac{e_{00}}{p_{0.}} \Rightarrow \frac{e_{1.}}{p_{1.}} > \frac{e_{0.}}{p_{0.}}.$$

A generalisation We now give a generalisation in which *Z* remains a binary random variable, while *X* and *Y* are now continuous non-negative random variables.

Assumption 1: The probability that *Z=1* conditional on *X* and *Y* increases if either *X* or *Y* increases.

$$P\Big(Z = 1\Big|X = a, Y = b\Big) \geq P\Big(Z = 1\Big|X = c, Y = d\Big) \text{ if } a \geq c, b \geq d.$$

Assumption 2: if $u>v$ then the conditional probability of *Y* given $X=u$ stochastically dominates the conditional probability of *Y* given $X=v$, that is

$$P\Big(Y \leq x\Big|X = v\Big) \geq P\Big(Y \leq x\Big|X = u\Big) \text{ if } u > v.$$

The second assumption is a strong version of positive correlation between *X*,*Y*. It implies that $E(Y \mid X=x)$ increases with *x*.

Lemma Put $A_i=a_1+a_2+\ldots a_i$, $C_i=c_1+c_2+\ldots c_i$. If $A_n=C_n$, $C_i \geq A_i$, $b_{i+1} \geq b_i$, $i=1,\ldots,n-1$ then $\sum_{i=1}^{n} a_i b_i \geq \sum_{i=1}^{n} c_i b_i$.

Proof

$$\sum_{i=1}^{n} c_i b_i = C_n b_n - \sum_{i=1}^{n-1} C_i (b_{i+1} - b_i) \leq A_n b_n - \sum_{i=1}^{n-1} A_i (b_{i+1} - b_i) = \sum_{i=1}^{n} a_i b_i.$$

Theorem 2 Under assumptions 1 and 2 the higher the value of *X*, the more likely *Z* is to occur.

Proof Suppose that Y can only take the values $y_1<y_2\ldots<y_n$. Then

$$\begin{aligned} P\left(Z=1\middle|X=u\right) &= \frac{P(Z=1,X=u)}{P(X=u)} = \sum_{i=1}^{n}\frac{P(Z=1,X=u,Y=y_i)}{P(X=u)} \\ &= \sum_{i=1}^{n}\frac{P\left(Z=1\middle|X=u,Y=y_i\right)P(X=u,Y=y_i)}{P(X=u)} \\ &= \sum_{i=1}^{n}P\left(Z=1\middle|X=u,Y=y_i\right)P\left(Y=y_i\middle|X=u\right) \\ &\geq \sum_{i=1}^{n}P\left(Z=1\middle|X=v,Y=y_i\right)P\left(Y=y_i\middle|X=u\right) \end{aligned}$$

by assumption 1. Now, putting

$$c_i = P\left(Y=y_i\middle|X=v\right),\ \ a_i = P\left(Y=y_i\middle|X=u\right)\ \ with\ \ u>v$$

assumption 2 implies that the conditions of the Lemma are satisfied. We thus have

$$P\left(Z=1\middle|X=u\right)\geq\sum_{i=1}^{n}P\left(Z=1\middle|X=v,Y=y_i\right)P\left(Y=y_i\middle|X=v\right)=P\left(Z=1\middle|X=v\right).$$

8 Comment on the conditions of Theorem 1

The conditions given in Theorem 1 (with *X*,*Y*, and *Z* binary again) are sufficient for (9), but not necessary, as can be seen from the following example in which (8) is violated, but (9) still holds.

The underlying probability distribution is fully defined by $P(X=0)=0.5$,

$$\mathrm{P}\left(\mathrm{Y}=1\middle|\mathrm{X}=0\right)=\mathrm{p}_{01}\Big/\mathrm{p}_{0.}=0.5>\mathrm{P}\left(\mathrm{Y}=1\middle|\mathrm{X}=1\right)=\mathrm{p}_{11}\Big/\mathrm{p}_{1.}=0.4$$

thus violating (8). Once again putting $Z_{ij}=_{def}P(Z=1\mid X=i, Y=j)$ and having $Z_{11}=0.8$, $Z_{01}=Z_{10}=0.5$, $Z_{00}=0.2$ we obtain

$$\begin{aligned} &\mathrm{P}\left(\mathrm{Z}=1\middle|\mathrm{X}=1\right)=\left(\mathrm{p}_{11}\mathrm{Z}_{11}+\mathrm{p}_{10}\mathrm{Z}_{10}\right)\Big/\mathrm{p}_{1.}=0.62> \\ &\mathrm{P}\left(\mathrm{Z}=1\middle|\mathrm{X}=0\right)=\left(\mathrm{p}_{01}\mathrm{Z}_{01}+\mathrm{p}_{00}\mathrm{Z}_{00}\right)\Big/\mathrm{p}_{1.}=0.35. \end{aligned}$$

