A diagnostic advice system based on pathophysiological models of diseases

W.J. ter Burg, P. Lucas,
E. ter Braak

UU-CS-1999-04

A Diagnostic Advice System based on Pathophysiological Models of Diseases[*]

Willem Jan ter Burg
University Medical Centre
Heideberglaan 100, 3584 CX Utrecht, The Netherlands
E-mail: W.J.terBurg@dit.azu.nl

Peter Lucas
Department of Computer Science, Utrecht University
Padualaan 14, 3584 CH Utrecht, The Netherlands
E-mail: lucas@cs.uu.nl

Edith ter Braak
Department of Internal Medicine, University Medical Centre
Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
E-mail: E.terBraak@digd.azu.nl

Abstract

Medical decision-support systems in which uncertainty plays an essential role are increasingly based on the formalism of probabilistic networks. Although this formalism is very powerful, the construction of actual networks is not straightforward, and requires the availability of clearly structured medical domain models as a starting point. In this paper it is argued that medical pathophysiological knowledge constitutes a good start for the development of such models, even though pathophysiological knowledge is semantically different from probabilistic knowledge. Two models concerning anaemia, which are part of a broad system covering the domain of anaemia, are discussed to illustrate the general approach.

Keywords: medical decision support, probabilistic networks, Bayesian belief networks.

1 Introduction

The proper clinical management of diseases in patients requires the handling of large quantities of uncertain medical knowledge. It is, therefore, not surprising that the representation and manipulation of uncertainty has been one of the key issues in the development of systems for medical decision support. Although a large number of different approaches to the representation of uncertain medical knowledge have been proposed in the past, it is increasingly believed by researchers that probability theory offers one of the best foundations for building

[*] This research is supported by the Technology Foundation STW, applied science division of NWO, and the technology programme of the Ministre of Economic Affairs. Participants are: W.J.P.P. ter Burg, E.W.M.T. ter Braak, O Ying-Lie, W.A.J.J. Wiegerinck, M.J. Nijman, J.P. Neijt, and H.J. Kappen.

medical decision-support systems [3, 4]. A major advantage of the probabilistic approach is the seamless integration with conventional medical statistics, such as the possibility to use clinical data for assessing uncertainty.

However, it is well-known that the construction of probabilistic models may impose unrealistic demands with respect to the amount of patient data required, a problem that is even more significant when dealing with rare disorders [3]. Such disorders are usually sparsely represented in clinical databases. The general problem of probability assessment may, in principle, be alleviated by adopting simplifying probabilistic independence assumptions, thus relaxing the requirements with respect to the number of patients needed. However, gathering relevant probabilistic independence information from various sources is a problem in itself.

The research described in this paper is part of a project in which a diagnostic advice system is being developed, covering the broad domain of anaemia. Lack of sufficient patient data, even at our large university hospital, made it necessary to investigate the potentials of modelling techniques to alleviate part of this problem. In this paper, we describe whether pathophysiological knowledge, as found in the medical literature, offers a sufficient basis to guide the development of probabilistic models. Moreover, we show how separately developed disease models can be integrated to obtain larger models. Pathophysiological models of two causes of anaemia, vitamin B_{12} and folic acid deficiency, have been designed for this purpose. The formalism of probabilistic networks (Bayesian belief networks) [2, 4] has been used for the representation of the resulting probabilistic models.

The structure of this paper is as follows. In the following section, the problem of diagnosing anaemia in patients is briefly reviewed. Next, in Section 3, the theory of probabilistic networks is introduced, and models of anaemia are described from various perspectives. In Section 4, the results of a preliminary evaluation of the resulting model are presented, and, finally, in Section 5 we discuss future research.

2 Clinical diagnosis of disease resulting in anaemia

In the context of this research, *anaemia* was defined as a blood hemoglobin level below the lower bound of the reference values, considered clinically relevant and prompting further diagnostic procedures. This condition may give rise to a variety of general symptoms and signs, like fatigue, paleness, palpitations, exercise-related shortness of breath, and dizziness. Anaemia in general may be due to: (1) impaired production of red blood cells, (2) loss of red blood cells, or (3) increased degradation of red blood cells. The processes of production and degradation of red blood cells have been unravelled the last century up to the molecular level. The same is true for many forms of anaemia for which the pathophysiology is described quite detailed in the medical literature.

