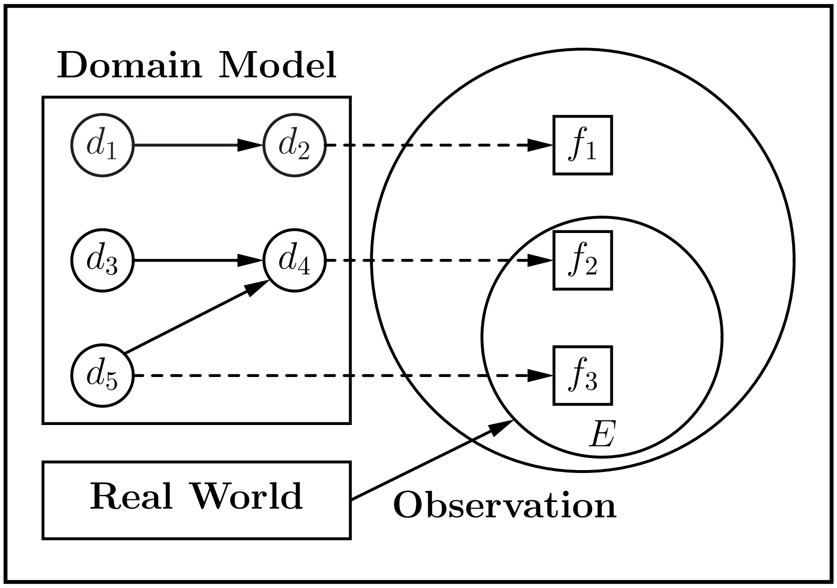
Structures in diagnosis

Lucas, Peter

1996

Structures in Diagnosis

from theory to medical application



Peter Lucas

Structures in Diagnosis

from theory to medical application

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Lucas, Petrus Johannes Franciscus

Structures in Diagnosis: from theory to medical application/
Peter Lucas
Thesis Free University of Amsterdam – with a summary in Dutch
ISBN 90-9009564-0
NUGI 855, 743
Subject headings: knowledge-based systems / expert systems /
medical decision making

VRIJE UNIVERSITEIT

Structures in Diagnosis

from theory to medical application

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus
prof.dr E. Boeker,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der wiskunde en informatica
op vrijdag 14 juni 1996 te 10.45 uur
in het hoofdgebouw van de universiteit,
De Boelelaan 1105

door

Petrus Johannes Franciscus Lucas

geboren te Voorburg

Promotor:	prof.dr J. Treur
Copromotoren:	dr F.M.T. Brazier
	dr F. van Harmelen
Referenten:	prof.dr L. Console
	prof.dr J.H.P. Wilson

To my parents and
to Linda,
for their patience . . .

Preface

It was the philosopher Jean-Jacques Rousseau's firm conviction that we live in the best possible world. A conviction which he never abandoned, not even after the catastrophic destruction of Lisbon due to the earthquake of 1755. I have, in contrast, always believed that there is not a unique best possible world, certainly not a best possible world of science. This thesis reflects this belief by devoting attention to both theory and practice, and to Computer Science as well as to Medicine, although its scope is limited to a single subject: diagnosis.

The origin of the thesis was the development of a medical expert system, called HEPAR, that is capable of assisting in the diagnosis of disorders of the liver and biliary tract. This work not only concerned the collection of relevant medical knowledge, but also the development of methods and software tools for the construction, evaluation and use of the system. The development of the HEPAR system was a joint project with Roelof Janssens. Without his continuous medical guidance, it would have been impossible. His many meticulous reviews of the gathered medical knowledge are the real basis of the quality of the system. Our regular meetings during the project were always a pleasure to me, not in the least because of our mutual interest in literature. The evaluation of the performance of HEPAR has been made possible by Rob Segaar and Paul Wilson, who were so kind to provide me with patient data, collected over a period of three years at Dijkzigt Hospital Rotterdam. Paul Wilson's help in the assessment of the evaluation results proved invaluable.

A small, insignificant attempt to formally describe the process of diagnosis was included in an earlier version of this thesis. Questions by Oscar Estévez raised by this formalization were the immediate cause for the expansion of what once was a single paragraph of text to more than 150 pages now. While it is likely that the size of my answer exceeds all his expectations, I sincerely hope that his questions have now been answered.

The last two years of the research described in this thesis have been carried out under the guidance of Jan Treur, Frances Brazier and Frank van Harmelen. Their critical support caused me to pursue research topics, which otherwise might have passed by unnoticed. I thank them all for the time and effort invested in this thesis. I also would like to thank Annette ten Teije and Ameen Abu-Hanna for their feedback on early parts of this thesis, and Luca Console for his insightful comments, which proved so useful in the production of the final version of the manuscript. Finally, I am grateful to Linda, who endured so many lonely weekends during the past two years; without her continuous encouragement and trust, this work might never have reached its end.

Peter Lucas
Utrecht, April 1996

Contents

Chapter 1

Introduction

Diagnosis pervades our lives. Even the fact that you are reading the first sentence of this thesis implies that you have made a diagnosis. Your diagnostic problem was whether or not this thesis was worth reading, with the possible solutions ‘yes’ and ‘no’. Apparently, your solution to this problem was affirmative.

Diagnosis is the subject of this thesis. Diagnosis is viewed as the interpretation of case-specific findings in the context of knowledge from a problem domain to obtain an indication of the presence and absence of defects or faults. Computer-based diagnosis was among the first applications investigated when digital computers became available more than four decades ago. Several aspects of diagnosis, however, are still unclear. In particular, it has been difficult to capture the concept of diagnosis in a precise, formal way. The formalization of diagnosis is the subject of study of the first part of this thesis. The main scientific contribution of this work is a mathematical framework of diagnosis within which various formal theories of diagnosis are analysed, and alternative forms of diagnosis are proposed.

The second part of this thesis deals with the development and evaluation of a diagnostic system in the field of hepatology, the medical field concerning the diagnosis and treatment of disorders of the liver and biliary tract. The main scientific contribution of this work is a diagnostic system that has been assessed with respect to its diagnostic accuracy.

In this chapter, the nature of the diagnostic process is sketched, and the various approaches to diagnosis described in the literature are introduced. Throughout this thesis, we shall primarily, but not only, focus on medical applications of diagnosis. Furthermore, a brief overview of diagnosis in the field of hepatology, the application field of the second part of our work, is presented. Other work in decision support of diagnosis in the field of hepatology is also described. The chapter is concluded by an overview of the thesis as a whole.

1.1 Diagnostic problem solving

Discovering what is wrong in a particular situation is one of the central activities in real life; this process is usually called *diagnosis* or *diagnostic problem solving*. It is, therefore, not at all surprising that the automation of diagnosis was one of the first subjects seriously studied when the first computer systems became available.

Diagnostic problem solving may be viewed as the process of the selective gathering and interpretation of information as evidence for or against the presence or absence of one or more defects in a system. This informal definition reveals the following aspects which are of central importance to diagnostic problem solving. Firstly, the *gathering* of information, and secondly, the *interpretation* of the gathered information for determining what is wrong, for example with a patient or a device. In medicine, the defects are disorders in a patient; in technical domains, defects are the faults or failures of a device. In medicine, the information-gathering process is usually carried out in a systematic, structured fashion, because there are an enormous number of diagnostic tests available to the clinician, that cannot all be carried out. Furthermore, some diagnostic tests cause discomfort to the patient, or even carry some risk of causing disease or death. By restricting the selection of diagnostic tests in early diagnosis to those that do no harm or cause little discomfort to the patient, as is common practice in medical diagnosis, diagnostic tests are performed only when necessary. In technical fields, it is sometimes impossible to gather certain information because of time constraints, costs involved, or physical impossibility. Although the information-gathering process is a characteristic feature of diagnosis, the interpretation of information as evidence for a diagnostic solution is a more basic aspect of diagnostic problem solving to which most of the thesis will be devoted.

The information-gathering process as well as related aspects, such as the order in which diagnostic hypotheses are generated and rejected or accepted, are sometimes referred to as *dynamic* aspects of diagnostic problem solving. They yield specific problem-solving behaviour. Establishing the actual diagnostic solution requires knowledge of what constitutes a diagnosis of a particular problem; these aspects are sometimes referred to as *static* aspects of diagnostic problem solving. This thesis focusses on these static aspects of diagnostic problem solving.

In general, diagnostic problem solving may be described using the scientific notion of the *empirical cycle*, which describes the framework underlying empirical research [Popper, 1959]. It states that empirical research encompasses (1) formulating a *hypothesis*, (2) *testing* that hypothesis and (3) *accepting* the hypothesis when it successfully passes the tests, or *rejecting* the hypothesis when it fails to pass the tests. The process may start again with (1), in which case the formulation of a hypothesis possibly involves *adjusting* a hypothesis previously rejected. In Figure 1.1, this view of diagnostic problem solving as an instance of the empirical cycle is depicted. Testing involves the application of procedures for the verification and falsification of a hypothesis using *observed findings* and domain knowledge. In general, a hypothesis may be a complex structure or mechanism. In diagnostic problem solving, however, a hypothesis is usually taken to be a collection of 'defects', assumed to be either present or absent. This simplification may not always be justified, for example because the defects may be interrelated to each other in some particular way, which could be part of the hypothesis. Nevertheless, this simplification is invariably made in diagnostic systems, and seems acceptable in the light of applications developed. A *diagnosis* may be conceived to be an accepted hypothesis concerning a particular defect or collection of defects; the results of diagnostic tests correspond to the observed findings.

The literature on diagnosis more or less follows the terminology and structure of the empirical cycle. For example, [Davis & Hamscher, 1988] views diagnostic problem solving

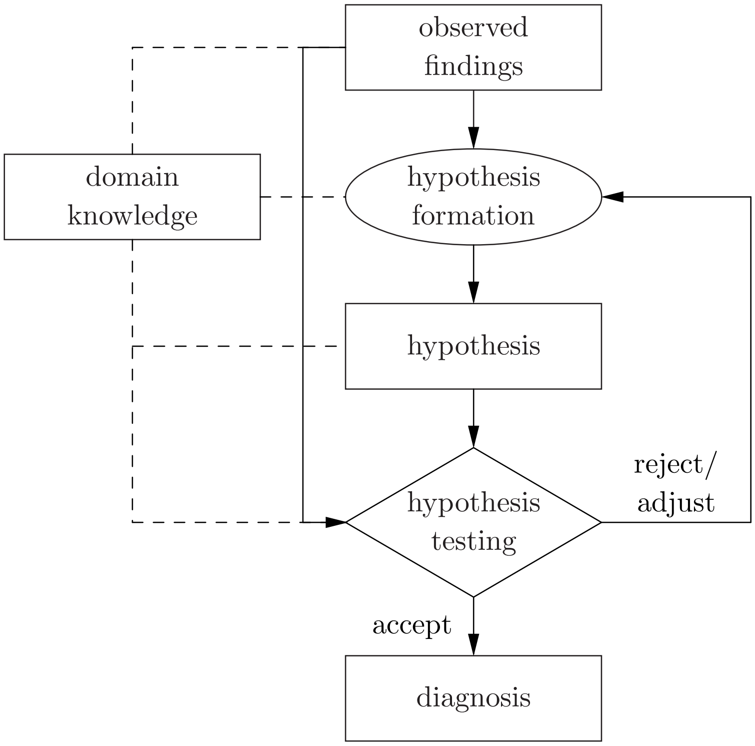


Figure 1.1: Diagnostic problem solving and the empirical cycle.

as three fundamental subproblems:

(1) Hypothesis generation (or hypothesis formation);

(2) Hypothesis testing;

(3) Hypothesis discrimination.

The subproblem of hypothesis discrimination concerns selecting from the hypotheses accepted on the basis of a measure of plausibility. This process may entail collecting additional observed findings. In this thesis, hypothesis discrimination will be called *diagnosis selection*.

The basic framework of diagnostic problem solving as the empirical cycle can be refined in several ways. For example, there may be an ordering on the set of hypotheses, such as an ordering from generic to specific, or an ordering by the value of a real-valued utility function associated with the hypotheses. A class of defects may be taken as a generic hypothesis, and a specific defect may be viewed as a specific hypothesis. Such orderings are especially useful in guiding the problem-solving process, information gathering included. For example, the process may be decomposed into several stages working from generic towards more specific hypotheses, or from hypotheses with high associated utility to those with low associated utility. It is well-known that guiding the problem-solving process, using information collected at earlier stages, may be quite effective in reducing the number of tests to be performed and may result in a step-wise reduction of the number of defects to be considered, due to the rejection of specific hypotheses motivated by the earlier rejection of associated generic hypotheses [Chandrasekaran & Mittal, 1983]. This approach to handling hypotheses and observable findings is an example of a so-called (diagnostic) *problem-solving strategy* [Newell & Simon, 1972]. Problem-solving strategies are beyond the scope of the theoretical work in this thesis, although the application of such strategies will be briefly discussed in the second part of the thesis.

1.2 Diagnostic systems

Many diagnostic systems, in particular in the field of medicine, are based on statistical methods. Since uncertainty plays an important role in many diagnostic problems, relevant work in the field of statistics is first reviewed. Next, relevant work in the field of artificial intelligence is described.

1.2.1 Statistical methods for diagnosis

In as early as the 1960s, computer programs were developed to investigate the applicability of a computerized version of Bayes' theorem to medical diagnosis [Spiegelhalter & Knill-Jones, 1984]. Diagnosis of liver disease was among the first subjects for which computer programs were constructed [Martin et al., 1960; Carlstrom et al., 1963]. Since then, the construction and validation of programs based on the application of Bayes' theorem has been undertaken by several research groups [Burbank, 1969; De Dombal et al., 1972; Gorry & Barnett, 1968; Heckerman, 1992; Knill-Jones et al., 1973; Malchow-Møller et al., 1986; Todd & Stamper, 1993]. A frequently cited medical application is the 'acute abdominal pain program' developed by De Dombal et al., a program capable of diagnosing causes of abdominal pain, such as perforated peptic ulcer [De Dombal et al., 1972; De Dombal, 1984; De Dombal et al., 1991]. In the specific form of Bayes' theorem used in the systems mentioned above, it is assumed that the elements in a diagnostic hypothesis are mutually exclusive and exhaustive, and that the observable findings that constitute the evidence are conditionally independent given the hypothesis. A resulting *diagnosis* is either a single disorder d with maximal posterior probability $P(d|E)$, where E denotes the set of observed findings, or a list of disorders, ordered by the posterior probability associated with each disorder. This form of Bayes' theorem is commonly called 'independence Bayes' or 'naive Bayes', since the conditions mentioned above, which must be fulfilled for its validity, are often not met in practice. If these assumptions fail to hold, then the computed probabilities cannot be interpreted as relative frequencies with respect to a population under study, although a diagnosis (disorders selected) may still be correct.

There is no consensus among researchers as to whether or not the simplifying assumptions severely affect the accuracy of diagnostic conclusions (cf. [Szolovits & Pauker, 1978; Heckerman, 1990; Todd & Stamper, 1993]). Presumably, the impact of the simplifications simply depends on the nature of the problem domain. Still, in almost any problem domain, certainly in medicine, dependencies among variables exist, and the representation of the domain in terms of probability or decision theory is semantically shallow, even if unrestricted probability theory is employed. These representation methods ignore the rich semantic relationships, e.g. causal or functional, in the domain, which might be explored in the process of diagnosis. From a computational point of view, the main advantages of the assumptions mentioned above are the relative ease of probabilistic assessment and the fact that the worst-case time complexity of the associated probabilistic algorithms is polynomial in the number of disorders and observable findings [Szolovits & Pauker, 1978]. This contrasts with the unrestricted situation, since probabilistic inference is $\mathcal{NP}$ hard in general [Cooper, 1990]. The reader is referred to [Lucas & Van der Gaag, 1991] and [Spiegelhalter & Knill-Jones, 1984] for a treatment of basic probabilistic concepts in diagnostic systems.

The required probabilities in systems based on the application of Bayes' theorem are usually estimated from a large (clinical) database. The size of the database and the number of items recorded per patient, used in the development of a statistical model, are important indicants for the number of conclusions that can be drawn by a statistical diagnostic system with sufficient accuracy. The probabilities may also be obtained from subjective estimates by experienced clinicians, although this may give rise to many difficulties. We will not go into the details of the problem of subjective probability assessment. The reader is referred to [Spiegelhalter et al., 1990].

Bayes' theorem has also been used as the basis for *scoring systems*. A scoring system need not be computer-based; several systems have been built which merely use a paper chart to be filled in by the physician. The model used in a scoring system is additive in nature, yielded by a logarithmic transformation of probabilities. A total score S that is computed for a patient is defined by

$$S = \sum_{j=0}^{m} \omega_j \tag{1.1}$$

where

$$\omega_j = \ln \frac{P(f_j|d)}{P(f_j|\neg d)}$$

$1 \leq j \leq m$, i.e. ω_j is equal to the logarithm of the likelihood ratio of a finding f_j given a disorder or defect d, and

$$\omega_0 = \ln \frac{P(d)}{P(\neg d)}$$

i.e. ω_0 is equal to the logarithm of the prior odds of the disorder d. The score S is equal to

$$S = \ln \frac{P(d|E)}{P(\neg d|E)}$$

i.e. the logarithm of the posterior odds of the disorder d given the evidence E. In a paper-based scoring system, the transformed probabilities ω_j, $0 \leq j \leq m$, constitute the chart. For a given patient, a total score S is computed by the sum of all individual scores; the most likely diagnosis is looked up in a table where logarithmically transformed odds are converted to probabilities using this score. For practical purposes, statistical systems based on the independence form of Bayes' theorem are only applicable if the number of variables used is relatively small, say in the order of ten to twenty.

As stated above, most of the conditions underlying the use of the independence form of Bayes' theorem are not fulfilled in medical practice, i.e. the findings in the patient are usually not conditionally independent given a disorder, the outcome space is not completely covered by the disorders distinguished, and the disorders are not necessarily mutually exclusive. The last decade, more advanced statistical methods have been introduced to deal with situations in which the assumptions mentioned above fail to hold. *Logistic models*, for example [Schmitz, 1986], can handle (in)dependencies among statistical variables; these models can be expressed as linear equations of the following form

$$T = \sum_{j=0}^{m} a_j \omega_j = \boldsymbol{a}^{\boldsymbol{t}} \boldsymbol{\omega} \tag{1.2}$$

where the vector $\boldsymbol{a}$ expresses dependencies among variables. As in equation (1.1), the quantities ω_j denote the contributions of the individual observed findings and prior knowledge in the computation of the score T. The elements in the vector $\boldsymbol{a}$ are called *shrinkage parameters*, because they cause a 'shrinkage' of the contribution to the overall evidence by the weights ω_j. Two frequently employed techniques to construct logistic models are *logistic discrimination* and *logistic regression* [Spiegelhalter & Knill-Jones, 1984; Schmitz, 1986]. At the end of the 1980s, other more advanced statistical techniques emerged for building diagnostic systems, in which the dependencies and independencies among variables distinguished in the domain can be represented explicitly, and not only implicitly as in logistic models. The (in)dependencies are taken into account in the probabilistic computations, using a graph representation of the probabilistic influences among the variables. These representation techniques are often referred to as *belief networks* [Pearl, 1988; Lauritzen & Spiegelhalter, 1987].

The main advantage of statistical methods for diagnosis over many other techniques that deal with uncertain knowledge (cf. [Lucas & Van der Gaag, 1991]) is that they have a sound mathematical basis. However, the statistical approaches to building diagnostic systems suffer from several limitations [McIntyre, 1986]:

- Statistical systems typically distinguish a small number of diagnostic categories. Only by resorting to subjective probability assessment is it practically possible to construct systems incorporating larger numbers of diagnostic categories. Examples of such systems are Pathfinder III [Heckerman, 1990; Heckerman, 1992] and Pathfinder IV [Heckerman, 1992; Heckerman et al., 1992; Heckerman & Nathwani, 1992], in which 60 and 63 disorders, respectively, are distinguished. These numbers approach the lower bound of the number of disorders in a narrow medical domain. In systems constructed from information extracted from a database, most often only general disease categories (i.e. disease groups) are distinguished. The number of distinguished disease categories is often less than twenty.

- Only a limited number of symptoms and signs are used in statistical diagnostic systems, usually only the symptoms and signs that several different diseases have in common. Again, this limitation is only enforced when probabilistic assessment is based on information from a database. Symptoms which are pathognomonic (typical) for a certain disease, such as Kayser-Fleischer rings in the cornea which may be observed in Wilson's disease, are normally not included in such systems; these systems do not operate at such a level of specificity.

- Statistical systems are difficult to transfer to other (medical) centres, because they are tailored to the characteristics of the population reflected in the database of patient cases. Usually, some form of local modification of a statistical system is necessary, which may require considerable effort [Segaar et al., 1988].

- Statistical systems provide only limited possibilities to represent the medical knowledge contained in medical textbooks; much of this knowledge is not statistical in nature. Still, some of this knowledge may be useful for building a diagnostic system. From a semantical point of view, statistical approaches are shallow.

Belief networks alleviate some of the limitations mentioned above, because they provide a better grip on the semantical analysis of a problem domain, and they are more suitable as a basis for the subjective assessment of probabilities [Breese et al., 1988; Heckerman, 1992]. Although the foundation of belief networks remains that of probability theory, with its associated limitations, the graph representation associated with a belief network is often viewed in qualitative, even causal terms. Belief networks are, therefore, frequently classified as knowledge-based systems, discussed in the next section. Belief-network theory with the associated field of decision theory provide a promising basis for developing practical diagnostic systems; yet, it is at present not completely clear how far the power and flexibility of the belief-network formalism reaches (cf. [Korver & Lucas, 1993]).

1.2.2 Knowledge-based systems for diagnosis

Knowledge-based systems are computer programs that embody knowledge, encoded by means of knowledge-representation formalisms. A knowledge-representation formalism is a formal language, with a syntax and semantics, suitable for the task at hand [Hayes, 1977]. *Expert systems* may be viewed as knowledge-based systems that are capable of producing advice in a particular problem domain, in a way and at a level comparable to that of human experts in the field [Lucas & Van der Gaag, 1991]. It is recognized by many researchers nowadays that expert systems only contain a very simplified model of expertise, although their level of competence may reach that of human experts in solving restricted problems.

The specified knowledge in an expert system is applied in order to solve problems by means of methods of automated reasoning, usually referred to as *inference* or *problem-solving methods*. In the program, the specified knowledge resides in the *knowledge base*, and the inference methods in the *inference engine* of the system.

Medicine was one of the first areas in which diagnostic expert systems were developed. Classical diagnostic medical expert systems are: INTERNIST-1 and its commercially available successor QMR, expert systems in the broad domain of internal medicine [Miller et al., 1982; Bankowitz et al., 1989], CASNET, an expert system for the diagnosis and treatment of glaucoma [Kulikowski & Weiss, 1982; Weiss et al., 1978], ANA, an expert system for digitalis therapy advice in cardiac arrhythmia [Silverman, 1975], ABEL, an expert system for the management of electrolyte and acid-base derangements [Patil, 1981;

Patil et al., 1982], and MYCIN, an expert system for the diagnosis and treatment of septicaemia and meningitis [Shortliffe, 1976]. A medical expert system similar to MYCIN, but more clearly structured due to the separation of knowledge that controls the reasoning from declarative diagnostic knowledge, is CENTAUR [Aikins, 1980; Aikins, 1983]. The MYCIN expert system has produced a major impetus to the evolution of the field of expert systems, in particular because the MYCIN project showed that the techniques used in the MYCIN program could be generalized for use in other problem domains [Shortliffe, 1976; Buchanan & Shortliffe, 1984]. Several other expert systems were developed by means of the EMYCIN (Essential MYCIN) system, which was obtained by adapting the original MYCIN system, by separating the inference engine and other facilities from its domain-specific knowledge base. Hence, the techniques applied to the MYCIN system appeared to be applicable to other domains than infectious disease. At present, programs such as EMYCIN are known as *expert system shells*. The success of the EMYCIN system inspired many researchers to develop similar systems.

Although all systems mentioned above can be viewed as knowledge-based systems for diagnosis, they are actually based on different, but related, principles, as is apparent from the descriptions available in the literature (e.g. [Buchanan & Shortliffe, 1984], [Clancey & Shortliffe, 1984], [Johnson & Keravnou, 1988] and [Szolovits, 1982] contain extensive descriptions of several diagnostic systems). Until recently, however, no theoretical framework was available to formally describe and compare the various underlying principles. At a conceptual level, it was evident that the knowledge bases of some of the systems primarily consisted of encoded human expertise in solving particular (medical) problems. Currently, the term *empirical associations* is often employed to denote such knowledge. The classical example of such a system is MYCIN [Shortliffe, 1976]. The knowledge bases of other systems, however, did not consist of such empirical associations, but rather captured a model of structure and behaviour in a domain. Such systems have been called *model-based*; the approach has also been called 'diagnosis from first principles' [Reiter, 1987]. The systems CASNET and ABEL, mentioned above, are early examples of model-based systems. Both systems contain a representation of disease progress in terms of cause-effect (causal) relationships. In a sense these cause-effect relationships capture the 'behaviour' of disease processes. Little or no structural knowledge is included in their knowledge bases. This is due to the fact that anatomical knowledge, the typical knowledge regarding structure in medicine, is merely implicitly embodied in the causal relationships. However, systems that employ structural medical knowledge exist. For example, the LOCALIZE system is an expert system that assists in the localization of peripheral nervous system lesions, by using knowledge of the structure of the nervous system [First et al., 1982].

The model-based approach to diagnosis has been particularly fruitful in exploring fault finding in electronic circuits. Early work in this field is described in [Brown et al., 1982], [Davis, 1984], [Genesereth, 1984] and [De Kleer, 1977]. The study of simple electronic circuits has yielded much insight into the nature of the diagnostic process. More importantly, one of the first formal theories of diagnosis emerged from this research: the theory of consistency-based diagnosis as proposed by R. Reiter [Reiter, 1987]. *Consistency-based diagnosis* offers a logic-based framework to formally describe the diagnosis of abnormal behaviour in a device or system, using a model of normal structure and functional behaviour. Basically, consistency-based diagnosis entails finding faulty device components that account for a discrepancy between predicted normal behaviour of the device and observed abnormal behaviour. The predicted behaviour is inferred from a formal model of normal structure and behaviour of the device.

Where consistency-based diagnosis focusses traditionally on fault finding, employing a model of normal behaviour, *abduction* has been the principal model-based technique for describing and analysing diagnosis using a model of abnormal behaviour in terms of cause-effect relationships [Console & Torasso, 1990a; Josephson & Josephson, 1994; Reggia et al., 1983; Peng & Reggia, 1990; Poole, 1988]. Early work on abduction has been done by H.E. Pople (cf. [Pople, 1973; Pople, 1977]) and D. Poole (cf. [Poole et al., 1987]). As mentioned above, causal knowledge has also been incorporated in the systems ABEL and CASNET, although diagnosis in these systems has not been described in terms of abduction. In *abductive diagnosis*, diagnostic problem solving consists of establishing a diagnosis using cause-effect relationships with a set of observed findings as the starting point. In abduction, a system reasons from effects to causes, instead of from causes to effects. Since, in such systems, reasoning from causes to effects can be accomplished using logical deduction, in a sense abductive reasoning is carried out in a direction reverse to that of deduction. As holds for consistency-based diagnosis, abductive diagnosis may also be looked upon as the core concept for a formal framework of diagnosis. The various frameworks described in the literature are either logic-based (cf. for example [Console & Torasso, 1991], [Konolige, 1992] and [Poole, 1988]) or based on set theory (cf. for example [Josephson & Josephson, 1994], [Peng & Reggia, 1990] and [Wu, 1991]).

1.2.3 A formal framework of diagnosis

The phrase 'model-based' is usually employed to denote expert systems in which diagnostic problem solving is based on using an explicit model of structure and behaviour. However, expert systems that primarily employ empirical associations are also based on some model of the problem domain. This was, in fact, already noted by Clancey in his analysis of the MYCIN system [Clancey, 1985]. Hence, the terminology, now widely in use, is confusing. Also, the phrase 'consistency-based diagnosis' and 'abductive diagnosis' are confusing, because testing consistency or performing abduction are techniques. They have some, but no inherent connection with diagnosis. As both consistency checking and abduction can be achieved by ordinary deduction (cf. [Reiter, 1987; Console et al., 1991]), these concepts do not offer perspicuous characterizations of diagnostic problem solving. Due to these limitations of current formalizations of diagnosis, in this thesis, the concept of diagnosis is studied using a set-theoretical framework, where many of the assumptions are less restrictive than in other frameworks. The main part of this thesis is devoted to exploring alternative notions of diagnosis, including those presented in the literature.

1.3 Diagnosis in hepatology

The first part of this thesis addresses the formal underpinning of the static aspects of diagnosis. The second part of this thesis focusses on the development and validation of an actual diagnostic expert system in the field of hepatology, called HEPAR [Lucas et al., 1989]. Hepatology is a subdiscipline of internal medicine concerned with the diagnosis and treatment of patients with disorders of the liver and biliary tract. Since diagnosing these disorders in patients is known to be difficult, several researchers have attempted to develop diagnostic systems to support the clinician in the clinical assessment of the patient, using various techniques. The most relevant studies will be discussed below. First, however, a brief introduction to this problem domain is provided.

1.3.1 Clinical aspects of hepatology

The basic problem in hepatology is to differentiate between disorders of the bile system (biliary obstructive disorders) and disorders of the liver cells (hepatocellular disorders), because diagnosis, treatment plans, and prognosis are quite different for the disorders in the two disease groups. For example, the typical treatment of acute cholecystitis (a biliary obstructive disorder) is surgical in nature, whereas acute hepatitis B (a hepatocellular disorder) is managed conservatively (i.e. with nonsurgical measures, such as medication). In recent decades, a large number of new diagnostic methods has become available to analyse patients with disorders of the liver and biliary tract. A number of viruses that cause several forms of viral hepatitis in the patient have been identified, and new accurate serological tests to appraise viral hepatitis in the patient have been introduced into clinical practice. Progress in immunology has yielded new tests for the detection of tissue-specific autoantibodies. New developments in ultrasound imaging of the liver and biliary tract have given clinicians an important tool, especially for the diagnosis and assessment of biliary obstructive disorders. Endoscopic retrograde cholangiopancreaticography (ERCP, visualization of the biliary tract and pancreas using an endoscope and image-forming techniques) has turned into an indispensable technique both for the diagnosis, treatment and follow-up of disorders of the pancreas and biliary system. Liver biopsy is now a frequently employed technique in the diagnosis and prognostic assessment of chronic liver disease. As a consequence, in well-equipped modern clinical centres, it is more likely now than ever before that the nature of the patient's disorder will be elucidated.