The reason is that the difference between $P(Y=1|X=0)$ and $P(Y=1|X=1)$ is much smaller than the difference between the Z_{ij} if, for *i* or *j*, 0 is replaced by 1. $P(Z=1|Y=1)=0.63>P(Z=1|Y=0)=0.36$ can also be shown.

9 Conclusions

Having stated our results in a mathematically precise way in the preceding 3 sections, let us now look at them in a more qualitative fashion. The conditions under which the theorems are proved could be stated roughly and informally somewhat as follows. Let us say that two indeterministic causes X and Y are *associated* if, in the binary case, the presence of one increases the probability of the presence of the other, and if, in the continuous case, they are strongly positively correlated. Two indeterministic causes X and Y are *opposed* if, in the binary case, the presence of one decreases the probability of the presence of the other, and if, in the continuous case, they are negatively correlated. The conditions of being associated or opposed are quite intuitive, and it would be easy to check from data whether they held in a particular case. If X and Y are associated indeterministic causes of Z, the two theorems show that it is a good strategy, in order to avoid Z, to make sure that either X or Y or both do not occur, or that their effects are pre-empted if they do occur.

There is, however, an objection to the claim that the results of our theorems justify the avoidance strategies we have described.[12] One of the results of theorem 1 is that $P(Z=1 \mid X=1)>P(Z=1 \mid X=0)$ [inequality ⋆ say], and this is taken to justify the strategy of trying to avoid the disease Z by setting X=0, i.e. giving up smoking. However, it could be objected that the inequality ⋆ might hold, but eliminating X will not reduce the probability of getting the disease. Let X=1 represent yellow teeth, and Z=1 lung cancer. The inequality ⋆ is satisfied, since people with yellow teeth tend to be smokers, and hence have higher rates of lung cancer. But whitening your teeth will not reduce your probability of getting lung cancer. This is not, however, a counter-example to the claims we have made, since, we are assuming not just inequality ⋆, but also that X causes Z, and tooth colour is not a cause of lung cancer. This alleged counter-example is instructive, however, because it reinforces the familiar point that statistical claims on their own are often not action guiding, and one may need causal assumptions as well as statistical ones to justify actions. It is indeed a fundamental characteristic of causes that they are action-related.[13]

What we have shown in this paper is that it is possible to develop a non-Markovian causal model for a well-known medical situation, that this model is both well corroborated empirically, and mathematically tractable, and that, in particular, we can draw action-guiding conclusions from the model. In the mathematical theories of causal networks so far developed, there has been an almost exclusive focus on cases in which the Markov condition is satisfied. So one of the important features of the two theorems just given is that they show that interesting results can be obtained in at least some cases in which the Markov condition is dropped. It seems likely that more interesting results covering this situation could be obtained in the case of more complicated networks – for example in the case of multi-causal forks with more than two indeterministic causes.

However, in pointing to a possible use of non-Markovian models, we do not want to make the dogmatic claim that it is impossible to obtain similar results using

[12] This objection was made, by an anonymous referee, to an earlier version of this paper. I have quoted the objection, more or less verbatim, from the referee's report.

[13] For more on this, see Gillies (2004).

Markovian models. We have done no more than pose a challenge to the advocates of the exclusive use of Markovian models to solve the problem dealt with in this paper within their preferred scheme. However, as stressed above in Section 5, a satisfactory solution requires that the model used is *both* mathematically tractable *and* empirically well corroborated. Our non-Markovian model of Fig. 5 satisfies both these conditions, and any satisfactory Markovian model would have to satisfy them both as well.

We would also like to recommend the type of example considered in this paper as a very suitable field of study for causal modellers. Heart disease is still the number one killer in most developed countries. Medical science has made considerable advances in its study, and these depend on the use of indeterministic causality and multi-causal forks. The very important Framingham study, which has carried out investigations into heart disease continuously since 1948 (see Levy and Brink 2005), has provided a whole mass of data concerning possible causal factors of heart disease. Yet strange to say, there have been very few attempts to create causal models for this data. One of the rare and admirable exceptions is Korb et al (2004).[14] Instead of causal models, only traditional statistical models have been employed on the Framingham data. Surely the development of causal, and hence action guiding, models here would be a step forward.