Examples of causes of anaemia are vitamin B_{12} and folic acid deficiency. Vitamin B_{12} and folic acid are both essential elements in DNA synthesis. Deficiency primarily affects cells with a rapid turnover, especially blood forming cells. Impaired DNA synthesis results in disturbed, ineffective formation of large red blood cells with certain features (megaloblastic cells), and in anaemia. Two diseases that may cause vitamin B_{12} deficiency are pernicious anaemia and atrophic gastritis. *Pernicious anaemia* may lead to vitamin B_{12} deficiency, because auto-antibodies destroy acid-producing cells (parietal cells) of the stomach and also intrinsic factor. Both gastric acid and intrinsic factor are required for the proper absorption of vitamin B_{12}. In *atrophic gastritis*, acid-producing cells and pepsinogen producing cells are destroyed, resulting in impaired absorption of vitamin B_{12}.

Information from the patient history and physical examination, as well as results of a battery of laboratory tests, and sometimes imaging techniques, are required for making a diagnosis in the domain of anaemia. It is not always clear which tests are most informative for effective diagnosis making; a decision-support system could be very helpful in this respect.

3 A probabilistic model of anaemia

The central question is how pathophysiological knowledge from the literature can be used as a basis for the representation of the uncertainty underlying diagnosis of anaemia. This means that we have to look for a mapping of pathophysiological knowledge to the formalism of probabilistic networks. The formalism of probabilistic networks is first briefly introduced. Next, we discuss how designed pathophysiological models can be mapped to probabilistic networks.

3.1 Probabilistic networks

A *probabilistic network*, also call Bayesian belief network, is a directed acyclic graph $G = (V_G, A_G)$, consisting of a set of vertices $V_G = \{V_1, \ldots, V_n\}$, representing discrete stochastic variables, and a set of arcs $A_G \subseteq V_G \times V_G$, representing stochastic influences among variables [3]. If a binary variable V assumes the value *true*, this is denoted by v; if it assumes the value *false*, this is denoted by $\neg v$. Assigning a value to a variable is called *instantiating* the variable.

On the set of vertices V_G is defined a joint probability distribution P that can be factorised according to the topology of the graph as follows:

$$P(V_1, \ldots, V_n) = \prod_{i=1}^{n} P(V_i | \pi(V_i))$$

where $\pi(V_i)$ denotes the set of parent vertices of vertex V_i. As this equation indicates, it is only necessary to represent local probabilistic information $P(V_i|\pi(V_i))$ to obtain the joint probability distribution P, because the variable V_i is assumed conditionally independent of all predecessors with the exception of the parents, given the parents.

A probabilistic network not only offers a compact representation of uncertain knowledge; it can also be viewed as an architecture for probabilistic inference. To determine the effect on the probabilistic information of evidence, several algorithms have been proposed that take a probabilistic network and a set of evidence E as input, and produce as a result an updated joint probability distribution [2, 4]. It appears that the marginal probability of a variable V_i in the Boolean algebra that results when evidence E is given, denoted by $P'(V_i)$, is exactly the same as the a posterior probability distribution given the evidence, i.e. $P(V_i|E)$:

$$P'(V_i) = P(V_i|E)$$

When the variable V_i represents a diagnostic category, the probability $P'(V_i)$ represents the updated, marginal probability distribution after processing evidence. It can be compared to the prior marginal probability $P(V_i)$, and this indicates to what extent a diagnosis is (dis)confirmed by this evidence E.