The rapid progress in hepatology, with the introduction of many new laboratory tests and diagnostic techniques, has also increased the dependency of the clinician with respect to these methods, and has apparently decreased the role of clinical experience in the diagnostic process. This is unfortunate, since a careful interpretation of data gathered by the medical interview, a thorough physical examination and a small number of laboratory tests frequently suffice for the formulation of a differential diagnosis. On the other hand, only clinicians who frequently encounter patients with liver disease, out of the vast range of possible aetiologies, will be capable of basing their decisions mainly on clinical experience. It therefore seems that the new techniques will be of particular value to the inexperienced clinician. This is only apparently true, since the careful selection of diagnostic procedures to be applied to the patient can be based on clinical experience only. Only by the physician's clinical experience can the number of supplementary diagnostic procedures, which are often uncomfortable for the patient and often expensive, be reduced. Typical examples of diagnostic procedures which should be introduced into the diagnostic process with great care are liver biopsy and ERCP.

It is widely recognized that diagnosis of disorders of the liver and biliary tract on purely clinical grounds is a difficult task [Matzen et al., 1984; McIntyre, 1986]. Clinicians have varying experience in this area, and are not all equally confident in dealing with these patients. There are several studies in the literature which report on the capability of experienced and less experienced clinicians in making a correct diagnosis for these patients. Most of these studies concern patients with jaundice classified by the clinician in two or three diagnostic categories. These studies indicate that experienced clinicians are able to differentiate between biliary obstructive and hepatocellular disease in 80 – 90% of their patients [Martin et al., 1960; Schenker et al., 1962; Berkowitz, 1964; Conn et al., 1979; Haubek et al., 1981; Theodossi et al., 1983; Theodossi, 1986]. The diagnostic accuracy is even larger if the results of ultrasound and liver biopsy are taken into account [Theodossi et al., 1983]. However, the overall diagnostic accuracy drops to about 45 – 65% when more specific diagnosis is attempted (using about ten different categories) [Stern et al., 1974; Theodossi, 1986]. Inexperienced clinicians seem to be able to make a correct diagnosis in jaundiced patients in less than 45% of cases [Theodossi, 1986].

Due to the importance of early, accurate diagnosis in diseases of the liver and biliary tract, interest in computer-based and other formal approaches to diagnosis in the patient was raised in the early decades of Computer Science. In the next section, some of the early work will be reviewed.

1.3.2 Statistical diagnostic systems in hepatology

An early example of a computer-based system for diagnosis in hepatology based on the application of Bayes' theorem was developed by F. Burbank in the late 1960s [Burbank, 1969]. His system was able to distinguish between six disorders of the liver and biliary tract: benign biliary obstruction, chronic active hepatitis, primary biliary cirrhosis, drug-induced jaundice, malignant extrahepatic obstruction and viral hepatitis. The system was tested using the original database of 52 patient cases, also used in the development of the system, each time leaving out the patient case being tested. The system correctly identified the cause of the patient's jaundice for 77% of the patients. Another, similar, early diagnostic system in the field of hepatology has been developed by Knill-Jones et al. [Knill-Jones et al., 1973].

Logistic models have been applied more recently in the development of classification systems in hepatology. Especially noteworthy is the clinical classification system developed by the Copenhagen Computer Icterus (COMIK) group, known as the Copenhagen Pocket Chart [Malchow-Møller et al., 1986; Malchow-Møller et al., 1987]. It has been based on the analysis of data of more than 1000 jaundiced patients. The Copenhagen Pocket Chart classifies a given jaundiced patient into one of four different diagnostic categories: acute non-obstructive, chronic non-obstructive, benign obstructive, and malignant obstructive jaundice, based on the values of 21 variables to be filled in by the clinician (cf. Table 8.3). The performance of this classification system has been investigated by research groups in several countries, such as in Sweden [Lindberg et al., 1987] and in the Netherlands [Segaar et al., 1988], using retrospective data from patients. These studies showed that, when taking the diagnostic conclusions of the clinician as a point of reference, the system is able to produce a correct conclusion (one of the four possible diagnostic categories), in about 75 – 77% of jaundiced patients.

The medical basis of the statistical models discussed above requires some further consideration. Many diseases of the liver and biliary tract manifest themselves by jaundice. Most diagnostic computer systems in hepatology, and databases on which they are based, therefore take jaundice of the patient as an entry condition. The development of a statistical model is usually done by collecting data of a set of items (variables), such as symptoms, signs and laboratory data of the patient, and the final diagnosis determined by a procedure taken as the 'gold' standard, for a large number of patients. The procedure selected as the 'gold' standard should have a high accuracy (sensitivity and specificity); in hepatology the histological examination of the liver and the results from ERCP are usually taken as such. Using some measure of predictive accuracy, it is next determined which variables contribute most to the discrimination of the diagnostic categories. These variables may be either discrete, for example the presence or absence of some symptom in the patient, or continuous, for example the concentration of some substance in the blood. Discrete variables are often binary, i.e. they may have one of two mutually exclusive values. Continuous variables may be transformed into a collection of binary variables by calculating one or more cut-off points, and separating the entire range of values into a finite set of (often two) intervals. For example, the cut-off point used in the Copenhagen Pocket Chart for alkaline phosphatase is 1000 U/l; the variable 'alkaline phosphatase' has been divided into two variables, one with the range $[400, 1000]$ and the other with the range $(1000, \infty)$. For the development of the Copenhagen Pocket Chart, likelihood ratios were applied to select 24 relevant variables from 107 initially given variables; this subset was again reduced to 21 relevant variables by using Bayes' theorem and logistic discrimination. In this way, the final logistic discrimination model was obtained. Below we discuss why developing the HEPAR system, given the availability of the statistical diagnostic systems discussed above, was considered worthwhile.

1.3.3 Expert systems in hepatology

A number of expert systems for the diagnosis of disorders of the liver and biliary tract have been developed in the past. MDX is a hepatological expert system in which disease hypotheses are represented as diagnostic tasks [Chandrasekaran & Mittal, 1983]. A *task* is a procedural description of an isolated problem-solving activity in terms of a collection of operations on data. If this task represents a procedure for diagnosing a group of disorders, it is referred to as a *generic task*. The main objective for the development of MDX has been to investigate the theory of problem solving according to generic tasks, rather than to develop a high performance diagnostic system. In MDX, the task of medical diagnosis is viewed as an activity which can be naturally subdivided into a number of subtasks, each task tailored to the solution of a specific part of the entire process of diagnosis. Subtasks in MDX have therefore been called *specialists*. The totality of diagnostic knowledge is distributed among various 'specialists'. There are, for example, 'specialists' for the diagnosis of extrahepatic and intrahepatic cholestasis (biliary obstruction and hepatocellular disease, respectively); the task of diagnosing hepatocellular disease is again subdivided into the task of diagnosing more specific groups of disorders, such as hepatitis and cirrhosis, etc. Each of the diagnostic 'specialists' contains declarative domain knowledge as well as control knowledge on how to deal with that knowledge. Part of the control knowledge has a local nature, in the sense that it only deals with the knowledge known to the 'specialist'; another part is concerned with passing control to other 'specialists'. Thus, problem-solving knowledge is actually embedded in the knowledge base, and not clearly separated from declarative knowledge, as was the case for MYCIN-like systems. An advantage of a knowledge-representation formalism as employed in MDX is the clear structure it imposes on its knowledge base. In contrast, when all knowledge is represented in a uniform, logic-based form, as is done in many rule-based systems, the resulting knowledge base may be difficult to maintain. However, the mixed procedural-declarative nature of the formalism may be taken to be a disadvantage. With respect to its capabilities as a diagnostic expert system, we remark that only a relatively small number of diseases are covered by MDX. Furthermore, the system has not been adequately validated with regard to either diagnostic performance or any other feature.

LITO1 is an expert system similar to the HEPAR system, in the sense that it incorporates hepatological knowledge encoded by means of production rules [Lesmo et al., 1984]. The main objective for the construction of the LITO1 system was to experiment with a sequential decision structure in order to reduce the number of laboratory tests applied to the patient [Milanese & Bona, 1984]. In this system, the diagnostic strategy is subdivided into three stages. In the first stage, only clinical data is employed in order to classify a patient into one of three categories, covering the absence (A), possibility (P), or suspicion (S) of liver disease. In the second stage, simple laboratory data are required to refine the previous rough classification. Laboratory tests are only requested for patients classified as belonging to the P group. The classification of all patients in this group having an outcome for at least one laboratory test lying outside the normal range (defined as $mean \pm 2 \cdot standard\text{-}deviation$) is revised into the S category. The remaining cases in the P category are reclassified as belonging to the A category. Finally, in the third stage of the classification process, the results of a large number of supplementary tests are considered for those patients who have been classified into the S category after the previous two stages. These patients are reclassified as belonging to the A (absence of disease) or D (presence of disease) group. The final conclusion of the expert system is an assignment of a patient case to one of two mutually exclusive categories. The performance of the LITO1 system has been estimated using data from 288 patient cases, consisting of 92 healthy subjects and 196 patients with liver disease. The system was capable of correctly classifying the patients with liver disease in 96.4% of the cases (category D); 76.1% of the healthy subjects were correctly classified (into category A). Therefore, for this test population, the false negative rate of the system was only about 4%, whereas the false positive rate was about 24%, which is unacceptably high. The successor of LITO1, called LITO2, is a more general diagnostic expert system covering the complete area of diagnosis of disorders of the liver and biliary tract [Cravetto et al., 1985]. The design of this expert system was inspired by the MDX and CENTAUR systems mentioned earlier. The designers considered existing production-rule systems unsuitable for developing such a large application, and therefore chose a task-like representation formalism, as in MDX and CENTAUR. As far as we know, the system has never been submitted to a formal evaluation of its performance, but it was later extended for educational purposes as the LIED system [Console et al., 1992].

HEPAXPERT-I is an expert system that is in routine use for the interpretation of serological tests for hepatitis A and B viral infections [Adlassnig & Horak, 1995]. This system, which only deals with a tiny portion of the hepatological domain, has been extensively validated.

Recently, in the EURICTERUS project, a research project funded by the European Union, a large database with data of patients with jaundice has been compiled [EURICTERUS, 1993]. A wide variety of techniques, such as probabilistic, neural-network and expert-system techniques, have been investigated. Unfortunately, only a limited amount of clinical data has been collected for individual patients. Furthermore, only 17 different disease entities have been distinguished in the systems developed.

1.3.4 The HEPAR project

Above, we have described relevant work in the field of hepatology. The work described in the second part of this thesis addresses the development and validation of the HEPAR system, an expert system for the diagnosis of liver and biliary disease [Lucas et al., 1989]. The knowledge base of this expert system consists of empirical associations gathered in close collaboration with the hepatologist dr. A.R. Janssens. The construction of this system commenced in 1984 as one of the first experimental applications of a rule-based expert system shell, called DELFI-2 [Lucas, 1986; Lucas & De Swaan Arons, 1987]. As a consequence, the development of the HEPAR system includes not only the implementation of a development environment, but also knowledge acquisition and system validation.

The assumptions underlying the construction of the HEPAR system were as follows [Lucas & Janssens, 1991a]:

- Clinicians not frequently confronted with patients with a liver or biliary tract disease could benefit from consulting a hepatological expert system;
- An expert system which assists the clinician in the initial assessment of the patient would be particularly valuable;
- The advice produced by the expert system should be detailed enough to be taken as the basis for the prognostic assessment and treatment of the patient's disease, possibly also supported by the computer.

As discussed in Section 1.3.1, the initial diagnostic assessment is clearly important for structuring the remainder of the diagnostic process. As shall be discussed in detail in Chapter 6, the application of simple data from the medical interview, physical examination, and laboratory tests has therefore been taken as the basis for the development of the HEPAR expert system. Using only these data, the system is capable of providing an initial assessment of the patient with a disorder of the liver or biliary tract. Based on the availability of more specific information, the HEPAR system is also capable of producing a differential diagnosis which may be used as input for the prognostic assessment of the patient. No use is made of information obtained by liver biopsy or ERCP. The HEPAR system may therefore be a valuable supportive tool for clinicians with insufficient training in the field of hepatology.

At the start of the HEPAR project, it was decided to undertake a validation study of the diagnostic performance of the resulting system. In order to render performance validation a feasible possibility, the formalism used for the representation of the medical knowledge in the HEPAR system was fixed early in the project. Two validation studies were accomplished using data from the Department of Internal Medicine II of the Dijkzigt University Hospital at Rotterdam. The two studies have been carried out between 1986 and 1990, in collaboration with a team responsible for the collection of the patient data, but not involved in the development of HEPAR. The test populations consisted of patients with jaundice, consecutively admitted to the hospital. The performance of the expert system was estimated by comparing its conclusions to the known clinical diagnoses of the patients. One important limitation of this approach to performance validation is that clinicians often based their final conclusion on information not available to the HEPAR system. Several cases were indeed encountered, in which the diagnosis was based largely on results from liver biopsy or ERCP, both being tests for which the results were not included in the HEPAR system.

In conclusion, the main reasons for developing the HEPAR system were the observed limitations of statistical methods to develop a system that could handle information at a similar level of detail as applied by clinicians in establishing an early diagnosis, and interest in studying the suitability of techniques from the field of expert systems to develop such a system. Although we had access to the COMIK database, which contains information of more than 1000 patients, there was no database available of sufficient size to be used as a basis to develop a statistical system with similar amount of detail as the HEPAR system.

1.4 Overview of this thesis

This chapter is concluded by giving an overview of the contents of this thesis.

Chapters 2 – 5 constitute the first part of this thesis. Chapter 2 provides a survey of the conceptual and formal basis of current theories of diagnostic problem solving. Diagnostic problem solving is primarily viewed as the process of handling qualitative information in establishing a diagnosis, and not as a process of probabilistic reasoning, although probabilistic information may come in at a particular stage. In the analysis of the various approaches, the meaning of the concept of 'diagnosis' is addressed. Several possible meanings are discussed.

In Chapter 3, a mathematical framework is developed with sufficient generality to capture the essential, static, aspects of diagnostic problem solving. The framework is used in an analysis of several existing theories of diagnosis underlying implementations in a number of different systems, which are described in Chapter 4.

The models of diagnosis described in the literature are shown to offer rather limited flexibility. Chapter 5 explores several possibilities of providing more flexible notions of diagnosis, in the light of practical limitations to the completeness and accuracy of encoded knowledge in real-life expert systems.

Chapters 6 – 8 constitute the second part of this thesis. Chapter 6 introduces the problem of diagnosis of disorders of the liver and biliary tract from a medical perspective, and reviews the basic notions and techniques in this field so that the reader who is not familiar with this field will be able to place this work in its proper medical context.

Chapter 7 of this thesis contains a detailed description of the techniques applied in the construction of the HEPAR system, and gives an overview of the medical knowledge included in the system. The method of knowledge acquisition and techniques applied in the development of the HEPAR system are also described.

In Chapter 8, two successive efforts in the validation of the diagnostic performance of the HEPAR system are described. These studies would not have been possible without the availability of a special testing environment, consisting of a collection of interrelated software tools, developed by the author, described in Chapter 7.

Chapter 9 summarizes the major conclusions of this thesis, and some suggestions for future research are proposed.

Part I

Structures in Diagnosis

Chapter 2

Formal Theories of Diagnosis

Medical diagnosis is an application field for which many diagnostic computer programs have been developed over a long period of time. Although these systems frequently dealt with similar, or related, problem domains, often their underlying principles differed considerably. In a sense, this was a consequence of the additional goal of the development of many of these systems to explore representation and reasoning methods. Only after researchers experienced that developing diagnostic systems was much more difficult than previously thought, was it recognized that the principles underlying diagnosis were actually poorly understood. Starting about halfway through the 1980s, a significant amount of research on conceptual and formal aspects of diagnosis was undertaken, with the aim of acquiring more insight into the nature of diagnostic problem solving. The major results of this research are reviewed in the present chapter. First, the conceptual basis of diagnosis, which brought researchers to propose their formalizations of diagnosis in the first place, is briefly reviewed. Next, the most influential formal theories of diagnosis are described. This chapter paves the way for the presentation of our framework of diagnosis, which is introduced in Chapter 3 and applied in chapters 4 and 5. Recall that we confine the treatment to static aspects of diagnosis.

2.1 Conceptual basis

Although the description of diagnostic problem solving given in Section 1.1 carries much of the flavour of the process of diagnosis, it is still an imprecise description and, in fact, several formal theories have been proposed to capture the concept of diagnosis more precisely. However, in doing so, researchers have become aware that there are actually various *conceptual models* of diagnosis, determined by the kind of knowledge involved. As stated in Chapter 1, diagnosis concerns the interpretation of observed findings in the context of knowledge from a problem domain. A good starting point for describing diagnosis at a conceptual level are the various sorts of knowledge that play a role in developing diagnostic applications.

The knowledge embodied in a diagnostic system may be based on:

(1) A description of the *normal* structure and functional behaviour of a system.

(2) A description of *abnormal* functional behaviour of a system; abnormal structure is usually not taken into account.

(3) An enumeration of defects and collections of observable findings for every possible defect concerned, without the availability of knowledge concerning the (abnormal) functional behaviour of the system.

(4) An enumeration of findings for the normal situation.

These types of knowledge may coexist in real-life diagnostic systems, but it is customary to emphasize their distinction in conceptual and formal theories of diagnosis. Similar classifications of types of knowledge appear in the literature on diagnosis, although often no clear distinction is made between the conceptual, formal and implementation aspects of diagnostic systems. For example, [Davis & Hamscher, 1988] and [Poole, 1988] distinguish diagnostic rule-based systems, by which they mean diagnostic systems based on knowledge of the third type mentioned above, from diagnostic systems incorporating knowledge of structure and behaviour. However, rule-based systems with a sufficiently expressive production-rule formalism can be used to implement any diagnostic system, including those based on knowledge of structure and behaviour.

An observed finding that has been gathered in diagnosing a problem is often said to be either a normal finding, i.e. a finding that matches the normal situation, or an abnormal finding, i.e. a finding that does not match the normal situation. Based on the four types of knowledge mentioned above, and the two sorts of findings, three different conceptual models of diagnosis are usually distinguished; they will be called:

- *Deviation-from-Normal-Structure-and-Behaviour diagnosis*, abbreviated to *DNSB diagnosis*,

- *Matching-Abnormal-Behaviour diagnosis*, abbreviated to *MAB diagnosis*, and

- *Abnormality-Classification diagnosis*, abbreviated to *AC diagnosis.*

A formal theory of diagnosis has been proposed for each of these conceptual models of diagnosis. In the remainder of this section, each of the three conceptual models of diagnosis will be introduced together with the formal theory of diagnosis proposed. The formal theories of diagnosis are discussed in depth in Section 2.2.

DNSB diagnosis. For diagnosis based on knowledge concerning normal structure and behaviour, no explicit knowledge is available about the relationships between defects of the system, on the one hand, and findings to be observed when certain defects are present, on the other hand. From a practical point of view, the primary motivation for investigating this approach to diagnosis is that in many domains little knowledge concerning abnormality is available, which is certainly true for new human-developed artifacts. For example, for a new device that has just been released from the factory, experience with respect to the faults that may occur when the device is in operation, is lacking. Thus, the only conceivable way in which initially such faults can be handled is by looking at the normal structure and functional behaviour of the device. Yet, even if knowledge concerning abnormal behaviour is available, exhaustive description may be too

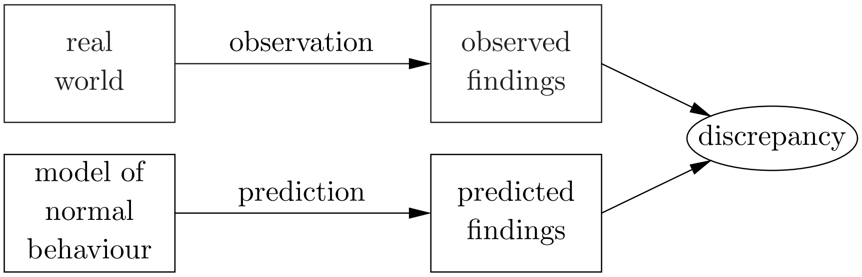


Figure 2.1: DNSB diagnosis.

cumbersome compared with a model of normal behaviour. For the purpose of diagnosis, the actual behaviour of a physical device, called *observed behaviour*, is compared with the results of a model of normal structure and behaviour of the device, which may be taken as *predicted behaviour*. Both types of behaviour can be characterized by findings. If there is a *discrepancy* between the observed and the predicted behaviour, diagnostic problem solving amounts to isolating the components in the device that are not properly functioning, using a model of the normal structure and behaviour of the device [Brown et al., 1982; Davis, 1984; Davis & Hamscher, 1988; Genesereth, 1984; De Kleer, 1977]. In doing so, it is assumed that the model of normal structure and behaviour is sufficiently accurate and correct. Figure 2.1 depicts DNSB diagnosis in a schematic way. DNSB diagnosis is frequently erroneously called model-based diagnosis in the literature, as if it were the only instance of model-based diagnosis. It is also called consistency-based dagnosis, but in this thesis this term is reserved for the corresponding formal theory of diagnosis. DNSB diagnosis has been developed in the context of troubleshooting in electronic circuits [Davis & Hamscher, 1988]. A well-known program that supports DNSB diagnosis, and includes various strategies to do so as efficiently as possible, is the *General Diagnostic Engine* (GDE) [De Kleer & Williams, 1987].

Above, we have reviewed the conceptual basis of diagnosis based on a model of normal structure and behaviour, which we have called DNSB diagnosis. The formal counterpart of DNSB diagnosis, called *consistency-based diagnosis*, originates from work by R. Reiter, [Reiter, 1987]; consistency-based diagnosis will be discussed in detail below. As far as known to the author, DNSB diagnosis-like approaches have been used in medical applications on a limited scale (cf. for example [Downing, 1993]); there is more work in which DNSB diagnosis has been applied to solve diagnostic non-medical, in particular technical, problems (cf. [Beschta et al., 1993; Dague, 1994; Hamscher, 1994; Ng, 1991; Sauthier & Faltings, 1992; Stefanini et al., 1993]).

MAB diagnosis. For diagnosis based on knowledge of abnormal behaviour, diagnostic problem solving amounts to simulating the abnormal behaviour using an explicit model of that behaviour. By assuming the presence of certain defects, some observable abnormal findings can be predicted. It can be investigated which of these assumed defects account for the observed findings by *matching* the predicted abnormal findings with those observed. In Figure 2.2, MAB diagnosis is depicted schematically. In most applications of MAB diagnosis, the domain knowledge that is used for diagnosis consists of cause-effect (causal) relationships. Two, strongly related, formal counterparts of MAB diagnosis have

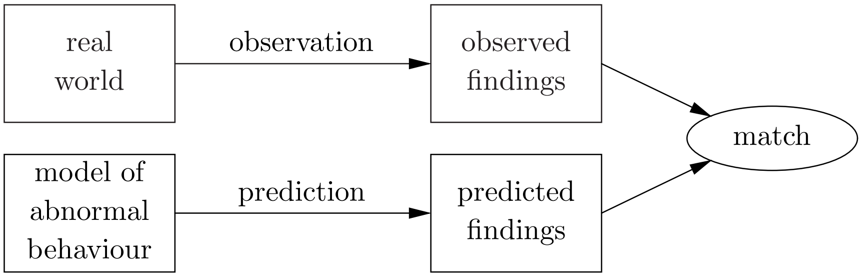


Figure 2.2: MAB diagnosis.

been proposed in the literature. The first formal theory, referred to as the *set-covering theory of diagnosis*, is based on set theory: causal knowledge is expressed as mathematical relations, used for diagnosis. This formal theory originates from work by J.A. Reggia and others [Reggia et al., 1983]. The second formal theory is based on logic. Early work in this area has been done by P.T. Cox and T. Pietrzykowski [Cox & Pietrzykowski, 1987], D. Poole [Poole et al., 1987], and by L. Console and P. Torasso [Console et al., 1989; Console & Torasso, 1990a]. Based on the type of reasoning employed to formalize MAB diagnosis, this theory of diagnosis is also referred to as *abductive diagnosis*.

AC diagnosis. Where DNSB and MAB diagnosis employ a model of normal or abnormal structure and behaviour for the purpose of diagnosis, the third conceptual model of diagnosis uses neither. The knowledge employed in this conceptual model of diagnosis consists of the enumeration of useful evidence that can be observed, i.e. observed findings, when a particular defect or defect category is present. This form of knowledge has been referred to as *empirical associations* in Chapter 1 (the phrase *compiled knowledge* is also employed) [Buchanan & Shortliffe, 1984]. Diagnostic problem solving amounts to establishing which of the elements in a finite set of defects have associated findings that account for as many of the findings observed as possible, as is shown in Figure 2.3. The enumeration of findings for the normal situation is sometimes also used in AC diagnosis; then, observed findings are classified in terms of present and absent defects. The main goal of AC diagnosis, however, remains the classification of observed findings in terms of abnormality. AC diagnosis is often referred to in the literature as *heuristic classification* [Clancey, 1985], although this term is ampler, since it also includes a reasoning strategy. The MYCIN system, [Shortliffe, 1976], is the classical system in which this conceptual approach to diagnosis has been adopted. AC diagnosis can be characterized in terms of logical deduction in a straightforward way. We shall refer to this formalization of AC

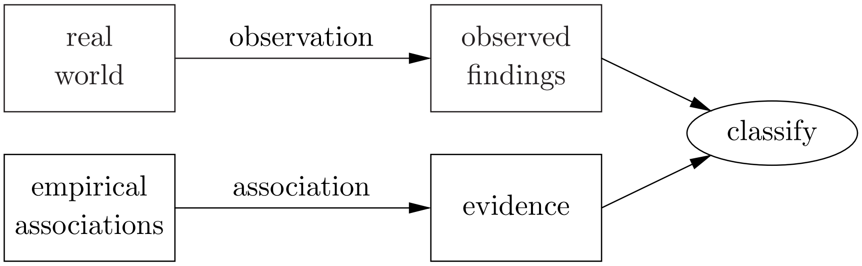


Figure 2.3: AC diagnosis.

diagnosis as *hypothetico-deductive diagnosis*.

Obviously, the various models of diagnosis discussed above can also be combined. To solve real-life diagnostic problems in a domain, it is likely that a mixture of conceptual models of diagnosis as distinguished above, will be required. The result is known as diagnosis with *multiple models* [Struss, 1992]. Several programs have been developed that offer limited possibility to carry out diagnostic problem solving using multiple models; examples of such programs are GDE$^+$ [Struss & Dressler, 1989] and Sherlock [De Kleer & Williams, 1989]. These programs use DNSB diagnosis as their core approach.

Although in the literature it is emphasized that the conceptual models of diagnosis discussed above focus on different forms of diagnosis, they have much in common. For example, the type of knowledge used in DNSB diagnosis can be viewed as an implicit, or intensional, version of the type of knowledge used in AC diagnosis (if restricted to normality classification), which is an explicit or extensional type of knowledge; the associations between normal observable findings and the absence of defects are hidden in the specified normal behaviour in DNSB diagnosis. DNSB and MAB diagnostic problem solving are based on some kind of simulation of behaviour; such simulation of behaviour is absent in AC diagnosis.

2.2 Formal theories

There have been several attempts to formalize the various conceptual models of diagnosis discussed above; most, but not all, of these formalizations are based on logic. The most important formal theories will be briefly reviewed.

2.2.1 Consistency-based diagnosis

The formal theory of diagnosis originally proposed by R. Reiter [Reiter, 1987], was motivated by the desire to provide a formal underpinning of diagnostic problem solving using knowledge of the normal structure and behaviour of technical devices, i.e. DNSB diagnosis. The theory of diagnosis may be viewed as the logical foundation of earlier work in DNSB diagnosis by J. de Kleer [De Kleer, 1977; De Kleer & Williams, 1987], Brown and colleagues [Brown et al., 1982], R. Davis and H. Shrobe [Davis & Shrobe, 1983; Davis, 1984], and M.R. Genesereth [Genesereth, 1984]. The logical formalization uses results from earlier work by R. Reiter, [Reiter, 1980], and J. McCarthy, [McCarthy, 1986], on nonmonotonic reasoning. We shall sometimes refer to this theory of diagnosis as Reiter's formal theory of diagnosis.

Reiter's theory of diagnosis was later extended by De Kleer et al. [De Kleer et al., 1992]; in this section, both formalizations will be introduced in a single, logical framework. Where appropriate, the differences between Reiter's original proposal, [Reiter, 1987], and the extensions proposed by De Kleer et al., [De Kleer et al., 1992], will be indicated. This formal theory of diagnosis is often referred to as the *consistency-based theory of diagnosis*, or *consistency-based diagnosis* for short. The theory is briefly described below.

The logical specification of knowledge concerning structure and behaviour in Reiter's theory is a triple $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS}, \mathrm{OBS})$, called a *system*, where

- SD denotes a finite set of formulae in first-order predicate logic, specifying normal structure and behaviour, called the *system description*;
- COMPS denotes a finite set of constants (nullary function symbols) in first-order logic, denoting the *components* (elements) of the system;
- OBS denotes a finite set of formulae in first-order predicate logic, denoting *observations*, i.e. observed findings.

It is, in principle, possible to specify normal as well as abnormal (faulty) behaviour within a system description SD, but originally SD was designed to comprise a logical specification of normal behaviour of the modelled system only, thus yielding the intended formalization of DNSB diagnosis. Slightly simplified, the essential part of a formal model of normal structure and behaviour of a system consists of logical axioms of the form

$$\neg\text{Abnormal}(c) \rightarrow f_{norm}$$

for each component $c \in$ COMPS and some finding f_{norm} that may be observed if the component c is normal, i.e. is nondefective. The axioms will be referred to as *normality axioms*. It is assumed that the finding f_{norm} may be observed in reality when component c of the device, that has been modelled in logic, is operating normally. Such an observed finding is called a *normality observation*. The subscript *norm* is used to emphasize that a particular finding represents a normal result; in Section 2.2.2 and further, the subscript *ab* is used to indicate an abnormal finding. These subscripts are only used for clarity and have no additional meaning; they will often be omitted. The predicate symbol 'Abnormal' is sometimes referred to as the *fault mode* (also behavioural mode) of the component [De Kleer & Williams, 1989]. The literal 'Abnormal(c)' denotes the component c to be defective if satisfied. Other predicate names, such as 'OK', 'Correct', are also employed in the literature, with similar intended meaning and use as the negation of an 'Abnormal' literal.

Diagnostic problem solving is formalized as a method for finding the source of unsatisfiability in the logical description of the (normal) functioning of a system when supplied with observed findings, where some of the observed findings are the result of a system defect in reality. Hence, unsatisfiability formalizes the notion of discrepancy in DNSB diagnosis as indicated in Figure 2.1. If it is assumed that the atom Abnormal(c) is *false*, i.e. the component c is functioning normally, an unsatisfiability will arise given the observed finding $\neg f_{norm}$. This result is interpreted in Reiter's theory as an indication that the defect may be localized in component c. This gives rise to the hypothesis that component c is defective, i.e. Abnormal(c) is *true*, and the unsatisfiability is resolved if the assumption that Abnormal(c) is *false* was its only source.