Acknowledgments Earlier drafts of this paper were read at the *International Workshop on Causal Inference in the Health Sciences*, which Maria Carla Galavotti and Raffaella Campaner organised in Bologna on 27-28 May 2011, and at a meeting of the Kent-UCL Causality group, held in UCL on 11 August 2011. Many comments were received at these meetings – some favourable, and some highly critical, indicating the controversial nature of the material. Later we received further comments, again some favourable and some highly critical, on subsequent drafts of the paper. We have tried to take into account both types of comment in revising the paper, and would like to thank those who made comments, particularly Carlo Berzuini, Raffaella Campaner, Nancy Cartwright, Brendan Clarke, David Corfield, Philip Dawid, Maria Carla Galavotti, Phyllis McKay Illari, Judea Pearl, Federica Russo, Jon Williamson, and several anonymous referees.

Appendix

Definitions of Terms Used

A *network* or *net* is a directed acyclic graph.

The nodes or vertices of a network are variables, which are denoted by capital letters, e.g. X, Y, Z, A, B, ….

If an arrow joins two nodes of a network A, B (see Fig. 11), then A is said to be a *parent* of B, and B is said to be a *child* of A. Children, children of children, etc. of A are said to be *descendants* of A.

If an arrow joining any two nodes A, B (see Fig. 11) of a network means that A has a causal influence on B, then the network is said to be a *causal network*.

If the set of variables of a network, X_1, X_2, …, X_n say, are random variables all defined on some underlying probability space and so having a joint distribution, then the network is said to be a *probability network*.

[14] A Google search for causal models for the Framingham data produced only Korb et al (2004). Of course we may have missed some other papers, but there cannot be many of these.

Fig. 11 Parent and Child

The *Markov Condition* is satisfied for a node A of a network, if A, conditional on its parents, is probabilistically independent of any other set of nodes in the network not containing any of A's descendants.

A probability network in which every node satisfies the Markov condition is said to be a *Bayesian network*.

In a Bayesian network, the parents of a node are said to *screen it off* from the other nodes of the network except its descendants.

If a causal network is also a probability network, it is said to be a *causal probability network*, *or causal model*. When the term 'causal network' is used in this paper with no further qualification, it will be assumed to be a causal probability network.

Note that when Pearl introduced the term Bayesian network or Bayes network in his 1985b, he used it to refer to causal Bayesian networks. In fact Pearl wrote (1985b, p. 330):

> "Bayes networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by conditional probabilities."

We here, following a later convention, defined Bayesian networks purely probabilistically, so that the arrows in a Bayesian network need not represent causal influences. However, when the term 'Bayesian network' is used in this paper with no further qualification, it will be assumed to be a causal Bayesian network.

A causal model, in which every node satisfies the Markov condition, is said to be a *Markovian causal model*.

A causal model, in which at least one node does not satisfy the Markov condition, is said to be a *non-Markovian causal model*.

References

Bandyoapdhyay, P. S., Nelson, D., Greenwood, M., Brittan, G., & Berwald, J. (2011). The logic of Simpson's paradox. *Synthese, 181*, 185–208.

Campaner, R., & Galavotti, M.C. (2007). Plurality in causality. In P. Machamer & G. Wolters (eds.), *Thinking about causes from Greek philosophy to modern physics*. University of Pittsburgh Press, Ch. 10, pp. 178–199

Cartwright, N. (1979). Causal laws and effective strategies. Reprinted in *How the laws of physics lie*. Oxford: Oxford University Press, 1983, pp. 21–43.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Oxford University Press.

Cartwright, N. (1995). False idealisation: a philosophical threat to scientific method. *Philosophical Studies, 77*, 339–352.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist, 84*, 242–264.

Codell Carter, K. (2003). The rise of causal concepts of disease. Case Histories. Ashgate.

Doll, R., & Peto, R. (1976). Mortality in relation to smoking: 20 years' observations on male British doctors. *British Medical Journal, 2*, 1525–1536.

Eells, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.

Galavotti, M. C. (2010). Probabilistic causality, observation and experimentation. In W. J. Gonzalez (Ed.), *New methodological perspectives on observation and experimentation in science* (pp. 139–155). A. Coruña: Netbiblo.

Gillies, D. (2004). An action-related theory of causality. *The British Journal for the Philosophy of Science, 56*, 823–842.