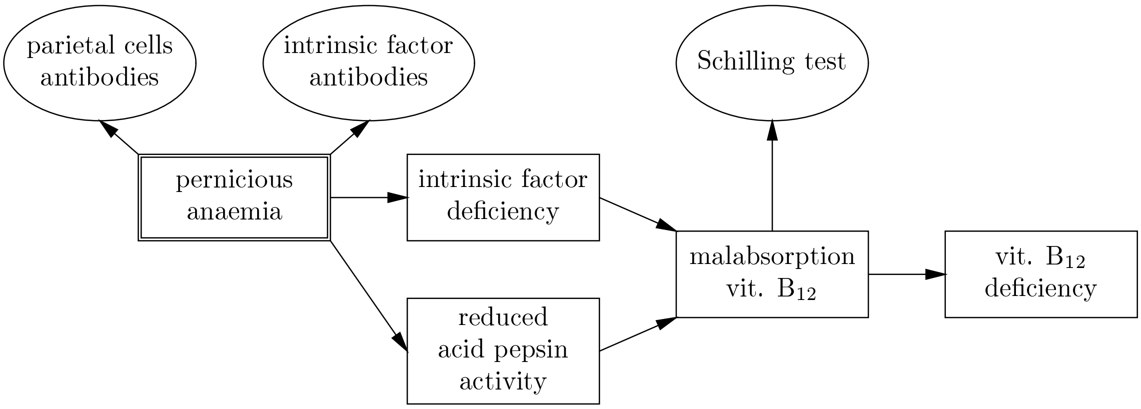
parietal cells antibodies
intrinsic factor antibodies
Schilling test
pernicious anaemia
intrinsic factor deficiency
malabsorption vit. B_{12}
vit. B_{12} deficiency
reduced acid pepsin activity

Figure 1: Small part of a qualitative model of vitamin B_{12} deficiency with disease category pernicious anaemia.

3.2 Design of pathophysiological models

Although the medical literature on anaemia includes elaborate descriptions of the mechanisms underlying diseases causing anaemia, no complete pathophysiological models are available. Such models were designed from scratch, using directed acyclic graphs as the main modelling tool; in the graphs, arcs had the meaning of cause-effect relationships, and vertices were used to represent disease states, test results and disease categories. A small part of a large pathophysiological model of vitamin B_{12} deficiency is shown in Figure 1. Ellipses indicate test results, single boxes indicate states, and the double box represents a diagnostic category. The resulting graphs can be seen as *qualitative*, *causal models* of anaemia.

3.3 Mapping pathophysiological to probabilistic models

Although causal knowledge is not the same as probabilistic knowledge, notions of causality may guide the development of probabilistic models [1]. Let $c_1 \Rightarrow e, \ldots, c_m \Rightarrow e$ represent multiple causes c_j for a given effect e, such as '*pernicious anaemia* ⇒ *intrinsic factor deficiency*'. The first step in the transformation concerns viewing cause c_j and effects e as stochastic variable V_{c_j} and V_e, respectively, with associated value domain. For most variables, the adopted value domain was binary, with as possible values *yes* and *no*, or *normal* and *abnormal*.

Since a causal relation ⇒ induces a stochastic dependency, causal direction can be employed as a heuristic in the modelling process. The main difficulty in the modelling process lies in the representation of possible interactions among the various causes; for the situation described above we would have to assess the probability distribution $P(V_e|V_{c_1}, \ldots, V_{c_m})$, which means assessing to what extent causes augment or weaken each other. As an example, consider the interaction between the causes *intrinsic factor deficiency* and *reduced acid pepsin activity* and the effect *malabsorption vit.* B_{12}. These variables, which are conditionally independent given *pernicious anaemia*, augment each other. In this way, the design of local interactions was guided by knowledge of pathophysiological interactions.

Except local interactions, modelling the combined (probabilistic) effect of several causes, interactions among causes also occur when considering combining separate disease models. Medical experts tend to divide knowledge concerning a medical domain into subdomains with relatively small overlap. When each subdomain is represented as a probabilistic network, the