Adopting the definition from [De Kleer et al., 1992], a diagnosis in the theory of consistency-based diagnosis can be defined as follows.

Definition 2.1 (*consistency-based diagnosis*)**.** *Let* $\mathcal{S} = (\text{SD}, \text{COMPS}, \text{OBS})$ *be a system. Let*

$$C_P = \{\text{Abnormal}(c) \mid c \in \text{COMPS}\}$$

be the set of all positive 'Abnormal' literals, and

$$C_N = \{\neg\text{Abnormal}(c) \mid c \in \text{COMPS}\}$$

be the set of all negative 'Abnormal' literals. Furthermore, let $C \subseteq C_P \cup C_N$ *be a set, such that*

$$C = \{\text{Abnormal}(c) \mid c \in D\} \cup \{\neg\text{Abnormal}(c) \mid \text{COMPS} \backslash D\}$$

for some $D \subseteq \text{COMPS}$*. Then,* C *is a* (consistency-based) diagnosis *of* $\mathcal{S}$ *if the following condition, called the* consistency condition*, holds:*

$$\text{SD} \cup C \cup \text{OBS} \nvDash \bot \tag{2.1}$$

i.e. $\text{SD} \cup C \cup \text{OBS}$ *is satisfiable.*

Here, $\nvDash$ stands for the negation of the logical entailment relation, and $\bot$ represents 'falsum'. The consistency condition (2.1) captures DNSB diagnosis in terms of consistency-based diagnosis under the assumption that the axioms in SD provide a completely accurate and correct representation of a physical system. In the formalization by De Kleer et al., [De Kleer et al., 1992], each literal $\text{Abnormal}(c) \in C$ is interpreted as being defective; a literals $\neg\text{Abnormal}(c) \in C$ indicates component c to be nondefective. In the original theory by Reiter, [Reiter, 1987], the set D above is taken as a diagnosis, with the extra requirement that D is minimal with respect to set inclusion. Then, each component c in a diagnosis D for which $\text{Abnormal}(c)$ is *true* is interpreted as being defective. According to expression (2.1), taking $D = \text{COMPS}$ leads to the trivial diagnosis that all components are defective (or the defective components are among the set of all components). Reiter, therefore, incorporated in the original theory the requirement that the set D must be a minimal set with respect to set inclusion, fulfilling the consistency condition. However, later it was recognized that minimality according to set inclusion is merely a measure of plausibility, which may not be appropriate when knowledge of abnormal behaviour is also included in the system description SD, and the minimality criterion was left out of the basic definition. Moreover, other measures of plausibility (cf. [Tuhrim et al., 1991] in the context of abduction) may also apply. The application of the formal theory by Reiter is illustrated by a classical example from the literature [Genesereth, 1984].

Example 2.1. Consider the logical circuit depicted in Figure 2.4, which represents a full adder, i.e. a circuit that can be used for the addition of two bits with carry-in and carry-out bits. The components X_1 and X_2 represent exclusive-OR gates, A_1 and A_2 represent AND gates, and R_1 represents an OR gate.

The system description, as provided in [Reiter, 1987], consists of the following axioms:

$$\begin{aligned}
\forall x(\text{ANDG}(x) \wedge \neg\text{Abnormal}(x) &\rightarrow \mathit{out}(x) = \mathit{and}(\mathit{in1}(x), \mathit{in2}(x))) \\
\forall x(\text{XORG}(x) \wedge \neg\text{Abnormal}(x) &\rightarrow \mathit{out}(x) = \mathit{xor}(\mathit{in1}(x), \mathit{in2}(x))) \\
\forall x(\text{ORG}(x) \wedge \neg\text{Abnormal}(x) &\rightarrow \mathit{out}(x) = \mathit{or}(\mathit{in1}(x), \mathit{in2}(x)))
\end{aligned}$$

which describe the (normal) behaviour of each individual component (gate), and

$$\mathit{out}(X_1) = \mathit{in2}(A_2)$$

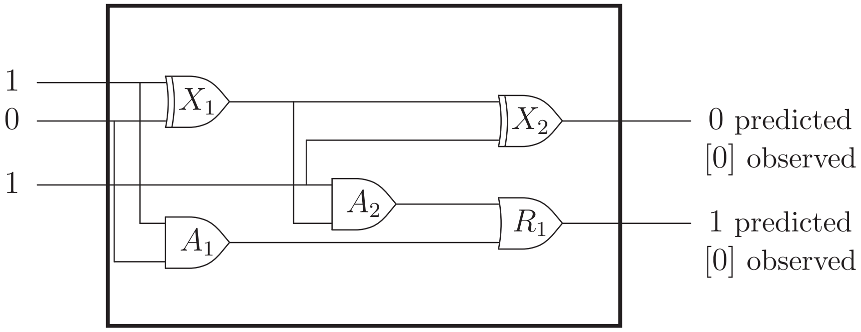


Figure 2.4: Full adder.

$$\begin{aligned}
out(X_1) &= in1(X_2)\\
out(A_2) &= in1(R_1)\\
in1(A_2) &= in2(X_2)\\
in1(X_1) &= in1(A_1)\\
in2(X_1) &= in2(A_1)\\
out(A_1) &= in2(R_1)
\end{aligned}$$

which gives information about the connections between the components, i.e. information about the normal structure, including some electrical relationships. Finally, the various gates are defined:

$$\begin{array}{l}
\text{ANDG}(A_1)\\
\text{ANDG}(A_2)\\
\text{XORG}(X_1)\\
\text{XORG}(X_2)\\
\text{ORG}(R_1)
\end{array}$$

Appropriate axioms for a Boolean algebra are also assumed to be available.

Now, let us assume that

$$\text{OBS} = \{in1(X_1) = 1, in2(X_1) = 0, in1(A_2) = 1, out(X_2) = 0, out(R_1) = 0\}$$

Note that $out(R_1) = 1$ is predicted using the model of normal structure and behaviour in Figure 2.4, which is in contrast with the observed output $out(R_1) = 0$. Assuming that $C = \{\neg\text{Abnormal}(c) \mid c \in \text{COMPS}\}$, it follows that

$$\text{SD} \cup C \cup \text{OBS}$$

is unsatisfiable. This confirms that some of the output signals observed differ from those expected under the assumption that the circuit is functioning normally. Using Formula (2.1), a possible diagnosis is, for instance,

$$\begin{aligned}
C' = \{&\text{Abnormal}(X_1), \neg\text{Abnormal}(X_2), \neg\text{Abnormal}(A_1),\\
&\neg\text{Abnormal}(A_2), \neg\text{Abnormal}(R_1)\}
\end{aligned}$$

since

$$\mathrm{SD} \cup C' \cup \mathrm{OBS}$$

is satisfiable. Note that, given the diagnosis C', no output is predicted for the circuit; the assumption Abnormal(X_1) completely blocks transforming input into output by the modelled circuit, because

$$\mathrm{SD} \cup C' \cup \mathrm{OBS} \backslash \{ out(X_2) = 0 \} \nvDash out(X_2) = 0$$

In a sense, this is too much, because there was no discrepancy between the predicted and observed output of gate X_2. Nevertheless, C' is a diagnosis according to Definition 2.1. $\Diamond$

It is interesting to look at consistency-based diagnosis in a more intuitive way. What the theory actually expresses is that if components that may be defective are removed from a system or device, and the resulting newly predicted behaviour, or no behaviour at all, does not contradict the observed behaviour, then a diagnosis has been established. This is a rather crude approach to diagnosis. Imagine that we have a formal model of an electrical device, including its electric plug, then simulating the removal of the plug from its socket, thus recovering satisfiability, will provide us with a diagnosis for a defective system. According to this approach, the plug will be identified as the culprit, which, of course, is absurd if the device was in operation prior to the removal of the plug, although incorrectly. K. Konolige, [Konolige, 1992; Konolige, 1994], refers to diagnoses produced by consistency-based diagnosis as *excuses*, to reflect that it may not be possible to explain such diagnoses in terms of cause-effect relationships.

In addition to a definition of consistency-based diagnosis, [De Kleer et al., 1992] introduces the concepts of partial diagnosis and kernel diagnosis. A *partial diagnosis* is an abbreviated representation for a set of diagnoses that have certain ‘Abnormal’ and ‘¬Abnormal’ literals in common. Partial diagnosis will be discussed in Chapter 4. A *kernel diagnosis* is simply a partial diagnosis that is minimal with respect to set inclusion.

In Section 2.2.5, the application of Reiter’s theory to the logical formalization of MAB diagnosis will be discussed. The techniques proposed by Reiter are not the only possible ways to formalize DNSB and MAB diagnosis; D. Poole has proposed other logical techniques for the same purpose in terms of his Theorist framework of default reasoning [Poole et al., 1987; Poole, 1990a; Poole, 1990b; Poole, 1994]. This work, however, bears great resemblance to the work by Reiter with respect to DNSB diagnosis, and to the work by Console and Torasso with respect MAB diagnosis, which will be discussed in the following section.

2.2.2 Abductive diagnosis

The formalization of MAB diagnosis has been thoroughly studied by L. Console and P. Torasso [Console et al., 1989; Console & Torasso, 1990a; Console & Torasso, 1990b]. In their theory, the abnormal behaviour of a system is specified in terms of abnormal states and resulting abnormal findings. Normal findings may also be included, but these are less useful for diagnosis, since an abnormal state is often causally related to a large number of normal findings. Diagnostic problem solving is formally described as the problem of accounting for a given set of observed findings, referred to in the theory as manifestations, by the simulation of abnormal behaviour. The simulation process is accomplished by deduction with logical axioms, describing abnormal behaviour, and assumed (abnormal) states.

The logical axioms in the formal theory by Console and Torasso are (definite) Horn clauses of the following form and meaning

$$\begin{aligned} \text{State}_1 \wedge \cdots \wedge \text{State}_n &\rightarrow f && (2.2)\\ \text{State}_1 \wedge \cdots \wedge \text{State}_n &\rightarrow \text{State} && (2.3)\\ \text{State}_1 \wedge \cdots \wedge \text{State}_n &\rightarrow d && (2.4) \end{aligned}$$

where State and State_i, $i = 1, \ldots, n$, are positive literals representing part of the internal state of a modelled system, d is a *defect* (or disorder), and f is a *finding*. It is assumed that the set of Horn clauses is hierarchical, i.e. no cyclic dependencies among atoms in clauses are allowed (which contrasts with the situation in logic programming where cyclic dependencies are almost the rule). In the original abductive theory of diagnosis by Console and Torasso, as described in [Console et al., 1989], a finding appearing in the conclusion of a logical implication usually represents an *abnormal* finding. Recall that such findings are sometimes denoted by f_{ab}, and that normal findings are sometimes denoted by f_{norm}. A state literal is employed for the simulation of the occurrence of abnormal behaviour using the logical specification. It corresponds to a parameter with a value. For example, if the parameter $pressure(blood)$ can take values *decreased*, *normal* and *increased*, then $pressure(blood) = increased$ corresponds to a state. The intuitive meaning of formulae of the form (2.2) is: 'presence of $\text{State}_1, \ldots, \text{State}_n$ *causes* the abnormal finding f', i.e. if $\text{State}_1, \ldots, \text{State}_n$ hold in the system, abnormal finding f must be observed. Formulae of the form (2.3) express that a collection of states is causally related to another state, i.e. if the states $\text{State}_1, \ldots, \text{State}_n$ occur then State occurs as well. The two logical axioms above are sometimes referred to as *abnormality axioms*. Note that the notion of causality is expressed in the theory using logical implication. Logical implication is employed to express a causal relationship between states and observable findings, and between states and states. Axioms of the form (2.4) can be viewed as *classification axioms* because they classify a collection of states as a particular defect. If sufficient state literals are assumed, a defect d can be derived, using axiom (2.4). In the theory by Console and Torasso, a defect is actually defined in terms of a collection of states. This can be expressed by using a bi-implication ($\leftrightarrow$) instead of an implication, as in axiom schema (2.4). However, when adopting this formalization for diagnosis, the implications from right to left ($\leftarrow$) are not involved. Classification axioms are not an essential ingredient of the theory of diagnosis by Console and Torasso; they are merely used to attach diagnostic labels to collections of states. Note that in the classification axioms, logical implication is used to express a classification instead of a causal relationship, as in the abnormality axioms. Due to the manifold uses of logical implication, the theory by Console and Torasso provides no clear logical meaning for the various relationships, including causality, underlying their theory of diagnosis. To express the theory by Console and Torasso in terms of defects and findings only, thus enabling us to analyse the essentials of the theory, states are identified with defects. Thus, axioms of the form (2.3) and (2.4) are collapsed into one axiom schema;

the classification axioms are given no further consideration. In the following, it shall be assumed that we have axioms of the following two forms:

$$d_1 \wedge \cdots \wedge d_n \rightarrow f \tag{2.5}$$

$$d_1 \wedge \cdots \wedge d_n \rightarrow d \tag{2.6}$$

where d, d_i, $i = 1, \ldots, n$, represent defects.

Console and Torasso also provide a mechanism in their logical formalization to weaken the causality relation. To this end, literals α are introduced into the premises of the axioms of the form (2.5) and (2.6), which can be used to block the deduction of a finding f or defect d if the defects d_i, $i = 1, \ldots, n$, hold true, by assuming the literal α to be false. The weakened axioms have the following form:

$$d_1 \wedge \cdots \wedge d_n \wedge \alpha_f \rightarrow f \tag{2.7}$$

$$d_1 \wedge \cdots \wedge d_n \wedge \alpha_d \rightarrow d \tag{2.8}$$

The literals α are called *incompleteness-assumption literals*, abbreviated to *assumption literals*. Axioms of the form (2.5) – (2.8) are now taken as the abnormality axioms.

In the sequel, let $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ stand for a *causal specification* in the theory of diagnosis by Console and Torasso, where:

- Δ denotes a set possible defect and assumption literals;
- Φ denotes a set of possible (positive and negative) finding literals;
- $\mathcal{R}$ stands for a set of logical (abnormality) axioms of the form (2.5) – (2.8).

A causal specification can then be employed for the prediction of observable findings in the sense of Figure 2.2.

Definition 2.2 (*prediction*). *Let $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ be a causal specification. Then, a set $H \subseteq \Delta$ is called a* prediction *for a set of observable findings $F \subseteq \Phi$ if*

(1) *$\mathcal{R} \cup H \vDash F$, and*

(2) *$\mathcal{R} \cup H$ is satisfiable.*

Hence, the notion of prediction formalizes the arrow in the lower half of Figure 2.2; the resulting set of findings F corresponds to the predicted (observable) findings in the same figure.

An *abductive diagnostic problem* $\mathcal{A}$ is now defined as a pair $\mathcal{A} = (\mathcal{C}, E)$, where $E \subseteq \Phi$ is called a *set of observed findings*. A set of observed findings corresponds to the box in the upper half of Figure 2.2.

Formally, a solution to an abductive diagnostic problem $\mathcal{A}$ can be defined as follows.

Definition 2.3 (*solution*). *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem, where $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ is a causal specification with $\mathcal{R}$ a set of abnormality axioms of the form (2.5) – (2.8), and $E \subseteq \Phi$ a set of observed findings. A set of defect and assumption literals $H \subseteq \Delta$ is called a* solution *to $\mathcal{A}$ if:*

(1) $\forall f \in E : \mathcal{R} \cup H \vDash f$ (covering condition);

(2) $\forall f \in E^c : \mathcal{R} \cup H \nvDash \neg f$ (consistency condition)

where E^c is defined by:

$$E^c = \{\neg f \in \Phi \mid f \in \Phi, f \notin E,\ f \textit{ is a positive literal}\}$$

In the work of Console and Torasso, the set $\mathcal{R} \cup H$ is called a 'world' if H is a prediction; the set $\mathcal{R} \cup H$ is called a 'final world' if H is a solution to an abductive diagnostic problem [Console et al., 1989; Console & Torasso, 1990a]. Note that the sets E and E^c are disjoint, and that if $f \in E$ then $\neg f \notin E^c$. The set of observed findings E is denoted by Ψ in [Console et al., 1989] and [Console & Torasso, 1990a], and denoted by Ψ^+ in [Console & Torasso, 1990b] and [Console & Torasso, 1991]. The set E^c (denoted by $\bar{\Psi}$ in [Console et al., 1989] and [Console & Torasso, 1990a], and denoted by Ψ^- in [Console et al., 1989] and [Console & Torasso, 1990a]) stands for findings assumed to be false, because they have not been observed (and are therefore assumed to be absent). But any finding may also be unknown. Thus, rather than providing a single definition, Console and Torasso provide in their articles several alternatives for this set E^c. The definition provided in Definition 2.3 above is just one of the alternatives.

Condition (1) is called the covering condition, because it requires that each observed finding is accounted for by a solution H. Note that any solution to a diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$ is a prediction for E according to Definition 2.2. Condition (2) is called the consistency condition, because it can be restated as follows

$$\mathcal{R} \cup H \cup E^c \nvDash \perp$$

A set of defects in a prediction H is also called a set of *perturbations* [Console & Torasso, 1990a]; in [Console & Torasso, 1991] the term *abducibles* is employed for literals that may be assumed as part of diagnostic problem solving.

In the original formulation of the theory only those defects (states) are admitted to H which do not appear in the conclusions of implications; such defects are called *initial defects* (initial states in the original theory). The covering condition defined above ensures that sufficient defects and assumption literals are assumed to account for all given observed findings. The consistency condition helps to ensure that not too many defect and assumption literals are assumed. Although it is only necessary to include an assumption literal α in a solution for implications $d \wedge \alpha_f \rightarrow f$ and $d \wedge \alpha_{d'} \rightarrow d'$ if the defect d is deducible from the assumed (initial) defects and assumption literals, Definition 2.3 does not always prevent their inclusion in a solution.

An entire solution H may be taken as a diagnosis, but following [Console et al., 1989], a diagnosis is considered to consist of the defect literals in a solution H.

Definition 2.4 (*abductive diagnosis*). *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem, where $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ is a causal specification. Let H be a solution to $\mathcal{A}$. Then, the set of all defects $D \subseteq H$ is called an* (abductive) diagnosis *of $\mathcal{A}$.*

Recall that in [Console et al., 1989], a diagnosis is obtained by applying the classification axioms (2.4); a distinction is therefore made in [Console et al., 1989] between a solution

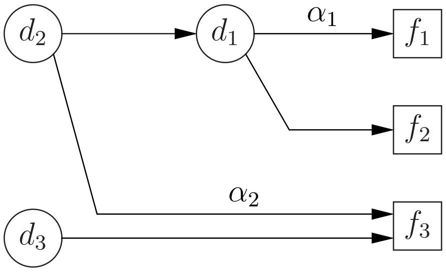


Figure 2.5: A knowledge base with causal relations.

H for which the covering and consistency conditions are satisfied, i.e. the set of defect and assumption literals contained in a 'final world' – this world is called a *causal explanation* – and the set of defects resulting from an explanation, which is called a diagnosis (originally, a solution). However, from a formal point of view, the distinction is not essential.

Example 2.2. Consider the causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, with $\Delta = \{d_1, d_2, d_3\}$, $\Phi = \{f_1, f_2, f_3, \neg f_1, \neg f_2, \neg f_3\}$, and the following set of logical formulae $\mathcal{R}$, representing medical knowledge concerning influenza and sport, both 'disorders' with frequent occurrence:

$$
\begin{array}{l}
d_1 \wedge \alpha_1 \rightarrow f_1 \\
d_2 \rightarrow d_1 \\
d_1 \rightarrow f_2 \\
d_2 \wedge \alpha_2 \rightarrow f_3 \\
d_3 \rightarrow f_3
\end{array}
$$

where the following medical meaning is given to the various symbols:

d_1:	fever
d_2:	influenza
d_3:	sport
f_1:	chills
f_2:	thirst
f_3:	myalgia (i.e. painful muscles)

For example, $d_2 \wedge \alpha_2 \rightarrow f_3$ means that influenza may cause myalgia; $d_2 \rightarrow d_1$ means that influenza always causes fever. For illustrative purposes, a causal knowledge base as given above is often depicted as a labelled, directed graph G, which is called a *causal net*, as shown in Figure 2.5. Suppose that the abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$ must be solved, where the set of observed findings is equal to: $E = \{f_2, f_3\}$. Then, $E^c = \{\neg f_1\}$. There are several solutions to this abductive diagnostic problem (for which the consistency and covering conditions are fulfilled):

$$
\begin{array}{l}
H_1 = \{d_2, \alpha_2\} \\
H_2 = \{d_2, d_3\} \\
H_3 = \{d_1, d_3\} \\
H_4 = \{d_1, \alpha_2, d_2\}
\end{array}
$$

$$H_5 = \{d_2, \alpha_2, d_3\}$$
$$H_6 = \{d_1, d_2, d_3\}$$
$$H_7 = \{d_1, \alpha_2, d_2, d_3\}$$

The following diagnoses correspond to these solutions:

$$D_1 = \{d_2\}$$
$$D_2 = \{d_2, d_3\}$$
$$D_3 = \{d_1, d_2\}$$
$$D_4 = \{d_1, d_3\}$$
$$D_5 = \{d_1, d_2, d_3\}$$

For example, the diagnosis $D = \{d_1, d_2\}$ means that the patient has influenza with associated fever. Restricting to initial defects would yield the solutions H_1, H_2 and H_5 and the diagnoses D_1 and D_2. Finally, note that, for example, the prediction $H = \{\alpha_1, \alpha_2, d_1, d_2\}$ is incompatible with the consistency condition. ♢

Because in this theory of diagnosis, the observable findings are logically entailed by the assumption of the presence of certain states, and the reasoning goes in a sense in a direction reverse to that of the logical implication, i.e. from the consequent to the premise, the theory is often referred to as the *abductive theory of diagnosis*, or *abductive diagnosis* for short.

Several researchers (cf. [Poole, 1988; Console et al., 1991]) have noted a close correspondence between abduction and the predicate completion of a logical theory, as originally proposed by K. Clark in connection with negation as finite failure in logic programming [Clark, 1978]. The characterization of abduction as deduction in a completed logical theory is natural, because computation of the predicate completion of a logical theory amounts to adding the only-if parts of the formulae to the theory, i.e. it 'reverses the arrow' which is exactly what happens when abduction is applied to derive conclusions. In an intuitive sense, predicate completion expresses that the only possible causes (defects) for observed findings are those appearing in the abnormality axioms. Where the characterization of abduction by means of the covering and consistency conditions may be viewed as a meta-level description of abductive diagnosis, the predicate completion can be taken as the object-level characterization, i.e. in terms of the original axioms in $\mathcal{R}$. [Poole, 1988] and [Console et al., 1991] note that, in contrast to the predicate completion in logic programming, predicate completion should only pertain to literals appearing as a consequence of the logical axioms in $\mathcal{R}$, i.e. finding literals and defect literals that can be derived from other defects and assumption literals. This set of defects and observable findings is called the set of *non-abducible* literals, denoted by A; the set $\Delta \backslash A$ is then called the set of *abducible* literals.

Let us denote the axiom set $\mathcal{R}$ by

$$\begin{aligned}\mathcal{R} = \{&\varphi_{1,1} \to a_1, \ldots, \varphi_{1,n_1} \to a_1,\\ &\qquad\vdots\\ &\varphi_{m,1} \to a_m, \ldots, \varphi_{m,n_m} \to a_m\}\end{aligned}$$

where $A = \{a_i \mid 1 \leq i \leq m\}$ is the set of non-abducible (finding or defect) literals and each $\varphi_{i,j}$ denotes a conjunction of defect literals, possibly including an assumption literal.

The predicate completion of $\mathcal{R}$ with respect to the non-abducible literals A, denoted by $\text{COMP}[\mathcal{R}; A]$ (cf. [Genesereth & Nilsson, 1987]), is defined as follows:

$$\begin{aligned}\text{COMP}[\mathcal{R}; A] = \mathcal{R} \cup \{ & a_1 \rightarrow \varphi_{1,1} \vee \cdots \vee \varphi_{1,n_1}, \\ & \vdots \\ & a_m \rightarrow \varphi_{m,1} \vee \cdots \vee \varphi_{m,n_m} \}\end{aligned}$$

The predicate completion of $\mathcal{R}$ makes explicit the fact that the only causes of non-abducible literals (findings and possibly also defects) are the defects and assumption literals given as a disjunct in the consequent. For example,

$$f_{ab} \rightarrow d_1 \vee \cdots \vee d_n$$

indicates that only the defects from the set $\{d_1, \ldots, d_n\}$ can be used to explain the observed finding f_{ab}.

Predicate completion of abnormality axioms with respect to a set of non-abducible literals can now be used to characterize diagnosis. Let ψ and ψ' be two logical formulae. It is said that ψ is *more specific than* ψ' iff $\psi \vDash \psi'$. Using the predicate completion of a set of abnormality axioms $\mathcal{R}$, we now have the following definition.

Definition 2.5 (*solution formula*). *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem and let* $\text{COMP}[\mathcal{R}; A]$ *be the predicate completion of $\mathcal{R}$ with respect to A, the set of non-abducible literals in $\mathcal{A}$. A* solution formula *S for $\mathcal{A}$ is defined as a most specific formula consisting only of abducible literals, such that*

$$\text{COMP}[\mathcal{R}; A] \cup E \cup E^c \vDash S$$

where E^c is defined as in Definition 2.3.

Hence, abductive diagnosis is transformed to hypothetico-deductive diagnosis (cf. Section 2.2.4). A solution formula is obtained by applying the set of equivalences in $\text{COMP}[\mathcal{R}; A]$ to a set of observed findings E, augmented with those findings not observed E^c, yielding a logical formula that includes all possible solutions according Definition 2.3, given the equivalences in $\text{COMP}[\mathcal{R}; A]$. The following theorem, which is proven in [Console et al., 1991], reveals an important relationship between the meta-level characterization of abductive diagnosis, as presented in Definition 2.3, and the object-level characterization of diagnosis in Definition 2.5.[1]

Theorem 2.1 (*[Console et al., 1991]*). *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem, where $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ is a causal specification. Let E^c be defined as in Definition 2.3, and let S be a solution formula for $\mathcal{A}$. Let $H \subseteq \Delta$ be a set of abducible literals, and let I be an interpretation of $\mathcal{A}$, such that for each abducible literal $a \in \Delta$: $\vDash_I a$ iff $a \in H$. Then, H is a solution to $\mathcal{A}$ iff $\vDash_I S$.*

Proof (sketch). ($\Rightarrow$): The set of defect and assumption literals H is a solution to $\mathcal{A}$, hence,

[1] Contrary to our treatment, in [Console et al., 1991], a solution H of an abductive problem $\mathcal{P}$ is defined by SLD resolution with the negation as finite failure rule, i.e. SLDNF resolution, such that $\mathcal{R} \cup H \vdash_{\text{SLDNF}} E \cup E^c$, i.e. the covering and consistency conditions are merged.

for each $f \in E$: $\mathcal{R} \cup H \vDash f$, and for each $f' \in E^c$: $\mathcal{R} \cup H \nvDash \neg f'$. The solution formula S is the result of rewriting observed findings in E and non-observed findings in E^c using the equivalences in $\mathrm{COMP}[\mathcal{R}; A]$ to a formula merely consisting of abducibles. Assume that S is in conjunctive normal form. A number of conjuncts in S is equivalent to an observed finding $f \in E$, that is logically entailed by $\mathcal{R} \cup H$, or to a non-observed finding $\neg f \in E^c$ that is consistent with $\mathcal{R} \cup H$. Hence, an interpretation I for which $\vDash_I H$, that falsifies each abducible in $\Delta \backslash H$, satisfying every $f \in E$ and each $\neg f \in E^c$ that has been rewritten, must satisfy this collection of conjuncts, i.e. S.
($\Leftarrow$): If S is in conjunctive normal form, S must be the result of rewriting observed findings $f \in E$ and non-observed findings in E^c to (negative or positive) abducibles, using the equivalences in $\mathrm{COMP}[\mathcal{R}; A]$. Since an interpretation I that satisfies H and S must also satisfy each finding $f \in E$ and those $\neg f \in E^c$ that have been rewritten to S, it follows that I can be chosen such that $\vDash_I E^c$, i.e. H must be a solution to $\mathcal{A}$. ◊

This theorem reveals an important property of the abductive theory of diagnosis. Sometimes, a solution to an abductive diagnostic problem is capable of satisfying a solution formula in the technical, logical sense.

Example 2.3. Reconsider the set of logical axioms given in Example 2.2. The predicate completion of $\mathcal{R}$ is equal to

$$\begin{aligned}\mathrm{COMP}[\mathcal{R}; \{f_1, f_2, f_3, d_1\}] &= \mathcal{R} \cup \{f_1 \rightarrow d_1 \wedge \alpha_1,\\ &\qquad\qquad d_1 \rightarrow d_2,\\ &\qquad\qquad f_2 \rightarrow d_1,\\ &\qquad\qquad f_3 \rightarrow (d_2 \wedge \alpha_2) \vee d_3\}\\ &= \{f_1 \leftrightarrow d_1 \wedge \alpha_1,\\ &\qquad d_1 \leftrightarrow d_2,\\ &\qquad f_2 \leftrightarrow d_1,\\ &\qquad f_3 \leftrightarrow (d_2 \wedge \alpha_2) \vee d_3\}\end{aligned}$$

Note that

$$\mathrm{COMP}[\mathcal{R}; \{f_1, f_2, f_3, d_1\}] \cup E \cup E^c \vDash (d_2 \wedge \alpha_2) \vee (d_2 \wedge d_3)$$

given that $E = \{f_2, f_3\}$ and $E^c = \{\neg f_1\}$. Although

$$\mathrm{COMP}[\mathcal{R}; \{f_1, f_2, f_3, d_1\}] \cup E \cup E^c \vDash \neg(d_1 \wedge \alpha_1)$$

the formula $\neg(d_1 \wedge \alpha_1)$, which is a logical consequence of $\neg f_1$ and $f_1 \leftrightarrow (d_1 \wedge \alpha_1)$, is not part of the solution formula $S \equiv (d_2 \wedge \alpha_2) \vee (d_2 \wedge d_3)$, because d_1 is non-abducible. It holds, in accordance with Theorem 2.1, that

$$\vDash_I H_i \;\; \Rightarrow \vDash_I (d_2 \wedge \alpha_2) \vee (d_2 \wedge d_3)$$

for $i = 1, 2, 5$, where H_i is a solution given in Example 2.2 consisting only of abducible literals, for suitable interpretations I. Here, it even holds that $H_i \vDash S$, because S does not contain any negative defects or assumption literals entailed by non-observed findings in E^c. ◊

Although the theory by Console and Torasso is restricted to reasoning with causal domain knowledge, other types of knowledge, referred to as *contextual information* by Console and Torasso, is also dealt with in the theory. Contextual information is incorporated to render the causal relation conditional on certain findings, e.g. in

$$d \wedge f \rightarrow f'$$

the finding literal f acts as a condition with regard to the causal relation between the defect d and the finding f'. For example, in a medical setting, many causal relations are age-specific; hence, the observed (normal) finding '$age \circ v$', where $\circ$ denotes an ordering predicate and v an integer, could be employed to express such conditional causality.