Gillies, D. (2011). The Russo-Williamson thesis and the question of whether smoking causes heart disease. In Illari, Russo, and Williamson, 2011, pp. 110–125.

Good, I.J. (1961). A causal calculus I. *British Journal for the Philosophy of Science*, 11, 305–318. Reprinted in I.J.Good, *Good thinking. The foundations of probability and its applications*. Minneapolis: University of Minnesota Press, pp. 197–217.

Good, I.J. (1962). A causal calculus II. *British Journal for the Philosophy of Science*, 12, 43–51. Reprinted in I.J.Good, *Good thinking. The foundations of probability and its applications*. Minneapolis: University of Minnesota Press, pp. 197–217.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations, 11, pp. 1–12. Reprinted in D.F. Hendry & M.S. Morgan (eds.), *The foundations of econometric analysis*. Cambridge University Press, 1995, pp. 477–490.

Hennig, C. (2010). Mathematical models and reality – a constructivist view. *Foundations of Science, 15*, 29–48.

Hesslow, G. (1976). Discussion: two notes on the probabilistic approach to causality. *Philosophy of Science, 43*, 290–292.

Hitchcock, C. (2001). A tale of two effects. *Philosophical Review, 110*(3), 361–396.

Hitchcock, C. (2010). Probabilistic causality. *Stanford encyclopedia of philosophy* (http://plato.stanford.edu).

Illari, Phyllis, McKay, Russo, Federica, Williamson, J. (eds) (2011). *Causality in the sciences*. Oxford University Press.

Kim, J.H. & Pearl, J. (1983). A computational model for combined causal and diagnostic reasoning in inference systems. *Proceedings of the 8th International Joint Conference on AI* (*IJCAI-85*), pp. 190–193.

Korb, K.B., Hope, L.R., Nicholson, A.E., Annick, K. (2004). Varieties of causal intervention. *Pacific Rim International Conference on AI'04*, pp. 322–331.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society B, 50*, 157–224.

Levy, D., & Brink, S. (2005). *A change of heart. Unraveling the mysteries of cardiosvacular disease*. New York: Vintage Books.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems. Theory and algorithms*. New York: John Wiley.

Pearl, J. (1982). Reverend Bayes on inference engines: a distributed hierarchical approach. *Proceedings of the National conference on AI, ASSI-82*, 133–136.

Pearl, J. (1985a). How to do with probabilities what people say you can't. *Proceedings of the Second IEEE Conference on AI Applications*. Miami, Fl., pp. 6–12.

Pearl, J. (1985b). Bayesian networks: a model of self-activated memory for evidential reasoning. *Proceedings of the Cognitive Science Society*, Ablex, pp. 329–34.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence, 29*, 241–288.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems. Networks of plausible inference*. San Mateo, California: Morgan Kaufmann.

Pearl, J. (2000). *Causality. models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, J. (2011). The structural theory of causation. In Illari, Russo, and Williamson, 2011, pp. 697–727.

Popper, K. R. (1963). *Conjectures and refutations. the growth of scientific knowledge*. London: Routledge & Kegan Paul.

Reichenbach, H. (1956). In M. Reichenbach (Ed.), *The direction of time*. Berkeley: University of California Press. 1971.

Russo, F. (2009). *Causality and causal modelling in the social sciences*. New York: Springer.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*(2), 157–170.

Russo, F., & Williamson, J. (2011). Generic versus single-case causality: the case of autopsy. *European Journal for Philosophy of Science, 1*, 47–69.

Salmon, W. (1978). Why Ask, : "Why?"? An inquiry concerning scientific explanation. Reprinted in Salmon, 1998, pp. 125–141.
Salmon, W. (1980). Probabilistic causality. Reprinted in Salmon, 1998, pp. 208–232.
Salmon, W. (1998). *Causality and explanation*. Oxford: Oxford University Press.
Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction and search*. New York: Springer Verlag.
Sucar, L. E., Gillies, D. F., & Gillies, D. A. (1993). Objective probabilities in expert systems. *Artificial Intelligence, 61*, 187–203.
Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.
Suppes, P. (1986). Non-Markovian causality in the social sciences with some theorems on transitivity. *Synthese, 68*(1), 129–140.
Twardy, C. R., & Korb, K. B. (2004). A criterion of probabilistic causality. *Philosophy of Science, 71*, 241–262.
Williamson, J. (2005). *Bayesian nets and causality.* Oxford: Oxford University Press.
Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford: Oxford University Press.