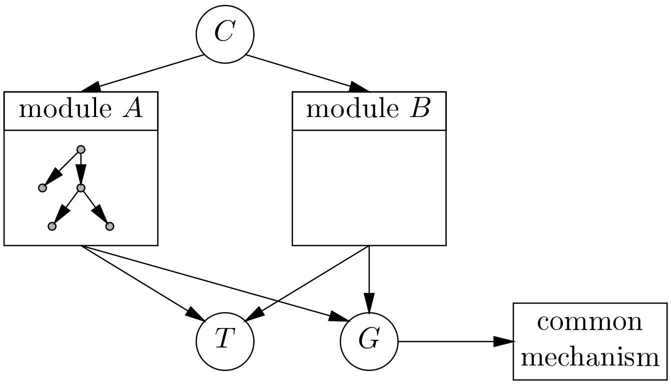

Figure 2: Modularisation of anaemia domain.

resulting combined model will have a *modular structure*, with some shared variables. The connection between the modules is accompliced by different types of shared variables. The possible situations are schematically depicted in Figure 2, where two modules describing separate disease mechanisms are shown. Each module contains one or more diagnostic categories, like *pernicious anaemia*. Note that we do not assume disorders to be mutually exclusive, because they have not been represented as values of a single stochastic variable.

The vertex with label C conditions the modules on particular evidence; in medicine, usually evidence concerning age or sex. The amount of probabilistic information to be specified is limited, because only prior probabilistic information $P(C)$ has to be specified. Propagation of this evidence through the modules can be done efficiently, if the module content is not too complex (i.e. the module graph is sparse). Note that the modules A and B can be developed separately, provided that the variable C is always instantiated, and that the variables T, G and all variables constituting the common mechanism are not instantiated. Under these conditions, the modules are independent. After separate constructions, the modules can be put together.

A major problem in putting modules together lies in the exponential number of probabilities that must be specified for vertices like T and G. Vertex T represent a test result; since many tests are shared by different diseases causing anaemia, there may be many incoming arcs to T. By using a logical OR operator for combining probabilistic influence, a quite compact representation can be obtained. Consider, for example, the situation in Figure 3.a, where three arcs enter the vertex D; the probability distribution $P(D|A,B,C)$ models an OR gate, i.e.

$$P(d|A,B,C) = \begin{cases} 0 & \text{if } A = \neg a, B = \neg b, C = \neg c \\ 1 & \text{otherwise} \end{cases}$$

The resulting behaviour of probabilistic influence of the stochastic variables A, B and C on D is called a *noisy-OR gate* [4]. It holds for the variable D modelling a noisy-OR gate that:

$$P(d) = \sum_{A,B,C} P(d|A,B,C)P(A,B,C) = 1 - P(\neg a)P(\neg b)P(\neg c)$$

Now, consider the situation in Figure 3.b, where two arcs enter the vertex D_1, and two arcs enter the vertex D_2; both $P(D_1|A,B)$ and $P(D_2|D_1,C)$ model an OR gate. We have that:

$$P(d_2) = \sum_{D_1,C} P(d_2|D_1,C)P(D_1,C) = 1 - P(\neg a)P(\neg b)P(\neg c)$$

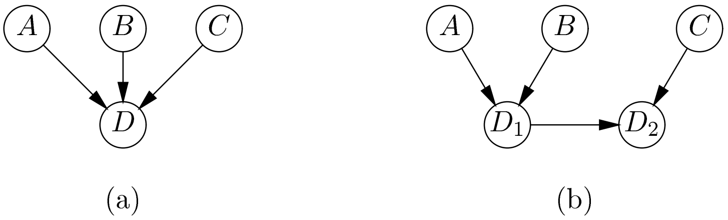

Figure 3: Disjunctive interaction.

Note that $P(d_1) = 1 - P(\neg a)P(\neg b)$ and that A, B and C, and also D_1 and C, are independent. We conclude that $P(d) = P(d_2)$: these two ways of achieving noisy-OR behaviour are equivalent. The second method of specifying a noisy-OR gate saves in the number of probabilities to be specified, which is linear in the number of incoming arcs instead of exponential, and also in the time complexity of probabilistic inference. In general, for binary variables we have that a specification of a probability distribution following the approach in Figure 3.a requires 2^{n+1} probabilities; for specifications similar to the one in Figure 3.b only $4n$ probabilities are required, where n is equal to the number of arcs. For diagnostic tests for which sensitivity and specificity information is available, probabilities $P(A)$, $P(B)$ and $P(C)$ may in fact represent test results for a single modular group of disorders, later to be combined to obtain a noisy-OR gate.