Above we have defined abductive diagnosis using propositional logic. The definition in terms of predicate logic reveals some additional subtleties, yielding various alternative definition for the set of findings not observed and assumed to be absent, E^c. Findings f are denoted in predicate logic using a predicate symbol p, indicating a particular group of findings or a test. For example, in 'Sign(*fever*)', the predicate symbol 'Sign' denotes a group of patient findings; in 'Serum_copper(*patient*, *high*)', the predicate symbol 'Serum_copper' indicates the result of a diagnostic test. The consequences of using predicate logic to define abductive diagnosis will be briefly introduced by means of the following example.

Example 2.4. Consider the following (partial) set of abnormality axioms $\mathcal{R}$, expressed in first-order predicate logic as follows:

$$\begin{array}{lcl} d_1 & \rightarrow & p(a) \\ d_1 & \rightarrow & q(b) \\ d_2 & \rightarrow & r(c) \end{array}$$

where the (ground) literals $p(a)$, $q(b)$ and $r(c)$ stand for findings, and the literals d_i, $i = 1, 2$, represent defects. The finding literals f, representing abnormal observable findings, are taken from the following set of positive finding literals:

$$\Phi_P = \{p(a), p(d), q(b), q(e), r(c), r(f)\}$$

and the set of negative finding literals is equal to

$$\Phi_N = \{\neg p(a), \neg p(d), \neg q(b), \neg q(e), \neg r(c), \neg r(f)\}$$

with $\Phi = \Phi_P \cup \Phi_N$. Now, let $E = \{q(b), r(c)\}$ be a set of observed findings. Usually, it is assumed that $E \subseteq \Phi_P$, because only positive findings can be accounted for by Horn clauses in $\mathcal{R}$. The set $E^c \subseteq \Phi$, representing the findings not observed is taken to be defined in accordance with Definition 2.3. In the present case, the set E^c is equal to

$$E^c = \{\neg p(a), \neg p(d), \neg q(e), \neg r(f)\}$$

Thus, test results denoted by the predicate symbol 'p' are assumed to be absent. Note that when applying this version of the consistency definition, obtained by the definition of E^c, the defect d_1 cannot be part of any diagnosis, because this would clash with the consistency condition. Although, on first thought, the set $\{d_2\}$ may seem to represent a diagnosis, it turns out that there exists no diagnosis at al. The reason is that

$$\mathcal{R} \cup \{d_2\} \not\models q(b)$$

i.e., the covering condition fails to hold.

A second alternative version of the theory is presented in [Console & Torasso, 1990b] and [Console & Torasso, 1991]. In these articles, the consistency condition is reformulated, by adopting another definition for the set E^c, as follows. The set $E^c \subseteq \Phi_N$ is defined by:

$$E^c = \{\neg\pi(t) \in \Phi_N \mid \pi(s) \in E, t \neq s\}$$

where π stands for a predicate symbol, and t and s are constants. The consistency condition remains the same, but its effects on the computation of a diagnosis differs, because of the altered definition of E^c. For the example diagnostic problem, the set E^c is equal to

$$E^c = \{\neg q(e), \neg r(f)\}$$

Note that the literals $\neg p(a)$ and $\neg p(d)$ are missing from this set, because none of the literals in the set of observed findings E has p as predicate symbol. Thus, the test results with respect to test 'p' are assumed to be unknown. A diagnosis in this case is $H = \{d_1, d_2\}$, because

$$\mathcal{R} \cup H \models \{q(b), r(c)\}$$

(in fact, the literal $p(a)$ is also entailed), and E^c is consistent with $\mathcal{R}$ and H. Note that $H = \{d_1, d_2\}$ yields an inconsistency if taken as a diagnosis using the first version of the consistency condition. $\diamondsuit$

The intuitive basis of the two versions of the consistency condition in abductive diagnosis, yielded by different logical interpretations of findings not observed, can be clarified in terms of diagnostic problem solving as follows. (We remark that this interpretation is the author's own, no such interpretation appears in the papers by Console and Torasso.) In the first version of the consistency condition, it is assumed that all findings associated with a defect, present in the real world, will be observed. If a finding is not included among the findings in the set of observed findings, it is assumed to be absent; absent findings are denoted by negative literals. The basic assumption is that all findings of defects that are absent will not be observed, i.e. are absent (if unique for the defect), hence, it can safely be assumed that all findings not observed are negative. Although this may not be justified in diagnostic problem solving – it could be more natural to take the findings as unknown – the assumption of the negative literals has the technical advantage of blocking the inclusion of defects that are not present in the real world according to the theory, because some observable finding associated with the defect is not included in the set of observed findings. This is precisely the effect required. Now, if, as in the example above, only part of the unique findings of a defect occurs among the set of observed findings, there must be something wrong, either with the abnormality axioms $\mathcal{R}$, or with the set of observed findings. It seems therefore justified that no diagnosis is established in this case. However, this result is only valid if one accepts as a basic assumption that every possible cause (defect) of a finding is included in the set of abnormality axioms $\mathcal{R}$, which also constituted the basis of the predicate completion discussed above (at the risk of ambiguity with respect to database theory, one might call this the closed world assumption of abduction).

The second version of the consistency condition in abductive diagnosis is similar to the first version, except that it is assumed that if no information concerning a specific diagnostic test is available,– recall that every test corresponds to a different predicate symbol – it is assumed to be unknown. Now, if some defect d is included in a solution H and

$$\mathcal{R} \cup \{d\} \vDash f$$

where $f \notin E$, this means that the model predicts that if the test is actually carried out, the finding f will be observed. If it is not observed, or turns out to be false, i.e. $\neg f$, some action needs to be undertaken, but no specific ideas concerning this situation appear in the papers of Console and Torasso. However, if the test has been carried out, i.e. there exists some finding f' with the same predicate symbol as f, and $f \notin E$, then again no diagnosis exists, because $\neg f \in E^c$ would hold.

The abductive theory of diagnosis discussed above may be viewed as a formalization of particular parts of the expert system shell CHECK [Console & Torasso, 1989; Torasso & Console, 1989]. This system can be used to build hybrid diagnostic systems for domains in which causal, hierarchical and heuristic knowledge coexist. As far as known to the author, CHECK has been used as an experimental platform on which various prototype systems have been developed, including diagnosis of automobile engine failure and diagnosis of liver disease; none of the systems built have been assessed with respect to their diagnostic accuracy. Consequently, Console and Torasso have not shown as yet that the application of causality as the primary modelling concept aids in building real-life applications.

2.2.3 Set-covering theory of diagnosis

Instead of choosing logic as the language for MAB diagnosis, as discussed above, others have adopted set theory as their formal language. This approach to the formalization of diagnosis is referred to as the *set-covering theory of diagnosis*, or *parsimonious covering theory* [Reggia et al., 1983; Allemang et al., 1987; Peng & Reggia, 1990; Wu, 1991]. The treatment of the set-covering theory of diagnosis in the literature deals only with the modelling of restricted forms of abnormal behaviour of a system.

The specification of the knowledge involved in diagnostic problem solving consists of the enumeration of all findings that may be present (and observed) given the presence of each individual defect distinguished in the domain; the association between each defect and its associated set of observable findings is interpreted as an uncertain *causal relation* between the defect and each of the findings in the set of observable findings. Instead of the terms ‘defect’ and ‘finding’ the terms ‘disorder’ and ‘manifestation’ are employed in descriptions of the set-covering theory of diagnosis. In the following, we have chosen to uniformly employ the terms ‘defect’ and ‘finding’ instead. The basic idea of the theory with respect to diagnosis is that each finding in the set of observed findings in a given diagnostic situation must be causally related to at least one present defect; the collected set of present defects thus obtained can be taken as a diagnosis. As with the theory of diagnosis by Console and Torasso, this reasoning method is usually viewed as being abductive in nature, because the reasoning goes from findings to defects, using causal knowledge from defects to findings.

More formally, the triple $\mathcal{N} = (\Delta, \Phi, C)$ is called a *causal net* in the set-covering theory of diagnosis, where

- Δ is a set of *defects*,
- Φ is a set of elements called *observable findings*, and
- C is a binary relation

  $$C \subseteq \Delta \times \Phi$$

  called the *causation relation*.

A *diagnostic problem* in the set-covering theory of diagnosis is then defined as a pair $\mathcal{D} = (\mathcal{N}, E)$, where $E \subseteq \Phi$ is a *set of observed findings*. It is assumed that all defects $d \in \Delta$ are potentially present in a diagnostic problem, and all findings $f \in \Phi$ will be observed when present. In addition, all defects $d \in \Delta$ have a causally related observable findings $f \in \Phi$, and vice versa, i.e. $\forall d \in \Delta \exists f \in \Phi : (d, f) \in C$, and $\forall f \in \Phi \exists d \in \Delta : (d, f) \in C$. No explicit distinction is made in the theory between positive (present), negative (absent) and unknown defects, and positive (present), negative (absent) and unknown findings. The causation relation is often depicted by means of a labelled, directed acyclic graph, which, as $\mathcal{N}$, is called a *causal net* [Peng & Reggia, 1990].

Let $\wp(X)$ denote the power set of the set X. It is convenient to write the binary causation relation C as two functions. Since in the next chapter, such functions are intensively employed, we adopt a notation that slightly generalizes the notation proposed in [Peng & Reggia, 1990]. The first function

$$e : \wp(\Delta) \to \wp(\Phi)$$

called the *effects function*, is defined as follows; for each $D \subseteq \Delta$:

$$e(D) = \bigcup_{d \in D} e(\{d\})$$

where

$$e(\{d\}) = \{f \mid (d, f) \in C\}$$

and the second function

$$c : \wp(\Phi) \to \wp(\Delta)$$

called the *causes function*, is defined as follows; for each $E \subseteq \Phi$:

$$c(E) = \bigcup_{f \in E} c(\{f\})$$

where

$$c(\{f\}) = \{d \mid (d, f) \in C\}$$

Hence, knowledge concerning combinations of findings and defects can be taken as being composed of knowledge concerning individual defects or findings (as shall become clear in chapters 3 and 4, this is not acceptable in general).

A causal net can now be redefined, in terms of the effects function e above, as a triple $\mathcal{N} = (\Delta, \Phi, e)$.

Given a set of observed findings, diagnostic problem solving amounts to determining sets of defects – technically the term *cover* is employed – that account for *all* observed findings. Formally, a diagnosis is defined as follows.

Definition 2.6 (*set-covering diagnosis*). *Let $\mathcal{D} = (\mathcal{N}, E)$ be a diagnostic problem, where $\mathcal{N} = (\Delta, \Phi, e)$ is a causal net and E denotes a set of observed findings. Then, a* (set-covering) diagnosis *of $\mathcal{D}$ is a set of defects $D \subseteq \Delta$, such that:*

$$e(D) \supseteq E \tag{2.9}$$

Due to the similarity of this condition with the covering condition in the abductive theory of diagnosis, condition (2.9) is called the *covering condition* in the set-covering theory of diagnosis. In the set-covering theory of diagnosis the technical term 'cover' is employed instead of 'diagnosis'; 'diagnosis' will be the name adopted in this section.

Since it is assumed that $e(\Delta) = \Phi$ is satisfied, i.e. any finding $f \in \Phi$ is a possible causal effect of at least one defect $d \in \Delta$, there exists a diagnosis for any set of observed findings E, because

$$e(\Delta) \supseteq E$$

always holds (explanation existence theorem, [Peng & Reggia, 1990]).

A set of defects D is said to be an *explanation* of a diagnostic problem $\mathcal{D} = (\mathcal{N}, E)$, with E a set of observed findings, if D is a diagnosis of E and D satisfies some additional criteria. Various criteria, in particular so-called *criteria of parsimony*, are in use. The basic idea is that among the various diagnoses of a set of observable findings, those that satisfy certain criteria of parsimony are more likely than others. Let $\mathcal{D} = (\mathcal{N}, E)$ be a diagnostic problem, then some of the criteria as mentioned in [Peng & Reggia, 1990; Tuhrim et al., 1991] are:

- *Minimal cardinality*: a diagnosis D of E is an explanation of $\mathcal{D}$ iff it contains the minimum number of elements among all diagnoses of E;

- *Irredundancy*: a diagnosis D of E is an explanation of $\mathcal{D}$ iff no proper subset of D is a diagnosis of E;

- *Relevance*: a diagnosis D of E is an explanation of $\mathcal{D}$ iff $D \subseteq c(E)$;

- *Most probable diagnosis*: a diagnosis D of E is an explanation of $\mathcal{D}$ iff $P(D|E) \geq P(D'|E)$ for any diagnosis D' of E.

We shall speak of a ‘minimal cardinality diagnosis’, an ‘irredundant diagnosis’, etcetera. Although not every diagnosis is an explanation, any diagnosis may be seen as a solution to a diagnostic problem, where diagnoses which represent explanations conform to more strict conditions than diagnoses that do not. The term ‘explanation’ refers to the fact that a diagnosis in the set-covering theory of diagnosis can be stated, and thus be explained, in terms of cause-effect relationships. A better choice in our opinion, would have been the adoption of the term ‘explanation’ for what is now called ‘cover’ in the theory, and to refer to what are now called ‘explanations’ by the name of ‘parsimonious explanations’.

For minimal cardinality, a diagnosis which consists of the smallest number of defects among all diagnoses is considered the most likely diagnosis. Minimal cardinality is a suitable parsimony criterion in domains in which large combinations of defects are unlikely to occur. For example, in medicine, it is generally more likely that a patient has a single disorder than more than one disorder. Irredundancy expresses that it is not possible to leave out a defect from an explanation without losing the capability of explaining the complete set of observed findings, i.e.

$$e(D) \not\supseteq E$$

for each $D \subset D'$, where D' is an irredundant diagnosis. The relevance criterion states that every defect in an explanation has at least one observable finding in common with the set of observed findings. This seems an obvious criterion, but note that the notion of uncertain causal relation employed in the set-covering theory of diagnosis does not preclude situations in which a defect is present, although none of its causally related observable findings have been observed. These three definitions of the notion of explanation are based on general set-theoretical considerations. In contrast, the most probable diagnosis embodies some knowledge of the domain, in particular with respect to the strengths of the causal relationships. We shall not deal with such probabilistic extensions of the set-covering theory of diagnosis any further.

Example 2.5. Consider the causal net $\mathcal{N} = (\Delta, \Phi, C)$, where the effects function e is defined by the causation relation C, i.e.

$$e(D) = \bigcup_{d \in D} e(\{d\})$$

where

$$e(\{d\}) = \begin{cases} \{f_1, f_2, f_3\} & \text{if } d = d_1 \\ \{f_1, f_3\} & \text{if } d = d_2 \\ \{f_2, f_4\} & \text{if } d = d_3 \end{cases}$$

The associated graph representation G_C of C is shown in Figure 2.6; it states, for example, that $e(\{d_1, d_2\}) = \{f_1, f_2, f_3\}$. Suppose the causal net represents medical knowledge, expressed by means of the following correspondence:

d_1:	influenza
d_2:	common cold
d_3:	pneumonia

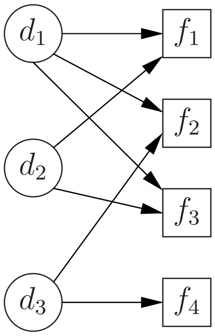


Figure 2.6: Causal net.

f_1:	cough
f_2:	fever
f_3:	sneezing
f_4:	dyspnoea (shortness of breath)

For example, a patient with influenza will be coughing, sneezing and have a fever; a patient with a common cold will show the same findings, except fever, and a patient with pneumonia will have a fever and dyspnoea. Based on the causal net C, the following causes function c is obtained:

$$c(E) = \bigcup_{f \in E} c(\{f\})$$

with

$$c(\{f\}) = \begin{cases} \{d_1, d_2\} & \text{if } f = f_1 \\ \{d_1, d_3\} & \text{if } f = f_2 \\ \{d_1, d_2\} & \text{if } f = f_3 \\ \{d_3\} & \text{if } f = f_4 \end{cases}$$

Suppose $\mathcal{D} = (\mathcal{N}, E)$ is a diagnostic problem, with $E = \{f_1, f_2\}$ a set of observed findings, then a diagnosis of E is $D_1 = \{d_1\}$, but $D_2 = \{d_1, d_2\}$, $D_3 = \{d_2, d_3\}$, and $D_4 = \{d_1, d_2, d_3\}$ are also diagnoses for E. All of these diagnoses are relevant diagnoses, because $c(\{f_1, f_2\}) \supseteq D_i$, $i = 1, \ldots, 4$. Irredundant diagnoses of E are $D_1 = \{d_1\}$ and $D_3 = \{d_2, d_3\}$. There is only one minimal cardinality diagnosis, viz. $D_1 = \{d_1\}$. Now suppose that $E = \{f_1\}$, then for example $D = \{d_1, d_3\}$ would not have been a relevant diagnosis, because

$$c(\{f_1\}) = \{d_1, d_2\} \not\supseteq D$$

$\diamondsuit$

Other, more domain-specific, definitions of the notion of explanation have only been developed recently. Such domain-specific knowledge can be effective in reducing the size of the set of diagnoses generated by a diagnostic system. For example, [Tuhrim et al., 1991] demonstrated that the use of knowledge concerning the three-dimensional structure of the brain by means of a binary adjacency relation in a neurological diagnostic expert system based on the set-covering theory of diagnosis, could increase the diagnostic accuracy of the system considerably.

In [Peng & Reggia, 1990], it is shown that the causation relation C can be extended for the representation of multi-layered causal nets, in which defects are causally connected to each other, finally leading to observable findings. By computation of the reflexive, transitive closure of the causation relation, $C^\star$, the basic techniques discussed above immediately apply.

In the set-theoretical formalization of diagnosis by Bylander et al., [Bylander et al., 1992], an effects function e is used to represent both the knowledge base and the method of diagnostic problem solving. In contrast to the theory by Peng and Reggia, the function e can be used to represent diagnostic interactions among defects, because the assumption that $e(D)$ is the union of function values $e(\{d\})$, for each $d \in D$, is not generally assumed. A diagnosis D is defined by $e(D) = E$, where E is a set of observed findings, i.e. every finding must be covered by the set of defects D.

INTERNIST-1/QMR is an example of an expert system with a basis related to the set-covering theory [Miller et al., 1982; Peng & Reggia, 1990]. However, the system is not a direct implementation of the theory reviewed above; in fact, the system predates the theory for about a decade. The way inference in the system has been implemented, bears some resemblance with the set-covering theory of diagnosis. However, it deviates from this theory in several respects, in particular by employing domain-specific heuristics in the inference process [Peng & Reggia, 1990]. RED is an expert system in the domain of blood bank antibody analysis [Josephson & Josephson, 1994; Punch III et al., 1990; Smith et al., 1985]. This system can also be described in terms of the set-covering theory of diagnosis, although several aspects of the system go beyond the theory, such as the representation of interactions among particular antibody reactions, requiring a generalization of the set-covering theory [Bylander et al., 1992]. Peirce is a domain-independent tool that generalized on the techniques used in RED [Punch III et al., 1990]. In [Tuhrim et al., 1991], an expert system for the diagnosis of brain lesions, based on the set-covering theory of diagnosis is described. Of the systems mentioned above, the last system is based most clearly on the principles described in this section.

2.2.4 Hypothetico-deductive diagnosis

The third approach to diagnosis mentioned in Section 2.1, AC (Abnormality Classification) diagnosis, originates from work by E.H. Shortliffe, B.G. Buchanan, W.J. Clancey and E.A. Feigenbaum in the MYCIN project [Shortliffe, 1976; Buchanan & Shortliffe, 1984; Clancey & Letsinger, 1984]. The knowledge incorporated in that expert system, and in similar systems for AC diagnosis, is based on the body of experience accumulated in handling a large number of cases, such as the patients a physician sees in medical practice. The knowledge is extracted from textbooks or human experts. We have called this type of knowledge empirical associations.

In most practical systems, including the HEPAR system [Lucas et al., 1989], that will be discussed in the second part of this thesis, the formal counterparts of empirical associations are organized according to some underlying model distinguished in the collection of empirical associations. A typical example is a distinction between families of disorders and specific disorders, i.e. a taxonomy of disorders, that can be exploited in problem solving. Hence, expert systems based on empirical associations are model-based like the other systems discussed above, because they are also based on a model of the problem domain, although the nature of the model is different. It is possible to characterize AC diagnosis in a more formal way. We shall refer to this formal counterpart of AC diagnosis as *hypothetico-deductive diagnosis*, a term suggested in [Campbell, 1976] and [Macartney, 1988].

A hypothetico-deductive diagnostic problem consists of a set of logical axioms $\mathcal{R}$ of the form

$$c_1 \wedge \cdots \wedge c_n \to q \tag{2.10}$$

where c_i and q represent either negative or positive defects and findings, represented in logic as negative or positive literals, and if every c_i is a finding, then q should be a defect. Logical implication in the formalization of empirical associations (2.10) may be viewed as a classification relation. A set of observed findings is represented as a set of ground literals, where each literal is of the finding type. For example, a typical logical axiom might be

$$f_1 \wedge \cdots \wedge f_m \to d$$

which expresses that a set of observable findings $F = \{f_1, \ldots, f_m\}$ represents necessary and sufficient evidence for establishing the presence of the defect d as part of a diagnosis. Examples of such implications, taken from the HEPAR system, are presented in Chapter 7, in the second part of this thesis. One difference between the theories of hypothetico-deductive diagnosis and abductive diagnosis is that, in hypothetico-deductive diagnosis, observed findings and defects need not be causally related to each other. Some of the findings may be interpreted as abnormal; other findings, such as, for example, age of a patient in a medical application, may not. The function of normal findings in empirical associations is similar to that of conditional causality introduced in Section 2.2.2, viz. to condition a particular piece of knowledge on a specific piece of evidence.

Now, let $\mathcal{B} = (\Delta, \Phi, \mathcal{R})$ denote an *associational specification*, where:

- Δ denotes a set of (positive and negative) defects,
- Φ denotes a set of (positive and negative) observable findings, and
- $\mathcal{R}$ denotes the logical representation of a set of empirical associations of the form (2.10).

A *hypothetical-deductive diagnostic problem* is then defined as a pair $\mathcal{H} = (\mathcal{B}, E)$, where $E \subseteq \Phi$ denotes a *set of observed findings*. A diagnosis based on empirical associations can be defined as follows.

Definition 2.7 (*hypothetico-deductive diagnosis*). *Let $\mathcal{H} = (\mathcal{B}, E)$ be a hypothetico-deductive diagnostic problem, where $\mathcal{B} = (\Delta, \Phi, \mathcal{R})$ is an associational specification, and E is a set of observed findings. Let $\Theta \subseteq \Delta$ be a set of defects, called a* hypothesis*. Then, $D \subseteq \Theta$ is called a* (hypothetico-deductive) diagnosis *of $\mathcal{H}$ if*

$$D = \{d \in \Theta \mid \mathcal{R} \cup E \vDash d\}$$

Note that, in contrast with the theories discussed above, a single hypothesis is initially given in hypothetico-deductive diagnosis; it stands for the defects that are initially given to be of interest. In the theory of hypothetico-deductive diagnosis, defects are logically entailed by the observed findings (usually implemented by a deductive calculus, hence the adjective hypothetico-*deductive*).

In contrast with the other theories of diagnosis, there are a large number of nonexperimental applications available that may be viewed as hypothetico-deductive diagnostic systems; the HEPAR system, [Lucas et al., 1989; Lucas & Janssens, 1991a], discussed in the second part of this thesis is an example.

2.2.5 Relationships between theories of diagnosis

Having described the various formal theories of diagnosis, the question arises in what sense these theories are related to each other. Several originators of theories of diagnosis have investigated the expressiveness of their theory for modelling other conceptual models of diagnosis than those for which the theory was originally designed. In this section, we summarize and comment on results found in the literature.

Reiter has shown that the framework of consistency-based diagnosis provides enough descriptive power to capture the set-covering theory of diagnosis [Reiter, 1987]. In Reiter's formalization, the normality axioms in the original theory of consistency-based diagnosis are changed into *abnormality axioms*, simply by replacing components by defects. These axioms have the following form

$$\neg\text{Abnormal}(d) \rightarrow \neg\text{Present}(d) \tag{2.11}$$

for each defect d, stating that under normal conditions defect d is not present, and

$$f_{ab} \rightarrow \text{Present}(d_1) \vee \cdots \vee \text{Present}(d_n) \tag{2.12}$$

for each observed abnormal finding f_{ab} and defects d_i, $1 \leq i \leq n$. Formulae of the form (2.11) express hypotheses, namely that a particular defect may be absent ($\neg\text{Present}(d)$) if it does not give rise to an inconsistency. As we have discussed in Section 2.2.2, formulae of the form (2.12) may be seen as the predicate completion, [Clark, 1978], of finding literals in formulae of the form

$$\text{Present}(d) \rightarrow f_{ab}$$

i.e. if $\mathcal{R}$ denotes the set of formulae of the last form, with defect literals $\text{Present}(d_1), \ldots, \text{Present}(d_n)$ in the premise, then the predicate completion $\text{COMP}[\mathcal{R}; f_{ab}]$ with regard to the finding f_{ab} is equal to

$$\text{COMP}[\mathcal{R}; f_{ab}] = \mathcal{R} \cup \{f_{ab} \rightarrow \text{Present}(d_1) \vee \cdots \vee \text{Present}(d_n)\}$$

This states that the only causes of the finding f_{ab} to be present (and observed) are the defects $d_1, \ldots, d_n$. As discussed above, this same kind of knowledge is expressed, although implicitly, in the abductive theory of diagnosis; it is also expressed in the set-covering theory of diagnosis, but the differences between the reasoning methods employed (consistency-based reasoning, logical abduction, and set covering) dictate a different representation (syntax) in all three formal theories. Informally, in the consistency-based diagnosis formalization of MAB diagnosis, diagnostic problem solving is carried out as follows. Given an observed finding f_{ab} associated with a defect d_i, $1 \leq i \leq n$, a disjunction

$$\text{Present}(d_1) \vee \cdots \vee \text{Present}(d_n)$$

is deduced, which is reduced by cancelling out atoms using axiom (2.11), assuming certain defects not to be present, i.e. Abnormal(d) is *false*, yielding a (subset minimal) diagnosis. The effect of axiom (2.11) corresponds to producing irredundant diagnoses in the set-covering theory of diagnosis, in the sense that a minimal diagnosis with respect to set inclusion is produced. Reiter shows that there exists a (subset minimal) diagnosis according to the consistency-based reformulation of the set-covering theory of diagnosis iff there exists an equivalent irredundant diagnosis in the set-covering theory (although at the time Reiter's result was published, the notion of irredundant diagnosis had not yet appeared in the literature) [Reiter, 1987].

Console and Torasso have studied the use of the consistency condition in abductive diagnosis for modelling DNSB diagnosis, i.e. diagnosis using a specification of a model of normal structure and behaviour in a way resembling the work of Reiter [Reiter, 1987; Console & Torasso, 1990b; Console & Torasso, 1991]. By taking the empty set for the set of observed findings that must be covered, the covering condition in abductive diagnosis becomes

$$\mathcal{R} \cup H \vDash E'$$

where $E' = \varnothing$; a diagnosis is the result of satisfaction of the consistency condition only, because the covering condition is always satisfied in this case. Thus consistency-based diagnosis in the sense of Reiter is obtained. However, the meaning of the logical axioms is entirely different from the meaning originally attached to the logical axioms, because they now represent normal behaviour of a device; d represents some normal state of a component of the device and a finding f in the conclusion of a Horn clause $d \to f$ represents a finding that may be observed when the component is in its normal state, i.e. f represents a normality finding f_{norm}. By varying between $E' = \varnothing$ and $E' = E$, for example by taking for E' the set of all abnormal findings f_{ab} occurring in E, DNSB and MAB diagnosis can be integrated within the same abductive framework [Console & Torasso, 1990b; Console & Torasso, 1991].

Finally, the set-covering theory has much in common with the abductive theory of diagnosis, in that both focus on the application of causal knowledge for diagnostic problem solving. In fact, as shall be shown in Chapter 4, if the abnormality axioms in the abductive theory of diagnosis are restricted to the weakened axioms, the theories yield similar notions of diagnosis. The abductive theory of diagnosis is more expressive in that different kinds of knowledge required to describe abnormal behaviour can be represented, and in that interactions among defects can be described. Chapter 4 provides a more in-depth account of the properties of the various theories of diagnosis. The technical characteristics of the various formal theories of diagnosis are summarized in Table 2.1.

Originator	Knowledge base specification	Knowledge base interpretation	Diagnosis
Reiter	functional relations	deduction	consistency
Console & Torasso	causality	abduction	covering
		deduction	consistency
Reggia et al.	causality	abduction	set covering
Bylander et al.	diagnostic relation	none	set covering
Shortliffe et al.	empirical associations	deduction	classification

Table 2.1: Comparison of formal theories of diagnosis.

We may conclude by saying that generalization of the formal theories of diagnosis discussed above has shown that there is no such thing as a unique formalization of a conceptual model of diagnosis. The formal theories can be applied to formalize conceptual models of diagnosis other than those for which they were originally designed. The resulting formalizations, however, often lack clarity.

2.3 Discussion

The overview of the various approaches to diagnostic problem solving presented above indicates that, on the one hand, several different formalizations of the same conceptual model of diagnosis exist, whereas, on the other hand, several different conceptualizations of diagnosis fit into the same formal framework. Unfortunately, the various conceptual models of diagnosis presented in the literature are still commonly referred to by the name of their formal counterpart, suggesting that a unique linkage does exist between a formal theory and a conceptual model of diagnosis.

Each formal theory of diagnosis discussed has originally been developed to capture one specific conceptual approach to diagnosis. This remains visible, in spite of attempts of generalization. They seem too intimately linked with their conceptual bases to be taken as genuine formal frameworks of diagnosis. The theory of consistency-based diagnosis as proposed by Reiter, [Reiter, 1987], and De Kleer et al., [De Kleer et al., 1992], provides a framework of both DNSB and MAB diagnosis, although the theory appears rather cumbersome for expressing MAB diagnosis. It does not provide a suitable basis for AC diagnosis. As the theory is based on the general notion of (un)satisfiability, it is not expressive enough to capture many of the essential features of diagnostic problem solving in a straightforward way. Unsatisfiability may be a suitable notion to describe deviation from the normal situation of a device, but as a model for the description of the relationships among defects and findings it is highly unnatural.

In the abductive theory of diagnosis proposed by Console and Torasso, specific assumptions are made with respect to the causal nature of the knowledge involved. Their formalization assumes that logical implication provides a suitable axiomatization of the notion of causality. However, only the transitive nature of logical implication seems to match the properties of causality; the reflexive and contrapositive properties of implication will not hold for all notions of causality. The interpretation of causal knowledge in the theory by Console and Torasso is achieved through the logical entailment or deduction relation, which is also used, together with the consistency and covering condition, to define notions of diagnosis. Hence, no clear distinction is made between the interpretation of a knowledge base – to determine what logically follows from the knowledge base – and applying this interpretation to determine a diagnosis. Furthermore, by the monotonicity of the deduction relation, certain types of knowledge, e.g. knowledge in which observable findings are cancelled due to interaction among defects, are precluded from formalization (at least in standard logic). The application of nonmonotonic inference rules, such as SLDNF resolution, offers additional flexibility, but at a cost. For example, the interpretation of negative information as failure to prove positive information under the negation as failure rule, may not be appropriate in every domain. The restriction to Horn clause or general Horn clause logic (normal logic programs) may also be too strong. The theory lends itself for describing a domain in terms cause-effect relations between changing values of parameters of processes. Not every domain for diagnosis can be described in these terms.

The set-covering theory of diagnosis aims, like the abductive theory of diagnosis, at describing a domain in terms of causality. Unlike the abductive theory of diagnosis, only a single concept of causality is employed, which is only made more expressive by the interpretation of causal relations as conditional probabilities [Peng & Reggia, 1990], yielding a formalism that is much alike the belief-network formalism [Lucas & Van der Gaag, 1991; Pearl, 1988]. Furthermore, a knowledge base consists only of a specification of single defects in terms of associated findings; in this way, it is not possible to model interactions among defects. Moreover, their notion of diagnosis (explanation) is fixed, with the exception of the notion of minimal diagnosis, which is variable in the theory. Finally, no clear distinction between the interpretation of a knowledge base in terms of causality and the process of diagnosis is made.

The diagnostic theory by Bylander et al., [Bylander et al., 1992], is more expressive than the set-covering theory by Reggia et al., [Peng & Reggia, 1990], in the sense that it is possible to express interactions among defects. However, both the specification of the knowledge base, the interpretation of the knowledge in the knowledge base, as well as using the knowledge base to establish a diagnosis, are expressed by means of a single function. Hence, no distinction between knowledge base interpretation and diagnosis is made.

As we have seen, in some of the formal theories above, diagnoses need not be unique, in which case certain criteria may be applied to select the diagnoses that best fit these criteria, better known as criteria of parsimony. An example of such a criterion is minimality according to set inclusion. Intuitively, the idea is to try to account for as many of the observed findings as possible with a minimal number of assumptions, such as defects assumed to be present. Several alternative diagnoses fulfilling the minimality criteria may then exist. Although minimality assumptions are made in all theories, with the exception of hypothetico-deductive diagnosis, we do not consider such minimality criteria to be essential features of notions of diagnosis, but rather to be refinements to basic diagnostic notions, used for diagnosis selection. The main reason why such minimality assumptions seem to be incorporated, is that most of the notions of diagnosis discussed above are too weak to limit the diagnoses generated to an acceptably small number. Designing notions of diagnosis that would be more restrictive could be an attractive alternative, which, however, has not been investigated systematically. Furthermore, note that the concept of subset-minimal diagnosis is only sound if the relationships between defects and findings accounted for is monotonic, a requirement nowhere stated explicitly in the literature. In the following chapter, it is shown that this requirement may be too strong for practical applications.

Consistency-based, abductive and set-covering diagnosis are all classified as nonmonotonic theories of diagnosis in the literature. The nonmonotonic nature of particular notions of diagnosis follows from the feature that the addition of a newly observed finding to a diagnostic model may require the removal of a previous diagnostic conclusion – to maintain satisfiability in consistency-based diagnosis, or to satisfy the covering or consistency condition in the abductive theory of diagnosis, or the covering condition in the set-covering theory of diagnosis. In the work by Reiter on consistency-based diagnosis, subset minimality makes it rather straightforward to investigate the meaning of diagnostic problem solving using default logic (cf. [Reiter, 1987]); Console and Torasso resort to circumscription to study abductive diagnosis for the same reason (cf. [Console & Torasso, 1990a]). However, as said above, there are other relationships in these theories of diagnosis that are monotonic (cf. Chapter 4). Hypothetico-deductive diagnosis a monotonic notion of diagnosis in the same sense as the other theories of diagnosis are nonmonotonic.

This chapter is rounded off by a number of observations regarding similarities between the various formal theories of diagnosis discussed above. These similarities can be made explicit by a number of parameters that can be used to characterize the formal theories of diagnosis. These parameters will turn out to be useful for designing a general framework of diagnosis, which incorporates the theories above as special instances. Based on the previous descriptions of the formal theories of diagnosis, the following parameters appear relevant:

- The way in which knowledge that is used for the purpose of diagnosis is specified, among which the possibility to express interactions among defects, and the interpretation of this knowledge for the purpose of diagnosis;

- The way in which observed findings are interpreted in terms of the described defects (or disorders, faults) yielding a diagnosis, e.g. due to the deviation of observed findings from expected normal findings or matching of observed findings with normal or abnormal findings. This includes the distinction in interpretation of negative, positive and unknown findings and defects (and what is actually meant by positive/negative findings and defects);

- The selection of diagnoses from a set of alternatives generated (by some criterion of parsimony), i.e. diagnosis selection, which is in accordance with diagnostic problem solving as described in the previous chapter;

- The gathering of findings in the diagnostic process.

In the next chapter, these aspects of diagnosis, with the exception of the gathering of findings, which is a dynamic aspect of diagnosis, will be incorporated in a set-theoretic framework of diagnosis. Note, however, that although set theory is chosen as a language to formalize various notions of diagnosis, this does not mean that the set-covering theory of diagnosis will be adopted as our framework.

Chapter 3

A Framework for Diagnostic Problem Solving

In the previous chapter, the conceptual basis of diagnosis was reviewed, together with several different formal theories of diagnosis as presented in the literature. In this chapter, a formal framework of diagnosis is developed, taking the conceptual basis of diagnosis as a starting point. The main goal is to obtain a framework of sufficient generality, such that the various formal theories of diagnosis can be shown to fit into the framework. As an immediate consequence of this endeavour, the various formal theories can be analysed in terms of the framework. This enables a comparison between the formal theories. Insight into the basic assumptions and associated restrictions of the formal theories of diagnosis, and also into the relationships between them, will be the main result of this work. The analysis will be undertaken in Chapter 4. In Chapter 5, a number of new notions of diagnosis, meant to capture flexible diagnosis in imperfect real-world knowledge bases, are explored in terms of the framework.

The framework will be built up gradually in terms of the most important static principles of diagnostic problem solving. Its underlying assumption is that diagnosis involves the interpretation of a knowledge base in terms of observable findings and possible defects. Given a set of observed findings and a hypothesis to be investigated, a diagnosis can be established using this diagnostic interpretation of a knowledge base. Hence, the type of knowledge represented in a knowledge base, the way in which this knowledge is interpreted in a diagnostic sense, as well as the interpretation of hypotheses and observed findings in the context of this knowledge, determine the diagnoses for a diagnostic problem.

The present chapter is organized as follows. In Section 3.1, the notion of 'evidence function' is proposed; it stands for an interpretation of a specification of knowledge for the purpose of diagnosis. An evidence function might be the result of a translation of a given knowledge base. Next, an evidence function is interpreted with respect to hypotheses and sets of observed findings, yielding diagnoses. This second interpretation is carried out by partial functions, called 'notions of diagnosis'. Evidence functions and notions of diagnosis are the essential building-blocks of our framework. Several properties of notions of diagnosis, useful to characterize and analyse formal theories of diagnosis, are studied in Section 3.2. In Section 3.3, we turn to the problem of selecting plausible diagnoses from a collection of alternative diagnoses, i.e. diagnosis selection. Finally, in Section 3.4, the framework of diagnosis is related to other work in the field of diagnosis.

3.1 Basic notions

The fundamental assumptions underlying the framework of diagnosis presented in this chapter are:

(1) a diagnostic problem can be solved through the observation of findings associated with defects that are either present or absent, and

(2) findings can be observed with complete certainty.

The term '*defect*' is used as the collective term for various related concepts from different fields, such as disorder or disease in medicine, and fault or failed component in technical applications. Findings are sometimes referred to as '*symptoms*', [Poole, 1988; Poole, 1994], or '*manifestations*', [Reggia et al., 1983; Peng & Reggia, 1990; Console et al., 1989]. Diagnostic knowledge represented in a knowledge base is viewed as a relationship between observable findings and possible defects; diagnostic problem solving is viewed as the problem of selecting from a set of possible defects (taken as a hypothesis), subsets that account for a given set of observed findings. Recall that this view of diagnostic problem solving is essentially static in nature.

The framework is composed of the following three basic ingredients, which are common to formal theories of diagnosis, as concluded in Section 2.3:

- an interpretation of a knowledge base in terms of relationships between (present or absent) defects and findings (called an 'evidence function', see Section 3.1.1);
- an interpretation of the defects-findings relationships captured in a knowledge base to establish diagnoses for given hypotheses and sets of observed findings (called a 'notion of diagnosis', see Section 3.1.4);
- criteria for the selection among the established diagnoses (called 'diagnosis selection', see Section 3.3).

In Figure 3.1, these three aspects of the framework are related to each other. For all three aspects, formal counterparts will be developed in the next subsections using set-theoretical principles.

Compared with the theories of diagnostic problem solving described in the previous chapter, few initial assumptions with respect to the nature of the problem domains to be modelled are made. In contrast to the literature on diagnosis based on set theory (e.g. [Peng & Reggia, 1990] and [Josephson & Josephson, 1994]), a knowledge base need not necessarily consist of causal relations. The domain knowledge may also consist of empirical associations between defects and findings, it may be a detailed functional description of the working of a device or it may consist of other meaningful descriptions. An advantage of this less restrictive approach is that, where in other literature the commonalities between different approaches are blurred by emphasizing specific aspects of one approach, here common and distinctive features are more clearly revealed. We shall profit from these aspects of the framework when analysing the various formal theories of diagnosis in Chapter 4.

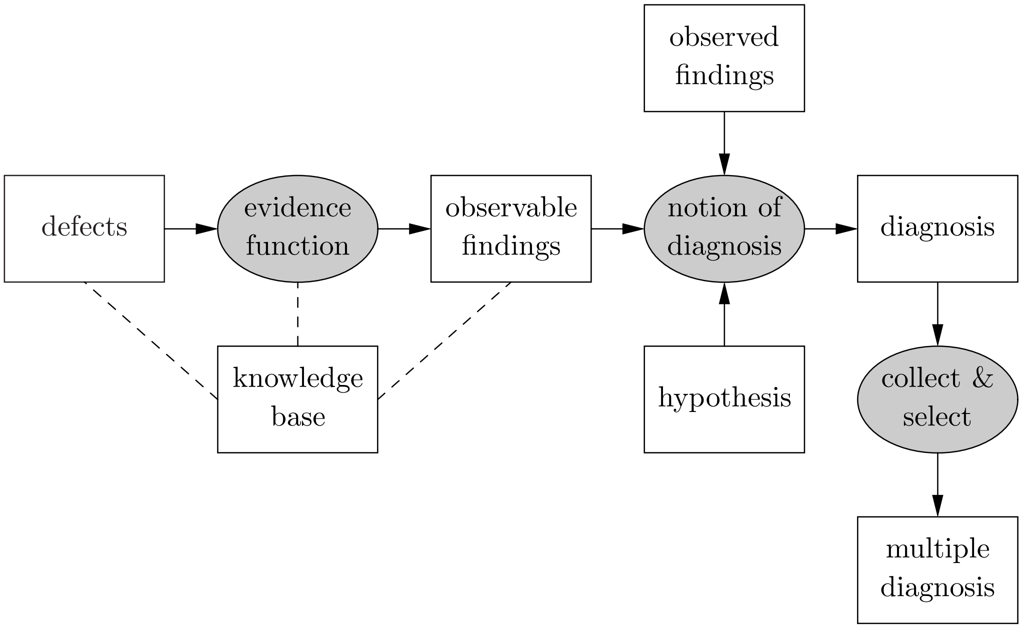


Figure 3.1: Schematic overview of the framework of diagnosis.

3.1.1 Defects and findings

Informally spoken, the notions of defect and finding constitute the basis of the framework from which all other notions in this section are constructed; together with two symbols denoting undefined situations, they form the diagnostic universe from which the terms for expressing a particular diagnostic problem must be selected.

Definition 3.1 (*diagnostic universe*). *A* diagnostic universe $\mathcal{U}$ *is a quadruple* $\mathcal{U} = (\mathcal{D}, \mathcal{F}, u, \perp)$*, where*

- $\mathcal{D}$ *is a set of elements, called* defects*,*
- $\mathcal{F}$ *is a set of elements, called* findings*.*

In addition, u *denotes a fixed set of elements, called the* undefined defects*, and* $\perp$ *denotes a fixed set of elements called the* undefined findings*;* u*,* $\perp$*,* $\mathcal{D}$ *and* $\mathcal{F}$ *are required to be mutually disjoint, with the exception of* $\mathcal{D}$ *and* $\mathcal{F}$*.*

In most diagnostic problems, the set of defects $\mathcal{D}$ and the set of findings $\mathcal{F}$ will be disjoint; however, as shall be discussed in Chapter 4, when representing the behaviour of a device, the dependence of the generated output on the entered input can only be represented by common elements from the sets $\mathcal{D}$ and $\mathcal{F}$. As shall become clear below, the sets $\mathcal{D}$ and $\mathcal{F}$ contain elements from problem domains, and the undefined defects and findings are used to express certain properties of the sets $\mathcal{D}$ and $\mathcal{F}$; hence, the symbols u and $\perp$ are meta-level symbols. From now on, it is assumed that the diagnostic universe $\mathcal{U}$ is fixed. In prospect, the undefined defects symbol u will be employed in Section 3.1.4 to denote that a diagnosis is undefined.

Defects are often denoted by the letter d and findings are denoted by the letter f, both possibly with a subscript; in examples the specific terminology of an example domain or problem is adopted.

Definition 3.2 (*diagnostic domain*)**.** *Let $\mathcal{U} = (\mathcal{D}, \mathcal{F}, u, \perp)$ be the diagnostic universe. A* diagnostic domain *Ω is a pair $\Omega = (\Delta, \Phi)$, where $\Delta \subseteq \mathcal{D}$ is a set of defects and $\Phi \subseteq \mathcal{F}$ is a set of findings.*

Usually, the sets Δ and Φ are assumed to be finite, because in the typical domains for which diagnostic knowledge-based systems are developed, the number of distinguished defects and findings is usually finite. The elements in the set of defects Δ and in the set of findings Φ are the building-blocks to describe the various elements involved in solving a specific problem of diagnosis. In the following, $\wp(\Delta)$ and $\wp(\Phi)$ will denote the power sets of Δ and Φ, respectively. The concept of diagnostic universe will return in Section 3.1.4, where it is used to express notions of diagnosis that hold for arbitrary diagnostic domains.

To be able of making a distinction between *present* and *absent* defects and findings, respectively, for representing a problem domain, a negation function is introduced. Positive and negative defects and findings are assumed to stand for the presence and absence of defects and findings, respectively, in reality. If a defect or finding is missing from a set of defects or findings, the defect or finding is assumed to be *unknown*.

Definition 3.3 (*negation*)**.** *Let $\Omega = (\Delta, \Phi)$ be a diagnostic domain. A* negation function *is a bijective function*

$$\neg : \Delta \cup \Phi \to \Delta \cup \Phi$$

such that the function composition $\neg \circ \neg = \iota$, where ι is the identity function. The set Δ is assumed to be partitioned into

- *Δ_P, the set of* positive *defects, and*
- *Δ_N, the set of* negative *defects*

such that $d \in \Delta_P$ iff $\neg(d) \in \Delta_N$. Similarly, the set of findings is assumed to be partitioned into

- *Φ_P, the set of* positive *findings, and*
- *Φ_N, the set of* negative *findings.*

such that $f \in \Phi_P$ iff $\neg(f) \in \Phi_N$.

We shall use the same symbol $\neg$ to refer to negation with respect to different diagnostic domains. In the following, the function value $\neg(x)$ will be abbreviated to $\neg x$. As a matter of convenience, members of Δ_N are frequently denoted by $\neg d$, such that $\neg(\neg d) = d \in \Delta_P$. Similarly, members of the set Φ_N are denoted by $\neg f$.

In the next definition, findings are associated with defects, yielding the notion of diagnostic specification, being the formal counterpart of a knowledge base in the framework.

Definition 3.4 (*diagnostic specification*)**.** *Let $\Omega = (\Delta, \Phi)$ be a diagnostic domain. A* diagnostic specification *Σ is a triple $\Sigma = (\Delta, \Phi, e)$, where e is a function*

$$e : \wp(\Delta) \to \wp(\Phi) \cup \{\perp\}$$

called an evidence function *with respect to* Ω*, for which the following hold:*

(1) *for each* $f \in \Phi$ *there exists a set* $D \subseteq \Delta$ *with* $f \in e(D)$ *or* $\neg f \in e(D)$ *(and possibly both);*

(2) *if* $d, \neg d \in D$ *then* $e(D) = \bot$*;*

(3) *if* $e(D) \neq \bot$ *and* $D' \subseteq D$ *then* $e(D') \neq \bot$*.*

If $e(D) \neq \bot$*, it is said that* $e(D)$ *is the set of* observable findings *for* D*.*

As has been touched upon above, an evidence function, and hence a diagnostic specification, is conceived as an *interpretation* of a knowledge base in terms of defects and findings for the purpose of diagnosis. How this function is to be interpreted for diagnosing a given problem must still be specified by a notion of diagnosis (cf. Section 3.1.4). The set $e(D)$ stands for the set of observable findings for a set D of simultaneously occurring (present or absent) defects. In terms of diagnostic problem solving, the set $e(D)$ consists of findings that may be interpreted in some way as 'evidence' for the occurrence of the set of defects D. Hence, the name 'evidence' function reflects the 'minimal' semantics that can be given to the function. A stricter meaning of the association between a set of defects and a set of findings is causality, i.e. the findings $F = e(D)$ are assumed to be caused by the combined occurrence of the defects in D. In medicine, a disorder or disease (a defect) is just a name for a collection of features that may be observed in the patient; then, the relationships between defects and observable findings need not be causal in nature. Hence, although the notion of evidence function has something in common with the notion of prediction of behaviour in other theories of diagnosis, it also captures other types of diagnostic knowledge. More about this will be said below.

Adopting the informal meaning of evidence functions as discussed above, the following can be observed. According to the definition above, we may have that both $f \in e(D)$ and $\neg f \in e(D)$, which simply means that these findings may alternatively occur given the combined occurrence of the defects in the set D. In some domains it might hold that if $e(\{d\}) = e(\{d'\})$, it follows that $d = d'$, i.e. the defects d and d' are taken as synonyms for the same defect. For example, if the defects stand for disorders in medicine, then two different names d and d' for which the equality holds, would normally be taken as different names for the same disorder. This situation is quite common in medicine. However, an evidence function is not assumed to be injective in general, because for non-singleton sets $D, D' \subseteq \Delta$, it is not precluded that $e(D) = e(D')$, although D and D' might be non-equivalent sets of defects. It is also not precluded that sets of defects may have several findings in common; thus, the sets $e(D)$ and $e(D')$, $D \neq D'$, need not be disjoint.

The undefined findings symbol $\bot$ is used to denote that the interpretation of a particular part of a knowledge base for the purpose of diagnosis is inconsistent. This may either be caused by conflicts among defects, or conflicts among findings.

Definition 3.5 (*consistency*). *Let* $\Sigma = (\Delta, \Phi, e)$ *be a diagnostic specification. If* $e(D) = \bot$*,* $D \subseteq \Delta$*, then the set of defects* D *is called* inconsistent*; otherwise,* D *is called* consistent.

The meaning of the empty set, $\varnothing$, of defects and findings is as usual, i.e. all defects and findings are unknown.

From the definition of the notions of evidence function and consistency it follows that inconsistency of the set of defects D may indicate that $d, \neg d \in D$. This is a form of inconsistency that is evident for syntactic reasons. However, it is also possible that D is inconsistent for other than syntactic reasons, for example, because D contains defects d and d' that are incompatible. In this situation, the inconsistency is a consequence of a semantic relationship between the defects d and d'. In several definitions, it will be convenient to consider only sets of defects that are consistent for syntactic reasons; hence, the following definition.

Definition 3.6 (*syntactic consistency*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, then the set of defects $D \subseteq \Delta$ is called* syntactically consistent *if for each defect $d \in D$: $\neg d \notin D$; otherwise, D is called* syntactically inconsistent.

In the following, the notion of maximal syntactic consistency will be employed to define particular evidence functions.

Definition 3.7 (*maximal syntactic consistency*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, then the set of defects $D \subseteq \Delta$ is called* maximally syntactically consistent *if D is syntactically consistent and there exists no $d \in \Delta$, $d \notin D$, such that $D \cup \{d\}$ is syntactically consistent.*

Sometimes, a knowledge base is only given or examined with respect to a subset of the entire set of defects Δ. For this purpose, the following definition is introduced.

Definition 3.8 (*restricted evidence function*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. A* restricted evidence function *of e with respect to the set $H \subseteq \Delta$, denoted by $e_{|H}$, is a function*

$$e_{|H} : \wp(H) \to \wp(\Phi) \cup \{\bot\}$$

such that for each $D \subseteq H$: $e_{|H}(D) = e(D)$.

From a general point of view, the expressive power of evidence functions is as large as infinite propositional logic; the function e may be viewed as similar to the conjunctive normal form of propositional formulae with defects and findings as literals. For example, the evidence-function representation of an implication $(d_1 \wedge d_2) \to (f_1 \vee f_2)$ would yield, among other function values, $e(\{d_1, d_2, \neg f_1\}) = \{f_2\}$. (Note that the argument $\{d_1, d_2, \neg f_1\}$ is allowed, because Δ and Φ need not be disjoint.) Hence, an evidence function is expressive enough to capture the sort of knowledge as represented in the logic theories of diagnosis, such as the abductive theory of diagnosis, discussed in the previous chapter. Consider the following example.

Example 3.1. In Figure 3.2, the graph representation of a logic specification of causal knowledge as, for example, employed in the abductive theory of diagnosis is depicted. The figure corresponds to the following logic specification:

$$d_1 \to d_2$$

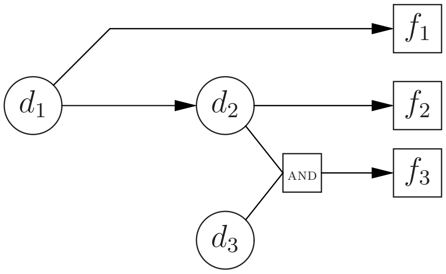


Figure 3.2: Causal net.

$$
\begin{aligned}
&d_1 \rightarrow f_1\\
&d_2 \rightarrow f_2\\
&d_2 \wedge d_3 \rightarrow f_3
\end{aligned}
$$

where d_1, d_2 stand for defects, and f_1, f_2 and f_3 are observable findings. Conceived as a specific medical application, the following meanings could be ascribed to the various elements:

$$
\begin{aligned}
d_1 &= \text{influenza}\\
d_2 &= \text{tracheobronchitis}\\
d_3 &= \text{asthma}
\end{aligned}
$$

$$
\begin{aligned}
f_1 &= \text{fever}\\
f_2 &= \text{sore throat}\\
f_3 &= \text{dyspnoea (shortness of breath)}
\end{aligned}
$$

In words: "influenza causes fever and infection of the trachea and bronchial tree, which causes sore throat, but if the patient suffers from asthma, dyspnoea will occur as well". Note that the specification incorporates causal knowledge, relating defect d_1 to defect d_2. Now, consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Delta_P = \{d_1, d_2, d_3\}$ and $\Phi_P = \{f_1, f_2, f_3\}$. The intended meaning of this specification with respect to diagnosis can be captured by means of the following evidence function e:

$$
e(D) = \begin{cases}
\{f_1, f_2\} & \text{if } D = \{d_1\}, \{d_1, d_2\}, \{d_1, \neg d_3\}, \{d_1, d_2, \neg d_3\}\\
\{f_2\} & \text{if } D = \{d_2\}, \{\neg d_1, d_2\}, \{d_2, \neg d_3\}, \{\neg d_1, d_2, \neg d_3\}\\
\{f_1, f_2, f_3\} & \text{if } D = \{d_1, d_3\}, \{d_1, d_2, d_3\}\\
\{f_2, f_3\} & \text{if } D = \{d_2, d_3\}, \{\neg d_1, d_2, d_3\}\\
\bot & \text{if } \{d_1, \neg d_2\} \subseteq D, \text{ or } D \text{ is syntactically inconsistent}\\
\varnothing & \text{otherwise}
\end{cases}
$$

For example, the function value $e(\{d_1\}) = \{f_1, f_2\}$ means that the findings f_1 and f_2 will be observed if the defect d_1 has occurred, and $e(\{d_1, d_3\}) = \{f_1, f_2, f_3\}$ means that the finding f_3 wil be observed by the combined presence of d_1 and d_3, in addition to the observable findings for d_1 and d_3 separately. The value $e(\{d_1, \neg d_2, d_3\}) = \bot$ indicates an impossible situation, because if d_1 is present, then d_2 cannot be absent (though, it may be unknown). Finally, if no findings at all will be observed, in spite of the presence or absence of certain defects, the function value of e is equal to the empty set. This is in accordance with the logic specification. For example, if only the defect d_3 is assumed to

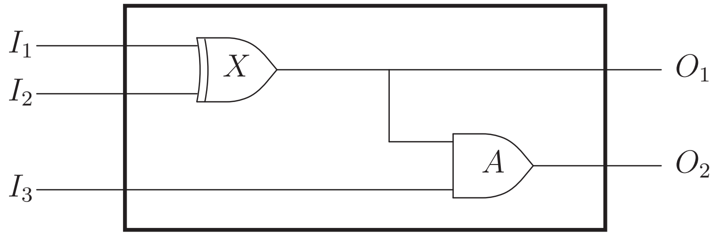


Figure 3.3: Logic circuit.

be present, no findings will be observed. The reader has probably noted that the evidence function above can be specified more tersely; in Section 3.1.3 techniques for the partial specification of evidence functions will be discussed in detail.

The evidence function e actually extends the logic specification above, by assuming that the specification is also intended to deal with negative defects. Another possibility would be to use only the restricted evidence function $e_{|\Delta_P}$ for translation of the logic specification. This would yield a considerable reduction in the specification, as well as preventing having to deal with inconsistent sets of defects. $\Diamond$

Basically, the evidence function e in the example above provides a description of defects in terms of findings that may be observed when the behaviour of a combination of defects D is studied in isolation, i.e. ignoring the findings produced by the defects $d \notin D$. The evidence function in the example captures a very specific type of causal knowledge; for this type of causality, it holds that when more defects are present, the same or more findings will be observed.

As has been discussed in Chapter 2, there are, in addition to causal knowledge, several other types of knowledge that can be incorporated in diagnostic system, e.g. knowledge of structure and normal or abnormal behaviour. This type of knowledge is usually employed for diagnosing device problems, where the behaviour of the device is observed by means of input and output signals. Since in many devices there are a fixed number of input and output channels, the number of observable findings associated with a set of defects is at least as large as the number of output channels. These assumptions are again reflected in the evidence function's structure. Of course, this only holds if all output channels are producing (not necessarily correct) output, irrespective of whether or not faulty components are present.

Example 3.2. Consider the logic circuit depicted in Figure 3.3. The circuit consists of an XOR (exclusive OR) gate X and an AND gate A. The presence of a defect in X is denoted by x; the absence of a defect in X is denoted by $\neg x$. A similar notation is employed to denote the presence or absence of a defect concerning gate A. The three inputs signals to the circuit are indicated by I_1, I_2 and I_3; O_1 and O_2 denote the two output signals. If $I_j = 1$, this will be denoted by i_j; an input equal to $I_j = 0$ will be denoted by $\neg i_j$. A similar convention is adopted for the output signals O_k. It is supposed that a defective gate produces an output signal that is complementary to the correct output signal. Suppose that the input signals to the circuit are $i_1, \neg i_2$ and i_3. Now, the output signals are represented as observable findings, and a component for which the presence or absence of a defect is unknown, is taken into account by assuming that the component is either defective or nondefective. Note that this description concerns both the structure as well as the normal and abnormal behaviour of the device. The following evidence function (only values for consistent sets of defects are provided) corresponds to the description above:

$$\begin{array}{rcl}
e(\{x, a\}) & = & \{\neg o_1, o_2\} \\
e(\{\neg x, a\}) & = & \{o_1, \neg o_2\} \\
e(\{x, \neg a\}) & = & \{\neg o_1, \neg o_2\} \\
e(\{\neg x, \neg a\}) & = & \{o_1, o_2\} \\
e(\{x\}) & = & \{\neg o_1, o_2, \neg o_2\} \\
e(\{\neg x\}) & = & \{o_1, o_2, \neg o_2\} \\
e(\{a\}) & = & \{o_1, \neg o_1, o_2, \neg o_2\} \\
 & = & e(\{\neg a\}) \\
 & = & e(\varnothing)
\end{array}$$

For example, $e(\{x\}) = \{\neg o_1, o_2, \neg o_2\}$ indicates that when the XOR gate is defective, and it is unknown whether or not the AND gate is defective, then the first output signal $O_1 = 0$ and the second output signal O_2 may be either 0 or 1, depending on whether the AND gate is defective or not. Hence, $e(\{x\})$ is defined with respect to the output of the entire circuit in Figure 3.3, not merely the output produced by the output channel directly connected to the XOR gate, i.e. O_1. For this circuit in general, the observable findings for $e(D)$ always include o_1, $\neg o_1$, or both, and o_2, $\neg o_2$, or both. In contrast with the assumptions underlying the evidence function given in Example 3.1, the behaviour of the system is described with respect to all elements of the entire system, and not in terms of isolated (defective) components. If, unlike the circuit in Figure 3.3, no interaction existed between a circuit's components, it is possible to define e in terms of observable findings associated with specific defective or nondefective components. ◊

The evidence function e in the example above expresses that when more defects are considered, i.e. taken to be either present or absent, the information contained in the set of observable findings will be more specific. The information for the model of normal and abnormal behaviour in the example above was represented by adopting fixed input signals, but this is not a fundamental limitation. By including the circuit inputs as extra defects into the set of defects Δ, it is possible to specify behaviour for any collection of inputs.