The variable G is also a shared variable, but since this concerns a common mechanism, and only few mechanisms among various modules are identical, the number of incoming arcs will not be very large. Rather, there will be a number of different common mechanisms shared by various modules, and there will be no need for use of the noisy-OR gate here.

The module structure shown in Figure 2 actually reflects the structure of the probabilistic model of two causes of anaemia, vitamin B_{12} and folic acid deficiency. Variables T are for example platelets and leukocytes counts; the variable G in this case stands for megaloblastic erythropoiesis, i.e. the production of megaloblastic cells, which occurs in both deficiencies.

4 Evaluation

Above, we have described the process of designing a probabilistic model, based on pathophysiological, qualitative knowledge from the medical literature. The constructed qualitative models provide strong support for the correctness of the structure of the resulting network model. However, by carrying out simulation experiments with patient cases, we have also investigated whether the actual probabilistic influences were as expected. Moreover, the a posteriori probabilities have been compared to frequency information (such as frequency of positive and negative test results, i.e. sensitivity and specificity, for certain diseases) from the literature. Based on the results of these experiments, the probabilistic model was slightly adjusted.

Furthermore, we have done a number of experiments with patient data, where in particular the effects of the systematic gathering of evidence, as is common practice in medicine, has been studied. The results for two patients are shown in Table 1. As can be seen, during three subsequent visits of a patient to the clinic, additional test results become available. For patient A the evidence slowly moves in the direction of pernicious anaemia when more specific information becomes available. In patient B, gastroscopy is the final test demonstrating

Patient A			**Patient B**		
Visit 1			*Visit 1*		
Hb	7.8	decreased	Hb	7.2	decreased
MCV	118	increased	MCV	104	increased
platelets	160	normal	reticulocytes	0.4 %	increased
leukocytes	4.1	normal	gastrointestinal symptoms		present
			neuropsychological symptoms		present
$P(\textit{pernicious anaemia})$	0.01			0.04	
$P(\textit{atrophic gastritis})$	0.01			0.03	
Visit 2			*Visit 2*		
haptoglobin	< 0.1	decreased	pentagastrin test		positive
LDH	667	increased	serum gastrin		increased
serum folate	21.1	normal	serum folate	17	normal
serum vit. B_{12}	59	decreased	serum vit. B_{12}	72	decreased
$P(\textit{pernicious anaemia})$	0.04			0.07	
$P(\textit{atrophic gastritis})$	0.03			0.07	
Visit 3			*Visit 3*		
TSH	1.4	normal	gastroscopy		atrophic gastritis
parietal cells antibodies	+++	yes			
intrinsic factor antibod.	+	yes			
$P(\textit{pernicious anaemia})$	0.82			0.00	
$P(\textit{atrophic gastritis})$	0.01			1.00	

Table 1: Results of the probabilistic model for two patients.

atrophic gastritis. The results for both patients illustrate the capability of the system to guide the medical decision-making process.

5 Discussion

In this paper, we have shown that pathophysiological knowledge from the literature can be used to support the process of making a medical diagnosis. We have also discussed how separate models of subdomains can be integrated to obtain larger models. Obviously, this approach will only work when elaborate pathophysiological descriptions are available in the literature. Clearly, this does not hold for medicine in general, but there are many medical domains where this approach will work.

The new detailed probabilistic models will be integrated with other disease models concerning anaemia, as part of the project's aim of achieving a diagnostic advice system that covers the broad domain of anaemia.

References

[1] M. Korver and P.J.F. Lucas, Converting a rule-based expert system into a belief network, Medical Informatics, 18(3) (1993) 219–241.

[2] S.L. Lauritzen and D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society (Series B), 50 (1987) 157–224.

[3] P.J.F. Lucas and L.C. van der Gaag, Principles of Expert Systems, Addison-Wesley, Wokingham, 1991.

[4] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman, San Mateo, California, 1988.