The two examples above were meant to convey some intuition concerning the expressive power of evidence functions for encoding knowledge that can be used for the purpose of diagnosis. One of the attractive features of evidence functions is that they provide an easy means for describing properties of diagnostic interpretations of knowledge bases in a precise, formal way. We are now ready to take a closer look at the notion of evidence function.

Above, we have adopted the view that an evidence function stands for the diagnostic interpretation of a knowledge base, i.e. an interpretation of a knowledge base in terms of defects and observable findings. This view, however, does not immediately provide us with a meaning that can be attached to the concept of evidence function. Now, let us adopt the idea that every knowledge base obtains a suitable meaning with respect to sets of cases with known collections of defects and findings. For example, a medical expert system would acquire its meaning relative to a large clinical database with data of patients from the same problem domain, or patients with particular disorders that have been encountered in the course of time. The same semantics seems appropriate as a basis for a formal semantics for evidence functions. Since the semantics is based on a universe of (idealized) historical cases, it is called a history-based semantics.

Definition 3.9 (*history-based semantics*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let $\mathcal{I}$ be a triple $\mathcal{I} = (\mathcal{C}, \delta, \varphi)$, called an* interpretation*, where*

- *C is a set of* cases*,*
- *δ is a surjective function*

  $$\delta : C \to \wp(\Delta)$$

  called a defects assignment function*, and*
- *φ is a function*

  $$\varphi : C \to \wp(\Phi)$$

  called a findings assignment function*,*

then the diagnostic specification Σ is called history-based *if for each consistent $D \subseteq \Delta$:*

(1) $\bigcup_{\substack{c \in C \\ \delta(c) = D}} \varphi(c) = e(D)$;

(2) $\forall c \in C$: $\varphi(c) \subseteq e(D) \Rightarrow e(\delta(c)) \subseteq e(D)$.

The first condition expresses that only the observed findings for each case in C known with (present and absent) defects D are relevant as observable findings in $e(D)$; other findings than these are considered irrelevant, and are discarded. This is typical for knowledge bases based on empirical associations, since these knowledge bases are based on experience obtained through handling many similar cases. The set of findings $\varphi(c)$ that has been observed for each case c consists of typical findings only. Cases that provide meaning to evidence functions that capture causal knowledge, or knowledge concerning normal or abnormal behaviour, will often have the same sets of observed findings for the same set of defects, i.e. if $\delta(c) = \delta(c'), c \neq c'$, then $\varphi(c) = \varphi(c')$. This, however, need not be true. For example, imagine a particular type of device that is known to produce various different abnormal behaviours for a given defect. Then, all findings that have been observed for devices with the given problem are gathered as the set of observable findings for the defect. The second condition expresses that if the observed findings associated with a case are included among the observable findings defined for a set of defects D, then the set of observable findings associated with the set of defects provided for the case must also be included among the set of observable findings defined for the set D. This condition ensures that the evidence function is sufficiently in agreement with the set of cases. The two conditions also ensure that the ordering among cases imposed by the findings assignment function is preserved by the evidence function.

Proposition 3.1. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification and let $\mathcal{I} = (C, \delta, \varphi)$ be a history-based interpretation for Σ. Then, for each $c, c' \in C$ it holds:*

$$\varphi(c) \subseteq \varphi(c') \Rightarrow e(\delta(c)) \subseteq e(\delta(c'))$$

Proof. From condition (1) of the definition of a history-based interpretation, it follows that $\varphi(c') \subseteq e(D)$ for $D = \delta(c')$, hence $\varphi(c) \subseteq e(D)$. Using condition (2) with $e(\delta(c')) = e(D)$, we get: $e(\delta(c)) \subseteq e(\delta(c'))$. ◊

This semantics encompasses both formalizations based on functional and causal models as well as approaches based on empirical associations, because in both approaches it is possible to collect case information from the past.

3.1.2 Properties of evidence functions

As has been argued above, an evidence function e may possess certain properties, determined by the (diagnostic) knowledge incorporated in the knowledge base on which it is based. In this section, an overview is provided of properties of evidence functions that will be useful for characterizing diagnostic knowledge. Some of these properties will be required in the analysis of the various formal theories of diagnosis in Chapter 4.

The various properties can be distinguished into *global* properties, i.e. properties that hold for the entire evidence function e, and *local* properties, i.e. properties that only hold for some sets of defects D.

Global properties

In descriptions of many problem domains, only positive findings, or positive findings and a few negative findings, are employed to characterize sets of defects. This situation has already been encountered in Example 3.1. By the definition of evidence function (cf. Definition 3.4), any finding that is included in the set of findings Φ, must appear, positively, negatively, or both, in some function value $e(D)$, $D \subseteq \Delta$. This explains why from Definition 3.4 it follows that $\bigcup_{D \subseteq \Delta, D\ consistent} e(D) = \Phi$ need not hold. Nevertheless, sometimes every positive *and* negative finding in Φ is covered by the evidence function e. The consequence is that such an evidence function is, in principle, dependent on the notion of diagnosis employed, capable of producing a diagnosis for any set of findings observed (cf. Section 3.1.4). It is convenient to introduce specific terminology for diagnostic specifications for which Φ is covered completely.

Definition 3.10 (*exhaustiveness*). *A diagnostic specification $\Sigma = (\Delta, \Phi, e)$ is called* exhaustive *if*

$$\bigcup_{\substack{D \subseteq \Delta \\ D\ consistent}} e(D) = \Phi$$

otherwise, it is called non-exhaustive.

Other global properties are based on set-theoretical relationships among the sets of defects and sets of observable findings associated with these sets of defects. We shall consider a number of such properties.

Monotonicity of the evidence function is a property that will be encountered several times in the analysis of theories of diagnosis in Chapter 4. It is defined as follows.

Definition 3.11 (*monotonicity*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. The evidence function e is called* monotonically increasing *if*

$$\forall D, D' \subseteq \Delta : D \subseteq D' \Rightarrow e(D) \subseteq e(D')$$

and e is called monotonically decreasing *if*

$$\forall D, D' \subseteq \Delta : D \subseteq D' \Rightarrow e(D) \supseteq e(D')$$

with D and D' consistent. If e is either monotonically increasing or decreasing, it is called monotonic*; otherwise, e is called* nonmonotonic*.*

If an evidence function is monotonically increasing, this means that the more defects are considered, the more (new) findings must be taken into account. The evidence function in Example 3.1, which was the result of the translation of causal knowledge into evidence-function representation, was monotonically increasing. If an evidence function is monotonically decreasing, this means that if more defects are considered, information concerning the observable findings of sets of defects will be more specific. We have encountered an example of such a function in Example 3.2, where knowledge concerning the normal and abnormal behaviour of a circuit was encoded. Note that what is often referred to as the 'nonmonotonicity of diagnosis' actually concerns the interpretation of observed findings in the process of diagnosis. This is an aspect completely different from the one considered here. Monotonicity in the present sense appears to be a feature underlying most current theories of diagnosis, although it is usually not explicitly mentioned (cf. Chapter 4).

Of special interest in the previous section was the representation of interactions among defects and findings in terms of an evidence function. If no interactions among defects and findings exist (except inconsistency among syntactically inconsistent defects), the evidence function conforms to the following definition.

Definition 3.12 (*interaction freeness*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. Then, the set of defects Δ is called* interaction free with respect to e *if*

$$e(D) = \bigcup_{d \in D} e(\{d\})$$

for each syntactically consistent set of defects $D \subseteq \Delta$. If in addition for each $d \in \Delta$: $e(\{d\})$ is nonempty, and for each $d, d' \in \Delta$, $d \neq d'$, it holds that

$$e(\{d\}) \cap e(\{d'\}) = \varnothing$$

the set Δ is called strongly interaction free*; otherwise, Δ is called* weakly interaction free*.*

We will sometimes simply say that the evidence function e is interaction free. Interaction freeness means that the observable findings associated with a collection of defects D are the same as the collected observable findings associated with each individual defect $d \in D$. Thus, by combining the observable findings for individual defects, the observable findings for combinations of defects are obtained. Although interaction freeness is presented here as a global property, we shall occasionally employ the phrase in a *local* sense, to express that two or more defects do not interact with each other, e.g. $e(\{d, d'\}) = e(\{d\}) \cup e(\{d'\})$. It is easy to show that an evidence function that is interaction free is also monotonically increasing.

Proposition 3.2. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, where Δ is interaction free, then e is monotonically increasing.*

Proof. Simply note that if $D \subseteq D'$, with consistent sets $D, D' \subseteq \Delta$, then

$$\begin{aligned} e(D') &= e(D \cup D') \\ &= \bigcup_{d \in D \cup D'} e(\{d\}) \\ &= \bigcup_{d \in D} e(\{d\}) \cup \bigcup_{d \in D'} e(\{d\}) \\ &= e(D) \cup e(D') \end{aligned}$$

From this, it follows that $e(D) \subseteq e(D')$. ◊

In [Bylander et al., 1992], interaction freeness is called ‘independence’; in this thesis, however, this term is reserved for notions of diagnosis to be discussed below. As a matter of convenience, function values $e(\{d\})$ of an evidence function that defines Δ to be interaction free, are sometimes simply denoted by $e(d)$. If a set of defects is strongly interaction free with respect to some evidence function e, this does not necessarily imply that the defects do not influence each other in one way or the other; it only means that these influences have not been captured in the function e explicitly, because the meaning attached to e does make these influences irrelevant with respect to diagnosis.

In some domains in which defects are interaction free, it holds that each defect is described in unique terms, i.e. for each defect $d \in \Delta$, the set of observable findings $e(d)$ is not contained in the set $e(D)$, if d is not included in D. It is shown that the evidence function restricted to consistent sets of defects is injective.

Proposition 3.3. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification such that Δ is interaction free with respect to e. Then, if for each $d \in \Delta$, and each consistent set $D \subseteq \Delta \backslash \{d\}$, it holds that $e(\{d\}) \not\subseteq e(D)$, then the restriction of the evidence functions e to consistent subsets of Δ is injective.*

Proof. It has to be proven that for consistent $D, D' \subseteq \Delta$, with $D \neq D'$, it holds that $e(D) \neq e(D')$. If $D \neq D'$, then there exists a defect $d \in D$ (or $d \in D'$ if $D \subset D'$, but reversing D and D' does not matter), such that $d \notin D'$. Hence, according to the assumption of the proposition: $e(d) \not\subseteq e(D')$. Since it holds by interaction freeness that $e(d) \subseteq e(D)$, it follows, also from interaction freeness, that $e(D) \not\subseteq e(D')$. From this, the result follows immediately. ◊

Given this proposition, the following corollary holds.

Corollary. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification such that Δ is strongly interaction free, then the restriction of the evidence function e to syntactically consistent sets of defects is injective.*

Proof. Simply note that if Δ is strongly interaction free, it holds that $e(\{d\}) \not\subseteq e(D \backslash \{d\})$ for each $D \subseteq \Delta$. ◊

Proposition 3.3 is also satisfied if for each $d \in \Delta$, $e(d)$ includes a unique observable finding (called a *pathognomonic* finding in medicine). Note that it is now possible to uniquely identify a set of defects D by its associated set of observable findings $F = e(D)$, due to the injective nature of e (but the set of defects may also be undefined). This yields a very simple form of diagnosis.

Local properties

There are a number of local properties of evidence functions that are the result of mapping a semantic relationship between (sets of) defects to relationships between sets of observable findings. A typical example of such a relationship is causality. For example, if the defect d is known to cause the defect d', it is, in terms of the associated evidence function, known that the set of observable findings for d contains all observable findings associated with d', i.e.

$$e(\{d'\}) \subseteq e(\{d\}) \tag{3.1}$$

In the abductive theory of diagnosis (cf. Section 2.2.2) the following would also hold:

$$e(\{d\}) = e(\{d, d'\}) \tag{3.2}$$

expressing that as d causes d', when d and d' are present together, precisely the same set of observable findings would be obtained as if only d was present and d' is unknown. From (3.1) and (3.2) it follows that

$$e(\{d, d'\}) = e(\{d\}) \cup e(\{d'\})$$

Hence, d and d' are assumed to be interaction free in the local sense; note that d and d' are only weakly interaction free. This appears to be a property of causality as employed in abductive diagnosis, but note that that interaction freeness will not hold in general (cf. Example 3.1).

There are various ways in which this result can be interpreted. In this thesis, it is assumed that we start with a specification of causal knowledge; an evidence function with the property given above will then be the result of translating this causal knowledge into uniform format. Another point of view might be to take causality as a notion which can be characterized in terms of observable findings. Then, the question arises whether properties (3.1) and (3.2) are necessary and sufficient conditions for the characterization of causality. This is a philosophical matter, which goes beyond the scope of this thesis; this issue shall not be pursued further.

Note that a causal specification $\mathcal{C}$ in the abductive theory of diagnosis with axiom set equal to

$$\mathcal{R} = \{d_1 \rightarrow d_2, d_2 \rightarrow f\}$$

is not distinguishable in terms of evidence functions from

$$\mathcal{R}' = \{d_1 \leftrightarrow d_2, d_2 \leftrightarrow f\}$$

because in both cases an interaction-free evidence function e with $e(\{d_i\}) = f$, $i = 1, 2$, results. This means that $\mathcal{R}$ and $\mathcal{R}'$ are similar with respect to their diagnostic interpretation. Note the correspondence to the predicate-completion interpretation of abductive diagnosis discussed in Section 2.2.2.

Starting with causality in a more general sense, a number of local properties of evidence functions will be examined.

(a) Influence interactions: the occurrence of some defects influences the occurrence of other defects, as reflected by the observable findings. The following two types of local interaction are distinguished:

- *Causality*: if the combination of defects D causes the set of findings F, then $F = e(D)$. The diagnostic view of knowledge of the sort 'the set of defects D causes the set of defects D'' as used in abductive diagnosis can be made precise in terms of the evidence function as follows:

  $$e(D') \subseteq e(D)$$

  for some consistent $D, D' \subseteq \Delta$, i.e. any finding that may be observed for the set of defects D' may also be observed for the set of defects D. Furthermore, it holds that

  $$e(D) = e(D \cup D')$$

  In Example 3.1 above, we discussed a simple causal relationship between two individual defects.

  Various other types of causal relations can be expressed in terms of evidence functions. For example, the values of the evidence function

  $$e(\{d_1\}) = e(\{\neg d_2\}) = e(\{\neg d_3\}) = \varnothing$$

  and

  $$e(\{d_2\}) = e(\{d_1, \neg d_3\}) = \{f_1\}$$
  $$e(\{d_3\}) = e(\{d_1, \neg d_2\}) = \{f_2\}$$

  express *nondeterministic causality* between the defect d_1 on the one hand, and d_2 and d_3, on the other hand, as depicted in Figure 3.4.

- *Correlation*: if the defects d and d', $d \neq d'$, are correlated, then if d has occurred then d' occurs as well, and vice versa, whereas if d is absent ($\neg d$), d' is also absent ($\neg d'$), and vice versa. Correlation of defects can be described by means of the evidence function as follows:

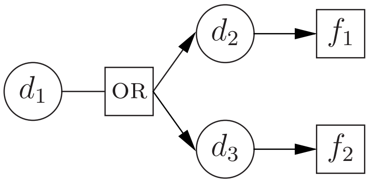


Figure 3.4: Nondeterministic causality.

$$\begin{aligned} e(\{d\}) &= e(\{d'\}) &&= e(\{d,d'\}) \\ e(\{\neg d\}) &= e(\{\neg d'\}) &&= e(\{\neg d,\neg d'\}) \end{aligned}$$

The conditions above are satisfied for positive correlation; negative correlation can be described by means of the condition

$$\begin{aligned} e(\{d\}) &= e(\{\neg d'\}) &&= e(\{d,\neg d'\}) \\ e(\{\neg d\}) &= e(\{d'\}) &&= e(\{\neg d,d'\}) \end{aligned}$$

(b) Synonymy: if the defects $d \in \Delta$ and $d' \in \Delta$ are synonymous, then $e(\{d\}) = e(\{d'\})$. This is commonly applied in medicine, as has been discussed above. If for each $d, d' \in \Delta$, $d \neq d'$, it holds that $e(\{d\}) \neq e(\{d'\})$, there are no synonymous defects. It is said that Δ (also e) is *synonym free*.

(c) Synergic interactions: these are interactions that augment, cancel, preclude, exclude, or complement local interactions among defects. The following types of interaction are distinguished:

- *Augmentation* (also referred to as *potentiation*): the combined occurrence of two or more defects in the set D gives rise to new observable findings in addition to those associated with the individual elements, or proper subsets of D, i.e.

  $$e(D) \supset \bigcup_{D' \subset D} e(D') \tag{3.3}$$

  for some consistent $D \subseteq \Delta$. It is interesting to note that (3.3) is yielded for monotonically increasing evidence functions, using the weaker condition:

  $$e(D) \not\subseteq \bigcup_{D' \subset D} e(D')$$

- *Cancellation* (also referred to as *fault masking* [Davis & Hamscher, 1988] or *antagonism*): the combined occurrence of two or more defects in the set D yields fewer observable finding when compared to the findings associated with the individual elements, or proper subsets of D, i.e.

  $$e(D) \subset \bigcup_{D' \subset D} e(D')$$

  for some consistent $D \subseteq \Delta$.

- *Augmented cancellation*: this notion combines the notions of augmentation and cancellation mentioned above, after weakening both conditions. The following holds:

  $$e(D) \not\subseteq \bigcup_{D' \subset D} e(D') \wedge e(D) \not\supseteq \bigcup_{D' \subset D} e(D')$$

for some consistent $D \subseteq \Delta$. For example, $e(\{d_1\}) = \{f_1\}$, $e(\{d_2\}) = \{f_2, f_3\}$, but $e(\{d_1, d_2\}) = \{f_3, f_4\}$; hence, the findings f_1 and f_2 are cancelled, and a new finding (f_4) is observable. Note that $e(\{d_1, d_2\}) \circ e(d_i)$, $i = 1, 2$, fails to hold for $\circ \in \{\subset, \supset\}$. This is a consequence of the dependence between augmentation and cancellation. The cancellation of findings causes augmentation to fail, and vice versa. Hence, the weakening of the two conditions in the notion of augmented cancellation.

- *Preclusion*: the presence of one or more defects in a combination implies that each element in some other combination of defects is assumed to be absent. This can be expressed by:

  $$e(\{d_1, \ldots, d_n\}) \supseteq e(\{\neg d'_1, \ldots, \neg d'_m\})$$

  This means that a set of present defects contains information pertaining to a set of absent defects. Note that if Δ is interaction free, it follows that

  $$e(\{d_1, \ldots, d_n\}) \supseteq e(\{\neg d'_i\})$$

  for each i, $1 \leq i \leq m$, $m \geq 1$. This yields a preclusion relation that is more easy to grasp, namely that a combination of defects D precludes some defect d:

  $$e(D) \supseteq e(\neg d)$$

- *Exclusion*: some combination of defects D cannot occur:

  $$e(D) = \bot$$

- *Complementation*: the observable findings associated with the absent defects $\neg d_1, \ldots, \neg d_n$, are the complements of those associated with the presence of those, i.e. if $e(\{d_1, \ldots, d_n\}) = \{f_1, \ldots, f_m\}$ then $e(\{\neg d_1, \ldots, \neg d_n\}) = \{\neg f_1, \ldots, \neg f_m\}$.

(d) Empirical associations: when the defects in the set D are simultaneously present, the findings in the set F may be observed, given $F = e(D)$. As mentioned above, knowledge based on empirical associations is often structured according to individual defects and families (categories) of defects. In the following definition, these concepts are formally introduced.

Definition 3.13 (*category*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. A defect $d \in \Delta$ is called* more specific *than a defect $d' \in \Delta$, denoted by $d \prec d'$, if $e(\{d\}) \subset e(\{d'\})$; d' is called* more general *than d. If $D \subseteq \Delta$ is a set of defects, such that there exists a defect $d' \in \Delta$ that is more general than d, for each $d \in D$, with*

$$e(\{d'\}) = \bigcup_{d \in D} e(\{d\})$$

then d' is called a category *for the defects $d \in D$.*

Hence, a category collects all findings of the defects with respect to which it is more general. The evidence-function representations of causal knowledge and of empirical associations have much in common, but there are a few differences. Firstly, the condition

$e(d) = e(d')$ fails to hold for empirical associations if d and d' are not synonymous. Secondly, a defect d for which $e(d) \supset e(d')$, for more than one defect $d' \in \Delta$, will be a category if the evidence function e stands for empirical associations, but, d will not be a category in general if e represents causal knowledge.

Example 3.3. Consider the disorders 'autoimmune chronic hepatitis' and 'acute hepatitis-A'. These two disorders are examples of hepatocellular disorders (i.e. disorders affecting the liver cells). Let us assume that these are the only two hepatocellular disorders known, and that 'hepatocellular disorder' is a category. From this we know that

$$e(\{\textit{autoimmune chronic hepatitis}\}) \subset e(\{\textit{hepatocellular disorder}\})$$

and

$$e(\{\textit{acute hepatitis-A}\}) \subset e(\{\textit{hepatocellular disorder}\})$$

hold. Using the notion of 'more specific defect' introduced above:

$$\textit{autoimmune chronic hepatitis} \prec \textit{hepatocellular disorder}$$

and

$$\textit{acute hepatitis-A} \prec \textit{hepatocellular disorder}$$

◊

Note that we have chosen to define a category in terms of all findings associated with a collection of more specific defects, instead of in terms of the common findings of those defects. According to the history-based semantics, for both choices all cases c provided with a category as diagnostic label could include all findings associated with the category, being all findings of more specific defects for the former choice and only common findings of more specific defects for the latter choice. The former choice seems more natural. Also observe that for a set of defects $\Delta_P = \{d_1, d_2, d_3\}$, with $d_1 \prec d_2$ and $d_3 \prec d_2$, given that $f \in e(\{d_1\})$ and $\neg f \in e(\{d_3\})$, it holds that $f, \neg f \in e(\{d_2\})$.

This concludes our list of possible interactions among defects, and their expression in terms of evidence functions. In the next chapter, the construction of evidence functions from specifications used in various formal theories of diagnosis is discussed in detail.

It should be noted that evidence-function values may not always provide explicit information about the *combinations* of findings that may be observed when a particular combination of defects occurs. For example, in Example 3.2 it is stated that the set of observable findings associated with a defective AND gate A is: $e(\{a\}) = \{o_1, \neg o_1, o_2, \neg o_2\}$. However, the function value does not make clear that if A is defective, either o_1 and $\neg o_2$, or $\neg o_1$ and o_2 may be observed, but $\neg o_1$ and $\neg o_2$ is an invalid possibility. This information can be expressed by means of an evidence function that is defined as follows:

$$e(\{a, o_1, \neg o_2\}) = e(\{a, \neg o_1, o_2\}) = \varnothing$$

and

$$e(\{a, \neg o_1, \neg o_2\}) = \bot$$

but this notation is not very perspicuous. When the evidence function is redefined as

$$e^s : \wp(\Delta) \to \wp(\wp(\Phi)) \cup \{\bot\}$$

such structural information can be represented as well, for example as follows: $e^s(\{a\}) = \{\{o_1, \neg o_2\}, \{\neg o_1, o_2\}\}$. We call such an evidence function e^s a *structural evidence function*; the original function is *unstructural*. Although it may be advantageous to have such structural information immediately available, it should be noted that the evidence function e in Example 3.2 does indeed contains this structural information. For example, the function value $e^s(\{a\})$ can be derived from $e(\{x, a\})$ and $e(\{\neg x, a\})$. For this reason, we shall only consider structural evidence functions when they offer some notational advantage. No fundamental increase in representational power is obtained by structural evidence functions.

3.1.3 Partial specification

When a domain satisfies certain properties, it may be sufficient to provide a partial specification of an evidence function. Partial specification has the advantage that it is not always necessary to explicitly specify, or compute, the exponential number of function values of the evidence function e; it suffices to provide only part of them explicitly. Any algorithm for diagnosis using an evidence function of the form discussed in the previous section, without simplifying assumptions, will be intractable. In [Bylander et al., 1992], in which the complexity of algorithms for abductive diagnosis is analysed, it is therefore assumed that the specification of an evidence function is polynomial in $|\Delta| + |\Phi|$. A *partial specification* of an evidence function e consists of a restriction of e, denoted by $\tilde{e}$, which is defined on a nonempty subset $V \subseteq \wp(\Delta)$, together with a number of computation rules expressing how function values $e(D)$ must be determined. If an evidence function is defined by means of a partial specification, it is called *partially specified*.

In domains for which not all function values $e(D)$ can be provided explicitly, such as in medicine, the condition that the specification of an evidence function is polynomial in size is usually fulfilled, be it for pragmatic reasons. In biomedical applications there is usually insufficient knowledge available to explicitly capture all interactions among defects, because the medical literature provides little information about the observable features of specific disorder combinations. In technical applications, the situation is less unfavourable, in the sense that often precise technical descriptions of the domain are available.

In several diagnostic theories, for example the set-covering theory of diagnosis [Peng & Reggia, 1990] (cf. Section 2.2.3), the partial specification includes a restriction of an evidence function to singleton sets, i.e. it suffices to define an evidence function in terms of the individual defects distinguished in the domain. If the associated computation rule expresses that the observable findings for non-singleton sets of defects can be taken as the union of the observable findings associated with their elements, the evidence function is interaction free. This limitation is enforced by some formal theories of diagnosis; it may

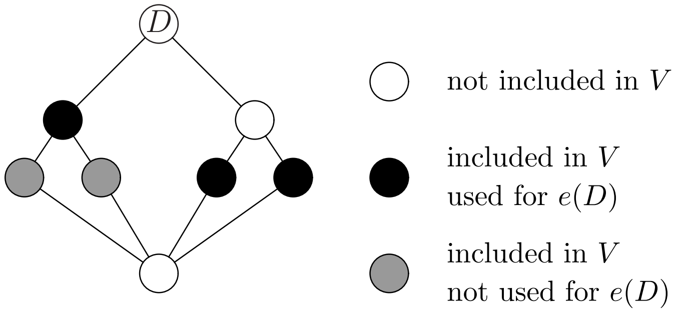


Figure 3.5: Part of a lattice used for bottom-up specification of an evidence function.

not be sanctioned by the characteristics of every problem domain, as we have seen in the previous section.

Although the extension of a partial specification to an evidence function is thus dependent on known evidence-function properties, expressed by means of computation rules, there are two extremes that deserve attention. The first useful way of partially specifying an evidence function is based on the assumption that when no explicit knowledge concerning the findings associated with a set of defects D is available, implicitly the largest proper subsets D' of D for which $\tilde{e}(D')$ is given, are taken to yield sufficient information concerning the interactions among the elements of D. This form of partial specification is called bottom-up partial specification.

Definition 3.14 (*bottom-up partial specification*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let $V \subseteq \wp(\Delta)\backslash\{\varnothing\}$ be a set, such that for each $d \in \Delta$: $\{d\} \in V$. Then, the function*

$$\tilde{e} : V \to \wp(\Phi) \cup \{\bot\}$$

is called a bottom-up partial specification *of e if:*

(1) *for each $D \in V$: $e(D) = \tilde{e}(D)$;*

(2) *for each $D \in \wp(\Delta)\backslash V$:*

$$e(D) = \bigcup_{\substack{D' \subset D,\, D' \in V \\ \forall D'' \in V,\, D'' \subset D \,:\, D'' \not\supset D'}} e(D')$$

Hence, by a bottom-up partial specification $\tilde{e}$ we mean a restriction of an evidence function e with appropriate computation rules to generate the function e from $\tilde{e}$. The principal idea is illustrated in Figure 3.5. Note that a restriction $\tilde{e}$ need not be unique; one can freely include subsets D of Δ in the domain of the restriction $\tilde{e}$ for which $e(D)$ could also be determined using condition (2) in the definition above. The intuitive idea of a bottom-up partial specification is that information concerning the interaction among defects is derived from the largest (with respect to $\subset$) proper subsets D' of a set of defects D, for which function values $\tilde{e}(D')$ have explicitly been given; the function value $e(D)$, when not explicitly given by $\tilde{e}$, is obtained as the union of all such $\tilde{e}(D')$. In the examples below this choice will be further clarified. For convenience, in the following, function values for syntactically inconsistent sets will be left out from the definition of bottom-up partial specifications $\tilde{e}$. From the definition of a bottom-up partial specification it follows that $e(\varnothing) = \varnothing$, i.e. there are no observable findings if there is no knowledge concerning defects. If the problem domain concerns the (faulty) behaviour of a device, a bottom-up partial specification amounts to specifying the isolated behaviours of parts of the device. Hence, a bottom-up partial specification is in line with a specification of causal knowledge as in the abductive theory of diagnosis by Console and Torasso, i.e. any diagnostic specification obtained from this theory can be described as a bottom-up partial specification (See Chapter 4).

Example 3.4. Reconsider Figure 3.2 and Example 3.1. The following bottom-up partial specification $\tilde{e}$ corresponds to the evidence function e defined in the example:

$$\tilde{e}(D) = \begin{cases} \{f_1, f_2\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2\} \\ \{f_1, f_2, f_3\} & \text{if } D = \{d_1, d_3\} \\ \{f_2, f_3\} & \text{if } D = \{d_2, d_3\} \\ \bot & \text{if } D = \{d_1, \neg d_2\}, \{d_1, \neg d_2, d_3\}, \{d_1, \neg d_2, \neg d_3\} \\ \varnothing & \text{if } D = \{d_3\},\ D = \{\neg d_i\}, i = 1, \ldots, 3 \end{cases}$$

For example,

$$\begin{aligned} e(\{d_1, d_2, d_3\}) &= \tilde{e}(\{d_1, d_3\}) \cup \tilde{e}(\{d_2, d_3\}) \\ &= \{f_1, f_2, f_3\} \end{aligned}$$

i.e. the findings associated with the combination of disorders (defects) influenza (d_1), tracheobronchitis (d_2) and asthma (d_3) are fever (f_1), sore throat (f_2) and dyspnoea (f_3). Note that the function values $e(\{d_1, d_2\})$, $e(\{d_i\})$, $i = 1, 2, 3$, are ignored in computing $e(\{d_1, d_2, d_3\})$ (although their inclusion would not have yielded another function value $e(\{d_1, d_2, d_3\})$).

Now, suppose that the following logical implication is added to the implications given in Example 3.1:

$$d_3 \rightarrow \neg f_2$$

i.e. when asthma is present, there can be no sore throat. Then, the evidence function must be redefined by including, for example, $e(\{d_2, d_3\}) = \bot$ as a function value. It is not possible to express such knowledge in the abductive theory of diagnosis by Console and Torasso, due to the restriction to (general) Horn formulae; the example, though, is in line with this theory. However, the exclusion of combinations of defects can be expressed in the abductive theory of diagnosis by logical integrity constraints (cf. [Console et al., 1991]). ◊

Hence, a bottom-up partial specification permits the encoding of causal knowledge in the sense of abductive diagnosis. However, bottom-up partial specifications also allow for representing nonmonotonic interactions and complementary findings representing alternative observable findings, e.g. f and $\neg f$, thus extending the repertoire of the types of

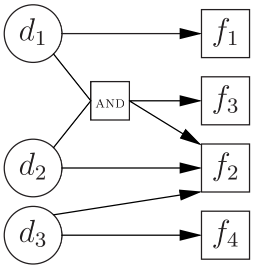


Figure 3.6: Partial evidence function $\tilde{e}$.

knowledge that can be used for diagnosis.

Example 3.5. Consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$ with $\Delta_P = \{d_1, d_2, d_3\}$, $\Phi_P = \{f_1, f_2, f_3, f_4\}$, where e is bottom-up partially specified by means of the function $\tilde{e}$, which is defined as follows:

$$\tilde{e}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2\} \\ \{f_2, f_4\} & \text{if } D = \{d_3\} \\ \{f_2, f_3\} & \text{if } D = \{d_1, d_2\} \\ \varnothing & \text{if } D = \{\neg d_i\},\ i = 1, \ldots, 3 \end{cases}$$

From this specification, it follows that $e(\{d_1\}) \not\subseteq e(\{d_1, d_2\})$; e is nonmonotonic. The corresponding diagram is shown in Figure 3.6. Note the difference between this diagram, which is a representation of the function $\tilde{e}$, and is not a causal graph as in the abductive theory of diagnosis, and the evidence-function interpretation of Figure 3.2. If we would interpret the diagram as in the abductive theory of diagnosis, we would have $e(\{d_1, d_2\}) = \{f_1, f_2, f_3\}$ which is not the case here.

In the abductive theory of diagnosis nonmonotonic interactions between defects have been expressed by negation as finite failure, yielding the following logical specification for the interaction between d_1 and d_2 ($\sim$ denotes negation by finite failure):

$$\begin{aligned} d_1 \wedge d_2 &\rightarrow f_2 \wedge f_3 \\ d_1 \wedge \sim d_2 &\rightarrow f_1 \\ \sim d_1 \wedge d_2 &\rightarrow f_2 \end{aligned}$$

Here, for example, it is stated that when d_1 is present and proving d_2 to be present has failed, which is therefore assumed to be absent, f_1 is observable, whereas the evidence-function value $e(\{d_1\}) = \{f_1\}$ expresses that d_2 is unknown. Hence, the meaning of the two specifications differs. ◊

As a prerequisite for bottom-up partial specification, it is assumed that at least knowledge concerning individual defects (i.e. singleton sets of defects) is available in a given diagnostic domain. This is not an unrealistic assumption, because in many problem domains knowledge concerning the possible abnormal behaviour resulting from an individual defect is the kind of knowledge most readily available. However, as shall be discussed below, there are other domains where strong interactions between defects exist, for which this assumption does not hold.

Note that an evidence function e that has been partially specified is, by convention, still a total function. If that prerequisite does not hold, the evidence function will be a partial function, for example, because there are defects in the domain for which no knowledge concerning their occurrence in isolation is available. In many theories of diagnosis it is then assumed that the set of observable findings associated with a set of defects is empty (cf. Chapter 4).

Example 3.6. Consider again the evidence function from the example above (Example 3.5). From this partial specification it follows that, for example, $e(\Delta_P) = \tilde{e}(\{d_1, d_2\}) \cup \tilde{e}(\{d_3\}) = \{f_2, f_3, f_4\}$. Note that neither $\tilde{e}(\{d_1\})$ nor $\tilde{e}(\{d_2\})$ play a role in determining $e(\Delta_P)$, because there is information available about the interaction between the defects d_1 and d_2 by the function value $\tilde{e}(\{d_1, d_2\})$. This function value provides partial information about the mutual influences among the defects in Δ_P; more precise information about the possible interactions between the members of Δ_P is unavailable; hence, $\{d_1, d_2\}$ and $\{d_3\}$ are assumed to be free of interaction, but the defects d_1 and d_2 are not. ◊

It follows that a bottom-up partial specification may provide information about the interaction between defects. In the extreme situation that no interaction between defects exists, it suffices to define a partial specification in terms of individual defects only.

Proposition 3.4. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, such that Δ is interaction free. Then, there exists a bottom-up partial specification $\tilde{e}$ of e with domain $V = \{\{d\} \mid d \in \Delta\}$.*

Proof. Note that if the domain of $\tilde{e}$, V, is defined as above, conditions (1) and (2) in Definition 3.14 simplify to the definition of interaction freeness; hence, the evidence function can be defined as follows

$$e(D) = \bigcup_{d \in D} \tilde{e}(\{d\})$$

for each syntactically consistent set $D \subseteq \Delta$. ◊

The basic idea of using bottom-up partial specifications for defining evidence functions was that in many domains knowledge concerning individual defects is the kind of knowledge readily available. An interesting question is whether or not it is possible to define an evidence function in terms of maximally consistent sets of defects instead of individual defects, where the same evidence function can also be defined using a bottom-up partial specification. Some additional concepts are needed to explore this possibility.

Definition 3.15 (*common findings assumption*)**.** *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. If for the evidence function e it holds that*

$$e(D) \supseteq \bigcap_{\substack{D' \supset D \\ D' \text{ consistent}}} e(D')$$

for each nonempty, consistent set $D \subseteq \Delta$, that is not maximally syntactically consistent, then it is said that e satisfies the common findings assumption.

The common findings assumption states that the common findings of combinations of defects D' that are supersets of another combination of defects, D, must be included among the observable findings of D.

It appears that an evidence function can indeed be defined in terms of sets of findings associated with maximally consistent sets of defects $D \subseteq \Delta$, but only when the evidence function is monotonically increasing and the common findings assumption is satisfied. The following, straightforward, but important, proposition states this result more formally.

Proposition 3.5. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification with monotonically increasing evidence function e for which the common findings assumption is satisfied, and for which $e(\varnothing) = \varnothing$. Let $\tilde{e} : V \to \wp(\Phi) \cup \{\perp\}$ be the restriction of e to V, where*

$$V = \{D \subseteq \Delta \mid D \textit{ is maximally syntactically consistent}\}$$

Then, for each nonempty consistent set $D \subseteq \Delta$, such that D is not maximally syntactically consistent:

$$e(D) = \bigcap_{\substack{D' \supset D,\, D' \in V \\ D' \textit{ consistent}}} \tilde{e}(D') \tag{3.4}$$

Proof. From the fact that e is monotonically increasing and that the common findings assumption holds, it follows that

$$e(D) = \bigcap_{\substack{D' \supset D \\ D' \textit{ consistent}}} e(D') \tag{3.5}$$

for each nonempty, consistent set $D \subseteq \Delta$, that is not maximally syntactically consistent. The remainder of the proof is by backward induction on the size of the set of defects D.

Basis: Let $n = |D|$, such that $D \cup \{d\}$ is a maximally consistent subset of Δ for some $d \in \Delta$, with $d \notin D$, then equality (3.4) holds by equality (3.5). (Note that $n = |\Delta_P| - 1$.)

Induction hypothesis: Assume that equation (3.4) holds for any set of defects $D \subseteq \Delta$, with $|D| = i$, $i = n, \ldots, k$, where $k \leq n$.

Induction step: Let $|D| = k - 1$, then

$$\begin{aligned}
e(D) &= \bigcap_{\substack{D' \supset D \\ D' \textit{ consistent}}} e(D') \\
&= \bigcap_{\substack{D' \supset D,\, |D'| = k \\ D' \textit{ consistent}}} e(D') && (e \text{ is monotonically increasing}) \\
&= \bigcap_{\substack{D' \supset D,\, |D'| = k \\ D' \textit{ consistent}}} \left[\bigcap_{\substack{D'' \supset D',\, D'' \in V \\ D'' \textit{ consistent}}} \tilde{e}(D'') \right] && (\text{induction hypothesis})
\end{aligned}$$

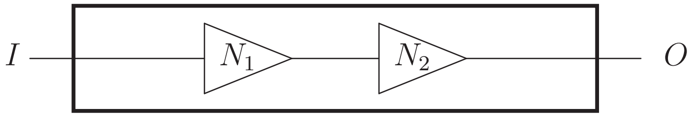


Figure 3.7: Two NOT gates in series.

$$= \bigcap_{\substack{D'' \supset D, D'' \in V \\ D'' \text{ consistent}}} \tilde{e}(D'') \qquad \text{(transitivity of } \supset\text{)}$$

The last step in the proof uses the fact that any chain $D_1 \supset \cdots \supset D_n$ from $D_1 = D''$ to $D_n = D$ contains a set D_i with $|D_i| = k$. Hence, the condition that $|D'| = k$ does not constrain the sets D'' taken into account in computing the intersection of $\tilde{e}(D'')$. This concludes the proof. ♢

Thus, it turns out that the function $\tilde{e}$ is a partial specification under the mentioned conditions of the evidence function e. It provides a characterization of e that differs from a bottom-up partial specification. However, from the condition $e(\varnothing) = \varnothing$, it follows that an evidence function as defined in the proposition above can also be defined using a bottom-up partial specification. The function values $e(D)$ for nonmaximal consistent sets can be simply computed. Note that the partial specification will include exactly 2^n function values, with $n = |\Delta_P|$. The proposition above holds a fortiori for sets of defects Δ that are interaction free with respect to an evidence function e. If an evidence function is monotonically increasing, but there are common findings of supersets D' of a set of defects D that are not included in $e(D)$, the proposition is not satisfied.

The second typical form of a partial specification of an evidence function is obtained by providing at least explicit function values for maximally syntactically consistent sets $D \subset \Delta$, and describing other combinations of defects D' by taking associated observable findings of defects $d \notin D$ into account. In the following example, this particular partial specification technique is introduced, using a diagnostic description of a logic circuit taken from [De Kleer et al., 1992].

Example 3.7. Consider the logic circuit depicted in Figure 3.7, which consists of two NOT gates (or inverters) in series. In [De Kleer et al., 1992], the problem of diagnosing faulty behaviour of the given logic circuit is described for an input signal fixed to $I = 0$, denoted here by $\neg i$, with resulting output signals equal to $O = 0$, denoted by $\neg o$, or $O = 1$, denoted by o, respectively. Again, output signals correspond to observable findings. The following behavioural assumptions are made in [De Kleer et al., 1992]. If a NOT gate N_i is defective, denoted by n_i, it will be either 0, or the input to the gate is shorted (unmodified) to its output; $\neg n_i$ designates that the NOT gate N_i is not defective. Given this information, the following restriction $\tilde{e}$ of e can be defined (we have disregarded the input, because it is assumed to be fixed):

$$\begin{aligned}
\tilde{e}(\{n_1, n_2\}) &= \{\neg o\} \\
\tilde{e}(\{\neg n_1, n_2\}) &= \{\neg o, o\} \\
\tilde{e}(\{n_1, \neg n_2\}) &= \{o\}
\end{aligned}$$

$$\tilde{e}(\{\neg n_1, \neg n_2\}) \;=\; \{\neg o\}$$

The complementary pair $\{\neg o, o\}$ is the result of the assumption above that there are two different, nondeterministic types of abnormal behaviour. The function $\tilde{e}$ is taken as a partial specification to generate e by assuming that $e(\{n_1\}) = \{\neg o, o\}$, etcetera, meaning that if it is unknown whether or not n_2 is defective, the possible output of the circuit, given n_1 to be defective, is $\{\neg o, o\}$. Thus, similar to Example 3.2, we have that $e(\{n_1\}) = e(\{n_1, n_2\}) \cup e(\{n_1, \neg n_2\})$. Interestingly, this partial specification indicates that if the observed output signal is equal to o, either $\{\neg n_1, n_2\}$ or $\{n_1, \neg n_2\}$ may be the case, which are precisely the diagnostic alternatives provided by De Kleer et al. However, it is not at all obvious from their example that for an output equal to $\neg o$, the set of defects $\{\neg n_1, n_2\}$ is a possibility as well. This information is immediately available from the evidence function e. ♢

This way of partially specifying an evidence function will be called top-down partial specification of an evidence function. A top-down partial specification is appropriate when it is not possible to describe defects with associated observable findings in isolation from other defects and associated findings, i.e. knowledge of the associated findings of the other defects, including their interaction, is needed to describe the defects. If the domain is a device, this assumption means that it is not possible to describe the (normal or abnormal) behaviour of a component in isolation from its environment. One could view the approach supported by top-down specification as a 'holistic approach', and the approach supported by bottom-up specification as a 'reductionistic approach'. Top-down partial specification is defined below.

Definition 3.16 (*top-down partial specification*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let $V \subseteq \wp(\Delta)\backslash\{\varnothing\}$ be a set, such that for each maximally syntactically consistent set $D \subseteq \Delta$: $D \in V$. Then, the function*

$$\tilde{e} : V \to \wp(\Phi) \cup \{\bot\}$$

is called a top-down partial specification *of e if:*

(1) *for each $D \in V$: $e(D) = \tilde{e}(D)$;*

(2) *for each $D \in \wp(\Delta)\backslash V$:*

$$e(D) = \bigcup_{\substack{D' \supset D,\, D' \text{ consistent},\, D' \in V \\ \forall D'' \in V, D'' \supset D : D'' \not\subset D'}} e(D')$$

Note that $e(D)$ is obtained by taking the union of all function values $e(D')$, where $D' \in V$ is a minimal proper superset of D, and no set $D'' \in V$ is smaller than D'. The principal idea is depicted in Figure 3.8. In Example 3.2 and Example 3.7, the behaviour of two logic circuits was studied using evidence functions that could have been generated by a top-down partial specification $\tilde{e}$, with

$$V = \{\{a, x\}, \{\neg a, x\}, \{a, \neg x\}, \{\neg a, \neg x\}\}$$

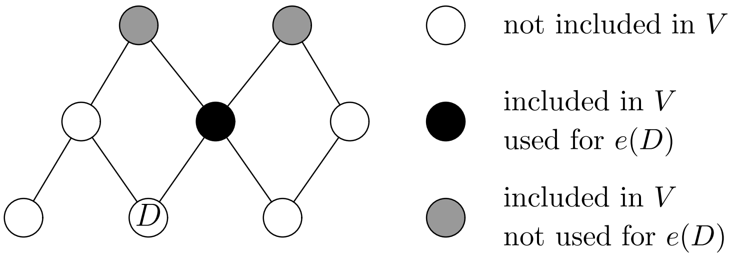


Figure 3.8: Part of a lattice used for top-down specification of an evidence function.

for Example 3.2 and

$$V = \{\{n_1, n_2\}, \{\neg n_1, n_2\}, \{n_1, \neg n_2\}, \{\neg n_1, \neg n_2\}\}$$

for Example 3.7. The assumption underlying an evidence function defined in this way is that it is sufficient to describe a domain in terms of the observable findings associated with all maximally consistent combinations of defects in the domain. This means that if the domain is a device consisting of components that may be defective, information about the isolated behaviour of individual components of the system has not been supplied. If a set of defects is described in terms of this special case of a top-down partial specification, we shall say that it is externally described.

Definition 3.17 (*externally described*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. The set of defects Δ is called* externally described *with respect to e if there exists a top-down partial specification $\tilde{e}$ for e with domain V, where for each $D \in V$: D is maximally syntactically consistent.*

Note that if Δ is externally described with respect to e, the definition of the evidence function can be simplified as follows. For each consistent $D \subseteq \Delta$:

$$e(D) = \bigcup_{\substack{D' \supseteq D, D' \in V \\ D' \text{ consistent}}} \tilde{e}(D')$$

It is easily shown that an evidence function for a set of defects that is externally described is monotonically decreasing.

Proposition 3.6. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, where Δ is externally described, then e is monotonically decreasing.*

Proof. If $D \subseteq D'$, with consistent $D, D' \subseteq \Delta$, then

$$\begin{aligned} e(D') &= e(D \cup D') \\ &= \bigcup_{\substack{D'' \supseteq (D \cup D'), D'' \in V \\ D'' \text{ consistent}}} e(D'') \\ &\subseteq \bigcup_{\substack{D' \supseteq D, D' \in V \\ D' \text{ consistent}}} e(D') \\ &= e(D) \end{aligned}$$

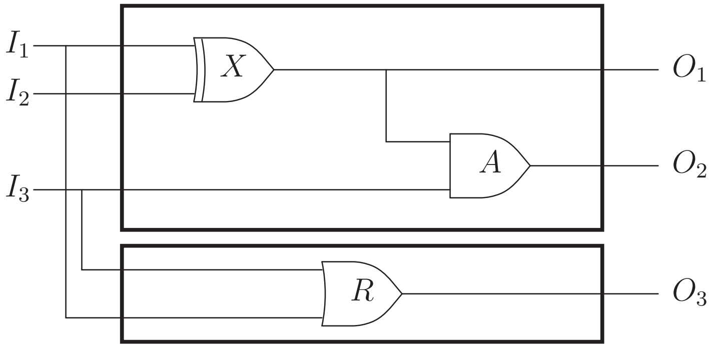


Figure 3.9: Logic circuit consisting of two separate modules.

From this, it follows that e is monotonically decreasing. $\Diamond$

Observe that top-down partial specification does not result in a significant reduction of the number of values to specified for the evidence function, because if $|\Delta_P| = n$, at least 2^n function values have to be specified.

Above, we have introduced two opposite ways to define evidence functions. Bottom-up partial specification appeared to be particularly suitable for generating evidence functions for defects between which a limited amount of interaction exists. By contrast, top-down partial specification is most suitable for generating evidence functions for defects which are strongly interrelated. One would expect that there are also evidence functions that lie somewhere between these two extremes, suitable for representing particular real-world knowledge; in the following example, we discuss a logic circuit that does.

Example 3.8. Consider the logic circuit depicted in Figure 3.9. The circuit consists of two separate modules. The first module contains an XOR gate, denoted by X, and an AND gate, denoted by A. The second module merely contains an OR gate, denoted by R. Suppose that a defective component reverses the output normally expected. Presence or absence of defects is denoted as in Example 3.2. Information concerning the behaviour of the two modules is separately available, i.e. there is no interaction between the two modules. As in Example 3.2, it is assumed that the input signals to the circuit are $i_1, \neg i_2$ and i_3. There are three output signals in the present case. The evidence function that captures the behaviour of the circuit is defined using a top-down partial specification $\tilde{e}$, that, however, has some characteristics of a bottom-up partial specification. The behaviour of the first module, consisting of an XOR and an AND gate, is described by means of the following function values for the restriction $\tilde{e}$:

$$\begin{array}{rcl}
\tilde{e}(\{x, a\}) & = & \{\neg o_1, o_2\} \\
\tilde{e}(\{\neg x, a\}) & = & \{o_1, \neg o_2\} \\
\tilde{e}(\{x, \neg a\}) & = & \{\neg o_1, \neg o_2\} \\
\tilde{e}(\{\neg x, \neg a\}) & = & \{o_1, o_2\}
\end{array}$$

For example, $e(\{x\}) = \{\neg o_1, \neg o_2, o_2\}$. The behaviour of the second module, which consists of an OR gate R only, is described by means of the following function values:

$$\tilde{e}(\{r\}) \quad = \quad \{\neg o_3\}$$

$$\tilde{e}(\{\neg r\}) \;=\; \{o_3\}$$

Now, in order to describe the observable findings (output) for the entire logic circuit, we take into account that the two modules do not interact. As a consequence, for example

$$\tilde{e}(\{x, a, r\}) = \tilde{e}(\{x, a\}) \cup \tilde{e}(\{r\})$$

and

$$\tilde{e}(\{x, a, \neg r\}) = \tilde{e}(\{x, a\}) \cup \tilde{e}(\{\neg r\})$$

hold. Hence, these values differ from those that would have been obtained for an externally described set of defects Δ. The structure of the restriction $\tilde{e}$ makes clear that there are two parts in the system, where each part is suitable for external description. This is exactly the real-world situation for this circuit. ◊

Bottom-up and top-down partial specification represent two variable instances of a spectrum of possible characterizations of diagnostic knowledge, with externally described and interaction-free defects as the two extremes. Bottom-up partial specification is especially suitable if little interaction among defects exists. Although top-down partial specification is employed when complex interactions between defects are involved, precise information concerning the nature of the interactions between sets of defects is embedded in the evidence function defined by the specification. For this reason, it may be worthwhile to identify to which extent defects are interaction free. This can be accomplished by transforming an evidence function generated by a top-down partial specification to an evidence function that can be generated by a bottom-up partial specification. The resulting evidence function may then be investigated with respect to interaction freeness. It is not always possible to determine the exact nature of interactions among defects by means of this transformation, simply because this information may have been lost due to the assumption that defects are interaction free.

The general transformation scheme is obtained by removing information concerning unknown defects from a function value $e(D)$, where $e(D)$ is generated by a top-down partial specification (cf. Definition 3.16). The observable findings that are removed are simply the observable findings associated with the smallest supersets D' of a set of defects D for which $\tilde{e}(D')$ is defined, with the exception of the observable findings these sets of defects have in common. In the following definition, the transformation is formally introduced.

Definition 3.18 (*top-to-bottom transformation*)**.** *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, where e is defined by means of the top-down partial specification $\tilde{e}$ with domain V. Let e' be an evidence function, obtained from the evidence function e as follows:*

(1) *For $D = \varnothing$: $e'(D) = \varnothing$;*

(2) *For each nonempty $D \subseteq \Delta$, with $e(D) \neq \bot$:*

$$e'(D) \;=\; e(D) \backslash$$

$$\left[\bigcup_{\substack{D' \supset D, D' \mathit{\ consistent}, D' \in V\\ \forall D'' \in V, D'' \supset D \,:\, D'' \not\subset D'}} e(D') \setminus \bigcap_{\substack{D' \supset D, D' \mathit{\ consistent}, D' \in V\\ \forall D'' \in V, D'' \supset D \,:\, D'' \not\subset D'}} e(D')\right]$$

(3) *For each nonempty* $D \subseteq \Delta$ *with* $e(D) = \bot$*:* $e'(D) = \bot$.

The resulting evidence function e' can be defined by a bottom-up partial specification $\tilde{e}'$, because the requirement $e'(\varnothing) = \varnothing$ holds. In order to keep the number of sets covered by the restriction $\tilde{e}'$ of a bottom-up partial specification as small as possible, a restriction $\tilde{e}'$ to generate e' could be devised iteratively. Starting with the singleton sets $\{d\}, d \in \Delta$, as members of V, and adding a set D if:

$$e'(D) \neq \bigcup_{\substack{D' \subset D, D' \in V\\ \forall D'' \in V, D'' \subset D, D'' \not\supset D'}} e'(D')$$

produces the required partial specification. In the following example, the transformation discussed is applied.

Example 3.9. Reconsider Example 3.7 and Figure 3.7. Application of the transformation discussed above, leads to the evidence function e', that is described by means of the following bottom-up partial specification:

$$\begin{aligned}
\tilde{e}'(\{n_1\}) &= \varnothing\\
\tilde{e}'(\{\neg n_1\}) &= \{\neg o\}\\
\tilde{e}'(\{n_2\}) &= \{\neg o\}\\
\tilde{e}'(\{\neg n_2\}) &= \varnothing\\
\tilde{e}'(\{\neg n_1, n_2\}) &= \{\neg o, o\}\\
\tilde{e}'(\{n_1, \neg n_2\}) &= \{o\}
\end{aligned}$$

There is insufficient information concerning the individual components available in the top-down partial specification, as is demonstrated by the fact that several function values are empty. For example, the function value $\tilde{e}'(\{n_1\}) = \varnothing$ indicates that it was not possible to isolate observable findings from the supersets of n_1 for n_1. Hence, it is not possible to describe n_1 separately from n_2 and $\neg n_2$. The bottom-up partial specification is quite extensive; only the function values $e(\{\neg n_1, \neg n_2\})$ and $e(\{n_1, n_2\})$ have not been specified explicitly. Obviously, Δ is not interaction free with respect to e' (which was, in fact, already evident from the problem description), because, for example, $e(\{\neg n_1, n_2\}) \neq \tilde{e}(\{\neg n_1\}) \cup \tilde{e}(\{n_2\})$. ◊

Note that if $D \notin V$, $D \neq \varnothing$, for the top-down partial specification $\tilde{e}$, then

$$e'(D) = \bigcap_{\substack{D' \supset D, D' \mathit{\ consistent}, D' \in V\\ \forall D'' \in V, D'' \supset D \,:\, D'' \not\subset D'}} e(D') \tag{3.6}$$

This transformation yields an evidence function that is defined in terms of common findings associated with sets of defects. It has, therefore, the tendency to reverse the structure of an evidence function. This might lead us to suspect that if an evidence function that is specified by means of a top-down partial specification is monotonically decreasing, then the evidence function e' obtained by the transformation above will be monotonically increasing. This conjecture, however, is false in general, as is shown by the following counter-example.

Example 3.10. Consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$ with the evidence function given by the following top-down partial specification $\tilde{e}$:

$$\tilde{e}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1, d_2, d_3\} \\ \{f_2\} & \text{if } D = \{d_1, d_2, \neg d_3\} \\ \{f_3\} & \text{if } D = \{d_1, \neg d_2, d_3\} \\ \{f_4\} & \text{if } D = \{d_1, \neg d_2, \neg d_3\} \\ \{f_1, f_2, f_3, f_4\} & \text{if } \neg d_1 \in D \text{ and } d, \neg d \notin D \\ \{f_1, f_2, f_4\} & \text{if } D = \{d_1, \neg d_2\} \\ \{f_1, f_3, f_4\} & \text{if } D = \{d_1, \neg d_3\} \end{cases}$$

For example, we have $e(\{d_1, d_2\}) = \tilde{e}(\{d_1, d_2, d_3\}) \cup \tilde{e}(\{d_1, d_2, \neg d_3\}) = \{f_1, f_2\}$. Note that the resulting evidence function e is monotonically decreasing. For the transformed evidence function e' it holds that $e'(\{d_1\}) = \{f_1\}$, but $e'(\{d_1, d_2\}) = \varnothing$. Hence, the evidence function e' is nonmonotonic. ♢

Equation (3.6) simplifies to

$$e'(D) = \bigcap_{\substack{D' \supseteq D \\ D' \in V}} \tilde{e}(D')$$

for each consistent $D \in \Delta$, $D \neq \varnothing$, if Δ is externally described with respect to e. This is exactly the partial specification referred to in Proposition 3.5; applying the transformation to an evidence function e, where Δ is externally described – hence, e is monotonically decreasing – yields an evidence function that is monotonically increasing.

Example 3.11. Consider a logic circuit consisting of an XOR gate X and an AND gate A in parallel, with input channels I_1 and I_2 for the XOR gate, and I_3 and I_4 for the AND gate. The circuit is depicted in Figure 3.10. The output signal of the XOR gate is denoted by O_1, and the output signal of the AND gate is denoted by O_2. We use the same convention to denote specific input and output signals as in Example 3.2. Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification used for the specification of the circuit's behaviour. The evidence function e is defined by the following top-down partial specification:

$$\tilde{e}(D) = \begin{cases} \{\neg o_1, \neg o_2\} & \text{if } D = \{x, a\} \\ \{\neg o_1, o_2\} & \text{if } D = \{x, \neg a\} \\ \{o_1, \neg o_2\} & \text{if } D = \{\neg x, a\} \\ \{o_1, o_2\} & \text{if } D = \{\neg x, \neg a\} \end{cases}$$

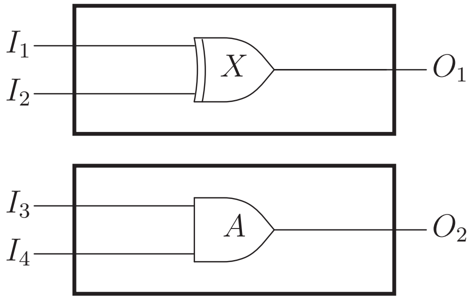


Figure 3.10: Two parallel gates.

In this case, Δ is externally described. The evidence function e' resulting from the transformation can be represented by the following bottom-up partial specification:

$$\tilde{e}'(D) = \begin{cases} \{\neg o_1\} & \text{if } D = \{x\} \\ \{o_1\} & \text{if } D = \{\neg x\} \\ \{\neg o_2\} & \text{if } D = \{a\} \\ \{o_2\} & \text{if } D = \{\neg a\} \end{cases}$$

Note that the evidence function e' is monotonically increasing; in fact, it expresses Δ to be interaction free. ◊

In the following, we shall often assume that a problem domain can be described by means of a bottom-up or top-down partial specification.

3.1.4 A formal notion of diagnosis

The interpretation of a knowledge base for the purpose of diagnosis is a central issue of any theory of diagnosis. Such interpretations were captured in the notion of evidence function. An evidence function, however, does not yet provide a means for diagnosing a given problem. Additional knowledge is required, stating how the function e is to be interpreted in the process of diagnosis. In other words, a separate diagnostic interpretation, which we call a ‘notion of diagnosis’, must be designed on top of the interpretation of the knowledge. This approach is in line with current views on the design of knowledge-based systems. It is believed that when designing a knowledge base, certain concepts, such as causality, may be conveniently employed to guide the modelling process. Its final diagnostic use, however, should not be taken into account in the early stages of the design. The idea has been proposed in literature on knowledge acquisition and modelling in recent years, where it has been investigated in an informal way (cf. for example [Abu-Hanna & Jansweijer, 1994], [Benjamins & Jansweijer, 1994], and [Wielinga et al., 1993]). In the present and following sections, the formal underpinning of this idea is explored.

Given a set of findings E, to be interpreted as the findings that have been *observed*, a diagnosis is defined as a consistent set of defects $D \subseteq \Delta$ that is included in a relationship between E and D, expressed by a mapping $\wp(\Phi) \to \wp(\Delta) \cup \{u\}$. In addition, to express whether a diagnosis is the result of an accepted, rejected or adjusted hypothesis $H \subseteq \Delta$ (cf. Section 1.1), restricted evidence functions $e_{|H}$ will be incorporated. Finally, to enable the application of the same diagnostic inference relation to other diagnostic specifications, generalization to the set of all diagnostic specifications is proposed. The resulting formalization is called a notion of diagnosis. In the following, $[A \to B]$ is used to denote a function space, i.e. the set of functions with domain A and codomain B.

Definition 3.19 (*function-space schema*)**.** *Let B be a set of elements, and let $X, Y \subseteq B$ and $t \in B$, then a* function-space schema *with respect to X, Y and t, denoted by $F_{X,Y,t}$, is defined as follows:*

$$F_{X,Y,t} = \bigcup_{\substack{V \subseteq X \\ W \subseteq Y}} [\wp(V) \to \wp(W) \cup \{t\}]$$

Function-space schemata can be conveniently employed to define notions of diagnosis.

Definition 3.20 (*notion of R-diagnosis*)**.** *Let $\mathcal{S} = \{(\Delta, \Phi, e) \mid \Delta \subseteq \mathcal{D}, \Phi \subseteq \mathcal{F}, e : \wp(\Delta) \to \wp(\Phi) \cup \{\bot\}\}$ be the set of all diagnostic specifications Σ based on the diagnostic universe $\mathcal{U} = (\mathcal{D}, \mathcal{F}, u, \bot)$. A* notion of diagnosis *R is then defined as a partial function*

$$R : \mathcal{S} \times F_{\mathcal{D},\mathcal{F},\bot} \to F_{\mathcal{F},\mathcal{D},u}$$

A partial function

$$R_\Sigma : F_{\Delta,\Phi,\bot} \to F_{\Phi,\Delta,u}$$

defined for some $\Sigma \in \mathcal{S}$, represents the notion of diagnosis R applied to $\Sigma = (\Delta, \Phi, e) \in \mathcal{S}$. For each $H \subseteq \Delta$, the function value $R_\Sigma(e_{|H}) \in F_{\Phi,\Delta,u}$, also denoted by $R_{\Sigma,e_{|H}}$, is a total function

$$R_{\Sigma,e_{|H}} : \wp(\Phi) \to \wp(\Delta) \cup \{u\}$$

called a diagnostic inference component*, or* diagnostic component *for short. It is assumed that $R_{\Sigma,e_{|H}}(E) \subseteq H$ if $R_{\Sigma,e_{|H}}(E) \neq u$.*

Recall that the symbol u denotes the undefined defects symbol; it expresses that a diagnosis is undefined. Thus, a notion of diagnosis R is obtained by fixing functions R_Σ for particular diagnostic specifications Σ. In turn, a function R_Σ is defined in terms of restricted evidence functions $e_{|H}$. Finally, a diagnostic inference component is defined in terms of sets of observed findings E and resulting sets of defects H'. As is reflected in the definition, a notion of diagnosis R is assumed to be parameterized with respect to diagnostic specifications Σ. Possibly, a notion of diagnosis R is only defined for diagnostic specifications having specific properties. It is also possible that a notion of diagnosis is completely determined by a single diagnostic specification, i.e. is completely domain-specific. Sometimes, diagnosis amounts to establishing defects for some system, using only knowledge concerning abnormality. This approach is adopted in domains where it is easier to obtain knowledge concerning defects than knowledge concerning the normal situation. The evidence function may then be restricted to the set of positive defects Δ_P, where positive defects express the presence of defects. Instead of providing a diagnostic component for every possible evidence function $e_{|H}$, it is also conceivable that diagnostic components are only defined for particular restrictions of e, e.g. only for $e_{|\Delta_P}$.

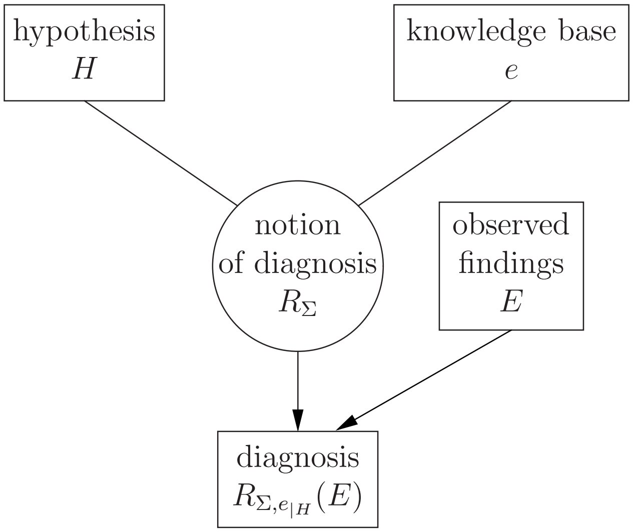


Figure 3.11: Schema of notion of diagnosis, diagnostic problem and solution.

Whereas the evidence function e corresponds to the (interpretation of a) knowledge base of an expert system, and E corresponds to the facts entered into the system, the function R_{Σ} stands for the diagnostic methods employed for deriving the conclusions that follow from the input facts and the knowledge base, i.e. this function corresponds to part of the reasoning methods of the system. A diagnostic component is the formal interpretation of an inference relation with respect to part of the knowledge base, selected by a diagnostic hypothesis. The function R_{Σ} may also be viewed as the formalization of the input-output relationships, in terms of parts of a knowledge base, of a diagnostic expert system.

The definition above is very unrestrictive; one conceivable restriction on the notion of diagnosis is obtained by assuming that for each nonempty $E \subseteq \Phi$, and each nonempty, consistent set $H \subseteq \Delta$, for which $R_{\Sigma,e_{|H}}(E) = H'$, with $H' \neq u$, it holds that $e_{|H}(H') \cap E \neq \varnothing$ if $e_{|H}(H') \neq \varnothing$. The condition $e_{|H}(H') \cap E \neq \varnothing$ simply means that the result H' of applying a diagnostic component should have at least some relevance with respect to the given set of findings E. This is a rather weak condition. More precise constraints on the notion of diagnosis shall be introduced below for specific notions of diagnosis. Next, a diagnosis is defined as the result of applying a diagnostic component to a set of observed findings.

Definition 3.21 (*diagnostic problem and solution*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let $E \subseteq \Phi$ be a set, called a set of* observed findings*, such that for each $f \in E$: $\neg f \notin E$. Let R be a notion of diagnosis. A* diagnostic problem *$\mathcal{P}$ is then defined as a pair $\mathcal{P} = (\Sigma, E)$; the R-*diagnostic solution*, or R-*diagnosis *for short, with respect to the set of defects $H \subseteq \Delta$ is defined as follows:*

$$R_{\Sigma,e_{|H}}(E)$$

In Figure 3.11, the idea underlying the definition of a notion of diagnosis R and diagnostic solution to a diagnostic problem is illustrated schematically. The diagnosis $R_{\Sigma,e_{|H}}(E)$

is sometimes simply denoted by a capital letter (possibly supplied with a subscript or quoted), such as H', if the underlying structure is not of essential importance. The set of observed findings E denotes findings that are present or absent at a given time, which is in contrast with the set of observable findings associated with a set of defects D, i.e. $e(D)$. These findings need not all be observed at the same time. The set $H \subseteq \Delta$ stands for a *hypothesis*, defined as a collection of defects, which is to be investigated. If application of the diagnostic component $R_{\Sigma,e_{|H}}$ yields as a result $H' = R_{\Sigma,e_{|H}}(E)$, it is said that:

(1) the hypothesis H is *accepted* if $H' = H$;

(2) the hypothesis H is *rejected* if $H' = u$;

(3) otherwise, the hypothesis H is said to be *adjusted*.

Adjustment of a hypothesis indicates that not all defects in H have passed when the hypothesis was tested against E, i.e. the result H' is taken as the adjusted version of the original hypothesis H.

The following example illustrates how the two definitions above can be applied.

Example 3.12. Consider a medical diagnostic problem, where a patient may have Cushing's disease (d_1) – a disease caused by a brain tumour, producing hyperfunctioning of the adrenal glands – and pulmonary infection (d_2); $\Delta_P = \{d_1, d_2\}$. We shall not enumerate all signs and symptoms associated with these medical problems; it suffices to note that moon face (f_1) is a sign associated with Cushing's disease, and fever (f_2) and dyspnoea (shortness of breath, f_3) are associated with pulmonary infection. We have: $\Phi_P = \{f_1, f_2, f_3\}$. However, in a patient in whom Cushing's disease and pulmonary infection coexist there usually is no fever. Thus, the following bottom-up partial specification $\tilde{e}$ of the evidence function e is obtained:

$$\tilde{e}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\} \\ \{f_2, f_3\} & \text{if } D = \{d_2\} \\ \{f_1, f_3\} & \text{if } D = \{d_1, d_2\} \\ \varnothing & \text{if } D = \{d\}, d \in \Delta_N \end{cases}$$

Note that cancellation occurs with the generated evidence function, as discussed in Section 3.1.2, because

$$e(\{d_1, d_2\}) \subset e(\{d_1\}) \cup e(\{d_2\})$$

Consider a notion of diagnosis U, that is defined as follows. Let $E \subseteq \Phi$ be a set of observed findings, then $U_{\Sigma,e_{|H}}(E) = H'$, with $H' \subseteq H$, if H' is the only subset of H such that $e_{|H}(H') \subseteq E$; otherwise, $H' = u$. This notion of diagnosis expresses that a diagnosis consists of a set of (positive) defects which, on the one hand, can account for at least part of all observed findings, and, on the other hand, every finding associated with the set of defects that is taken as a diagnosis has been observed. Furthermore, there is only one such subset of the given hypothesis H. Some interesting diagnostic conclusions are: $U_{\Sigma,e_{|\Delta_P}}(\{f_2, f_3\}) = \{d_2\}$, i.e. a patient with only fever and dyspnoea has pulmonary infection, $U_{\Sigma,e_{|\Delta_P}}(\{f_1, f_2\}) = u$, i.e. there exists no diagnosis accounting for both moon face and fever as signs, and finally, $U_{\Sigma,e_{|\Delta_P}}(\{f_1, f_3\}) = \Delta_P$. Hence, in the first case, the hypothesis has been adjusted, in the second case, the hypothesis $H = \Delta_P$ is rejected, and in the last case, the hypothesis $H = \Delta_P$ has been accepted. $\Diamond$

The definition of the notion of R-diagnosis above is sufficiently general to permit the expression of one or more of the following aspects:

- Relationships among function values $e(D)$, on the one hand, and the set of observed findings E, on the other hand;
- Restriction of the application of the evidence function e to defects in the subset H, expressing 'consulting' a knowledge base with respect to a limited number of defects, taken as a hypothesis;
- Relationships among the defects in H (possibly also involving relationships among the findings in E);
- A resulting set of defects $D = R_{\Sigma,e_{|H}}(E)$, taken as a diagnosis.

A form of diagnosis yielding a unique diagnostic solution to a diagnostic problem, using a single diagnostic component, will be referred to as *single diagnosis*. The meaning of this term differs from the meaning of the term 'single fault diagnosis' mentioned in literature (cf. [De Kleer & Williams, 1987; Peng & Reggia, 1990; Reiter, 1987]), where it means a diagnosis consisting of a single defect.

Note that it is possible that

$$R_{\Sigma,e_{|H}}(E) = R_{\Sigma,e_{|H'}}(E)$$

for $H \neq H'$; although the two diagnoses are equal in this case, they are the result of two different diagnostic components, differing with respect to the hypothesis tested. The fact that it is possible to distinguish between hypotheses and the result of testing the hypothesis is one of the advantages of the formalism.

As remarked above, Definition 3.20 imposes very few constraints with respect to the properties that must be satisfied by a reasonable notion of diagnosis. One desirable property that, however, usually fails to hold, is that a notion of diagnosis respects the evidence function e.

Definition 3.22 (*R respects e*). *Let R be a notion of diagnosis defined for the diagnostic specification $\Sigma = (\Delta, \Phi, e)$. It is said that R* respects e *if*

(1) *for each set of observed findings $E \subseteq \Phi$, there exists a set $H \subseteq \Delta$ such that $e(R_{\Sigma,e_{|H}}(E)) = E$,*

(2) *for each consistent $D \subseteq \Delta$, there exists a set $H \subseteq \Delta$, such that $R_{\Sigma,e_{|H}}(e(D)) = D$ and for each $H' \not\supseteq H$: $R_{\Sigma,e_{|H'}}(e(D)) = u$.*

This means that a function that is taken as the inverse of the evidence function e, which must be bijective (excluding inconsistent sets of defects and sets $E \subseteq \Phi$ with complementary findings), is composed of function values $R_{\Sigma,e_{|H}}(E)$, where the set $H \subseteq \Delta$ need not be fixed. Of course, the two conditions above will also hold if there exists a function $R_{\Sigma,e_{|H}}$ with fixed H that can be taken as the inverse. This definition imposes very strong constraints on a notion of diagnosis. Now, suppose that a diagnosis $R_{\Sigma,e_{|H}}(E) \neq u$ is produced for at least one hypothesis $H \subseteq \Delta$. Furthermore, suppose that the notion of diagnosis does not respect the evidence function e, because $e(R_{\Sigma,e_{|H}}(E)) \neq E$, for each $H \subseteq \Delta$ for which $R_{\Sigma,e_{|H}}(E)$ is defined. Two reasons can be distinguished to explain this situation from a conceptual point of view:

(1) If findings $f \in e(R_{\Sigma,e_{|H}}(E))$ exist with $f \notin E$ – the findings f are observable (predicted) but have not been observed – then one of the following holds:

   a. The findings f represent findings for the normal (abnormal) situation; they have been omitted, because only abnormal (normal) findings have been included in E for reasons of economy. We may also have that $\neg f \in E$.

   b. No attempt has been made to observe findings f, but if the observation had been made, they would have been observed.

   c. It has been attempted to observe findings f, but the findings could not be observed. This might occur when the findings f are only observed occasionally for the combination of defects $D = R_{\Sigma,e_{|H}}(E)$.

(2) If findings $f \in E$ exist with $f \notin e(R_{\Sigma,e_{|H}}(E))$. The findings f not accounted for, may be due to defects (present or absent) that have not been incorporated in the knowledge base. For example, an observed finding may be a normal finding, where the evidence function incorporates only knowledge of the abnormal situation, such as in MAB diagnosis. Similarly, an observed finding may be an abnormal finding, where the evidence function only deals with findings for the normal situation, such as in DNSB diagnosis. It may also hold that $\neg f \in e(R_{\Sigma,e_{|H}}(E))$.

Hence, a notion of diagnosis R that does not respect an evidence function e may still offer a sensible interpretation of e. The situation becomes more complicated if we want to take into account that real-world knowledge bases are usually imperfect. This aspect of diagnosis is treated in Section 3.2.2, and Chapter 5.

If a notion of diagnosis respects an evidence function, and, in addition, an evidence function is interaction free, the following proposition holds.

Proposition 3.7. *Let R be a notion of diagnosis defined for the diagnostic specification $\Sigma = (\Delta, \Phi, e)$ that respects e, where e is interaction free. Then,*

$$R_{\Sigma,e_{|H}}(E) = R_{\Sigma,e_{|H}}(E') \cup R_{\Sigma,e_{|H}}(E'')$$

for each set of observed findings E, E' and E'', with $E, E', E'' \subseteq \Phi$ and $E = E' \cup E''$, and $H \subseteq \Delta$.

Proof. Since e is bijective if restricted to consistent sets of defects D, we know that there exist sets D, D' and D'' such that $E = e(D)$, $E' = e(D')$ and $E'' = e(D'')$, with $E = E' \cup E''$. Then, using the fact that e is interaction free: $e(D) = e(D') \cup e(D'') = e(D' \cup D'')$.

Therefore, $D = D' \cup D''$, because e is injective. From the fact that R respects e it follows that

$$\begin{aligned} R_{\Sigma,e_{|H'}}(E') \cup R_{\Sigma,e_{|H''}}(E'') &= R_{\Sigma,e_{|H'}}(e(D')) \cup R_{\Sigma,e_{|H''}}(e(D'')) \\ &= D' \cup D'' \\ &= R_{\Sigma,e_{|H}}(e(D' \cup D'')) \\ &= R_{\Sigma,e_{|H}}(E) \end{aligned}$$

for some consistent $H, H', H'' \subseteq \Delta$. Furthermore, since R respects e and $R_{\Sigma,e_{|H}}(E) = D$ it follows that $e_{|H}(D) = E$ ($D \subseteq H$ holds by definition). Similarly, from $R_{\Sigma,e_{|H'}}(E') = D'$ we have $e_{|H'}(D') = E'$. Moreover, because $D' \subseteq D$ it follows that $D' \subseteq H$, hence $e_{|H}(D') = E'$. Therefore, $R_{\Sigma,e_{|H'}}(E') = R_{\Sigma,e_{|H}}(E')$. Analogously, $R_{\Sigma,e_{|H''}}(E'') = R_{\Sigma,e_{|H}}(E'')$. $\Diamond$

Hence, it turns out that if a notion of diagnosis R respects an interaction-free evidence function e, the set of observed findings can be partitioned, such that each subset can be accounted for separately by the same diagnostic component. Note that if we have an evidence function e for which $f, \neg f \in e(D)$, for some $D \subseteq \Delta$, then R cannot respect e, due to the fact that E cannot contain complementary findings, at least, if $R_{\Sigma,e_{|H}}(E)$ is to be interpreted as a diagnosis.

It is also possible to define notions of diagnosis R for which the interaction-freeness of the evidence function is taken into account, although it need not be satisfied that R respects e. Consider the following example.

Example 3.13. A notion of diagnosis Q is defined, such that for each diagnostic specification $\Sigma \in \mathcal{S}$, for which e expresses Δ to be interaction free, the following holds: for each $d \in H$, $H \subseteq \Delta_P$: $e_{|H}(d) \subseteq E$ iff $d \in Q_{\Sigma,e_{|H}}(E)$ given that there exists no $d' \in H$, $d' \neq d$, with $e_{|H}(d') \subseteq E$ and $e(d) \cap e(d') \neq \varnothing$; otherwise, $Q_{\Sigma,e_{|H}}(E) = u$, i.e. undefined. Recall that due to the condition of interaction freeness, it holds that $e(D) = \bigcup_{d \in D} e(d)$, and $e(d) \cap e(d') \neq \varnothing$ implies $e(D) \cap e(D') \neq \varnothing$ for $d \in D$ and $d' \in D'$; hence, only singleton sets $\{d\}$, $d \in \Delta_P$, need be considered. Intuitively, this notion of diagnosis only accepts defects as part of a diagnosis if:

(1) every observable finding associated with a defect has actually been observed;

(2) none of the findings associated with a defect in H can be accounted for by another defect from H.

The last condition would, for example, be acceptable in a broad medical domain in which it is thought that if the patient has more than one disorder, those disorders must be from different medical subdomains (hence, a bit simplified, findings would be disjoint). Every finding is uniquely accounted for by a defect. Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem, where $\Sigma = (\Delta, \Phi, e)$ with, $\Delta_P = \{d_1, d_2\}$, $\Phi_P = E = \{f_1, f_2\}$, and e is given by the bottom-up partial specification $\tilde{e}$:

$$\tilde{e}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\} \\ \{f_1, f_2\} & \text{if } D = \{d_2\} \\ \varnothing & \text{if } D = \{\neg d_i\}, i = 1, 2 \end{cases}$$

Note that, for example, $e(\{d_1, d_2\}) = \{f_1, f_2\}$. Using the definition above, the following cases can be distinguished for Q, where it is assumed that Q is only defined for $H \subseteq \Delta_P$, and each function $Q_{\Sigma,e_{|H}}$ is represented as a set of pairs:

- $H = \varnothing$; then,

  $$\begin{aligned} Q_{\Sigma,\varnothing} = \{&(\varnothing, \varnothing),\\ &(\{f_1\}, \varnothing),\\ &(\{f_2\}, \varnothing),\\ &(\{f_1, f_2\}, \varnothing)\} \end{aligned}$$

  The diagnosis is equal to $Q_{\Sigma,\varnothing}(E) = \varnothing$.

- $H = \{d_1\}$; then,

  $$\begin{aligned} Q_{\Sigma,e_{|\{d_1\}}} = \{&(\varnothing, \varnothing),\\ &(\{f_1\}, \{d_1\}),\\ &(\{f_2\}, \varnothing),\\ &(\{f_1, f_2\}, \{d_1\})\} \end{aligned}$$

  The diagnosis is equal to $Q_{\Sigma,e_{|\{d_1\}}}(E) = \{d_1\}$.

- $H = \{d_2\}$; then,

  $$\begin{aligned} Q_{\Sigma,e_{|\{d_2\}}} = \{&(\varnothing, \varnothing),\\ &(\{f_1\}, \varnothing),\\ &(\{f_2\}, \varnothing),\\ &(\{f_1, f_2\}, \{d_2\})\} \end{aligned}$$

  The diagnosis is equal to $Q_{\Sigma,e_{|\{d_2\}}}(E) = \{d_2\}$.

- $H = \{d_1, d_2\}$; then,

  $$\begin{aligned} Q_{\Sigma,e_{|\{d_1,d_2\}}} = \{&(\varnothing, \varnothing),\\ &(\{f_1\}, \{d_1\}),\\ &(\{f_2\}, \varnothing),\\ &(\{f_1, f_2\}, u)\} \end{aligned}$$

  Hence, the diagnosis is equal to $Q_{\Sigma,e_{|\{d_1,d_2\}}}(E) = u$, because $e(d_1) \cap e(d_2) \neq \varnothing$. Clearly, the empty or undefined diagnosis need not be obtained for each $E \subseteq \Phi$; for example, the diagnosis is equal to the set $\{d_1\}$ if the set of observed findings E is equal to $E = \{f_1\}$.

$\Diamond$

From now on, for reasons of convenience, tuples containing the empty set as a diagnosis will be left out from the diagnostic components $R_{\Sigma,e_{|H}}$.

It is often possible to employ a somewhat simpler notion of diagnosis, if diagnostic components can be decomposed into smaller components, which collectively yield the same result as the original diagnostic component. The following example illustrates the basic idea.

Example 3.14. The following notion of diagnosis S is defined for interaction-free diagnostic specifications $\Sigma \in \mathcal{S}$. Let $E \subseteq \Phi$ be a set of observed findings, then

$$S_{\Sigma,e_{|H}}(E) = \bigcup_{H' \subseteq H,\, e_{|H}(H') \subseteq E} H'$$

for each consistent $H \subseteq \Delta$. The intuitive idea underlying this notion of diagnosis is that only defects in a hypothesis H that have all their associated findings included as observed findings are admitted as part of a diagnosis. Consider again the diagnostic problem $\mathcal{P}$ from the previous example (where $E = \{f_1, f_2\}$). In addition to the situation in which $H = \varnothing$, and in which H contains negative defects, the following cases can be distinguished (recall that pairs having empty second components will be omitted):

- $H = \{d_1\}$; then,

  $$\begin{aligned} S_{\Sigma,e_{|H}} = \{&(\{f_1\}, \{d_1\}),\\ &(\{f_1, f_2\}, \{d_1\})\} \end{aligned}$$

  Hence, $S_{\Sigma,e_{|H}}(E) = \{d_1\}$.

- $H = \{d_2\}$; then,

  $$S_{\Sigma,e_{|H}} = \{(\{f_1, f_2\}, \{d_2\})\}$$

  Therefore, $S_{\Sigma,e_{|H}}(E) = \{d_2\}$.

- $H = \{d_1, d_2\}$; then,

  $$\begin{aligned} S_{\Sigma,e_{|H}} = \{&(\{f_1\}, \{d_1\}),\\ &(\{f_1, f_2\}, \{d_1, d_2\})\} \end{aligned}$$

  Therefore

  $$S_{\Sigma,e_{|H}}(E) = \{d_1, d_2\}$$

Note that

$$S_{\Sigma,e_{|\{d_1,d_2\}}}(E) = S_{\Sigma,e_{|\{d_1\}}}(E) \cup S_{\Sigma,e_{|\{d_2\}}}(E)$$

In fact, this property holds in general, i.e. for every set of observed findings E and any bipartition of the hypothesis H. ◊

This particular property of a notion of diagnosis is now formally defined. The notion of diagnosis in the example above is said to satisfy the independence assumption, because, by means of this definition, (possible) interactions among defects expressed by means of the evidence function are ignored.

Definition 3.23 (*independence assumption*). *Let R be a notion of diagnosis. It is said that R fulfils the* independence assumption *if for each diagnostic specification $\Sigma \in \mathcal{S}$ for which R_Σ is defined, and for each pair of consistent sets of defects $H, H' \subseteq \Delta$ and each set of observed findings $E \subseteq \Phi$ it holds that*

$$R_{\Sigma,e_{|H\cup H'}}(E) = R_{\Sigma,e_{|H}}(E) \cup R_{\Sigma,e_{|H'}}(E)$$

with $R_{\Sigma,e_{|H\cup H'}}(E) \neq u$.

Although the independence assumption states that it is sufficient to investigate functions $e_{|\{d\}}$, $d \in \Delta$, this does not mean that from the independence assumption it follows that e is interaction free. Generally speaking, the independence assumption states that if a diagnostic relationship between a function value $e_{|H}(D)$ and E is satisfied, then for each $d \in D$ the same relationship is assumed to hold between $e_{|\{d\}}(\{d\})$ and E as well. It is possible that the independence assumption is satisfied for a notion of diagnosis, but that the set of defects for which the notion of diagnosis is defined, is not interaction free.

Example 3.15. Reconsider the notion of diagnosis S from Example 3.14. Suppose that the definition of S is restricted to the following diagnostic specification $\Sigma = (\Delta, \Phi, e)$, with bottom-up partial specification $\tilde{e}$:

$$\tilde{e}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2\} \\ \{f_1, f_2, f_3\} & \text{if } D = \{d_1, d_2\} \\ \varnothing & \text{if } D = \{\neg d_i\}, i = 1, 2 \end{cases}$$

Note that

$$S_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_1\}) = S_{\Sigma,e_{|\{d_1\}}}(\{f_1\}) \cup S_{\Sigma,e_{|\{d_2\}}}(\{f_1\}) \qquad (3.7)$$

The independence assumption is satisfied for S, because

$$e(\{d_i\}) \subseteq e(\{d_1, d_2\})$$

$i = 1, 2$, is satisfied; therefore, if $e(\{d_1, d_2\}) \subseteq E$, then $e(d_i) \subseteq E$, $i = 1, 2$. If the evidence function e was defined as above, except that $\tilde{e}(\{d_1, d_2\}) = \{f_3\}$, condition (3.7) would fail to hold. Hence, the independence assumption fails to hold if the notion of diagnosis S is not restricted to specific diagnostic specifications. ◊

When the independence assumption is satisfied for a notion of diagnosis R, the notion of diagnosis can be simplified. The simplification amounts to providing only the diagnostic interpretation of an evidence function with respect to individual defects. Based on the independence assumption, a simplified form of a notion of diagnosis is defined.

Definition 3.24 (*independent form*). *Let R be a notion of diagnosis, then the* independent form *of a notion of diagnosis R, denoted by R^i, is defined as the smallest relation:*

$$R^i \subseteq \mathcal{S} \times \wp(\mathcal{F}) \times (\mathcal{D} \times \wp(\mathcal{F}))$$

such that for each $\Sigma \in \mathcal{S}$, and each $d \in \Delta$ and set of observed findings $E \subseteq \Phi$: if $R_{\Sigma,e_{|\{d\}}}(E) \neq u$, it holds that $R_{\Sigma,e_{|\{d\}}}(E) = \{d\}$ iff $(\Sigma, E, (d, F)) \in R^i$ with $F = e(d)$.

If a tuple $(\Sigma, E, (d, F))$ is included in a relation R^i, this will also be written as $R^i(\Sigma, E, (d, F))$. Note that from the definition above (and from the definition of a notion of diagnosis), it follows that if $R^i(\Sigma, E, (d, F))$ fails to hold, $R_{\Sigma, e_{|\{d\}}}(E) = \varnothing$. The relation R^i is defined as the smallest relation, because only the smallest relation is uniquely determined by R. For specific diagnostic specifications Σ, we shall often write R^i_Σ for the independent form of a notion of diagnosis R with respect to Σ; this set consists of tuples of the form $(E, (d, F))$.

It is easily shown that under the independence assumption, a notion of diagnosis can be translated into its independent form.

Proposition 3.8. *Let R be a notion of diagnosis for which the independence assumption holds, and for which $R_{\Sigma, e_{|H}}(E) \neq u$. Then,*

$$R_{\Sigma, e_{|H}}(E) = \{d \in H \mid R^i(\Sigma, E, (d, F)),\ F = e(d)\}$$

for each $\Sigma \in \mathcal{S}$ for which R_Σ is defined, and each set of observed findings $E \subseteq \Phi$, where R^i is the independent form of R.

Proof. Let $D(H, E) = \{d \in H \mid R^i(\Sigma, E, (d, F)),\ F = e(d)\}$. If the independence assumption holds, then

$$R_{\Sigma, e_{|H}}(E) = \bigcup_{d \in H} R_{\Sigma, e_{|\{d\}}}(E)$$

Let $R_{\Sigma, e_{|\{d\}}}(E) = H'$, then $H' = \{d\}$ or $H' = \varnothing$, because $R_{\Sigma, e_{|H}}(E) \neq u$. Let $F = e(d)$, $d \in \Delta$. If $R_{\Sigma, e_{|\{d\}}}(E) = \{d\}$ then, by definition, $R^i(\Sigma, E, (d, F))$ holds, hence $d \in D(H, E)$; otherwise, $R_{\Sigma, e_{|\{d\}}}(E) = \varnothing$, and $R^i(\Sigma, E, (d, F))$ fails to hold. Conversely, if $R^i(\Sigma, E, (d, F))$ then $R_{\Sigma, e_{|\{d\}}}(E) = \{d\}$, thus $d \in R_{\Sigma, e_{|H}}(E)$. $\Diamond$

A diagnostic component of a notion of diagnosis R for which the independence assumption holds, continues to be written as $R_{\Sigma, e_{|H}}$, but we shall sometimes employ the independent form R^i of R.

Example 3.16. Reconsider the notion of diagnosis S from Example 3.14. Collecting all information expressed by the example function S_Σ (due to space considerations, S_Σ is restricted to positive defects only), yields the following independent form S^i_Σ:

$$S^i_\Sigma = \{(\{f_1\}, (d_1, \{f_1\})), (\{f_1, f_2\}, (d_1, \{f_1\})), (\{f_1, f_2\}, (d_2, \{f_1, f_2\}))\}$$

The simplification of S_Σ to S^i_Σ is possible, because the independence assumption is satisfied for S. The definition of the possible diagnostic solutions can be restated as follows:

$$S_{\Sigma, e_{|H}}(E) = \{d \in H \mid S^i(\Sigma, E, (d, F))), F = e(d)\}$$

For example, if $H = \{d_1, d_2\}$ and $E = \{f_1, f_2\}$, then $S_{\Sigma, e_{|H}}(E) = \{d_1, d_2\}$, because both $S^i(\Sigma, E, (d_1, e(d_1)))$ and $S^i(\Sigma, E, (d_2, e(d_2)))$ hold. $\Diamond$

Next, the notion of Δ-monotonicity is defined for a notion of R-diagnosis, which is a property of a notion of diagnosis in line with the independence assumption.

Definition 3.25 (Δ-*monotonicity*). *A notion of diagnosis R is called* Δ-monotonic *if for each diagnostic specification $\Sigma \in \mathcal{S}$ for which R_Σ is defined, each consistent set of defects $H \subseteq H'$, with $H, H' \subseteq \Delta$, and each set of observed findings $E \subseteq \Phi$, it holds for the diagnostic problem $\mathcal{P} = (\Sigma, E)$ that if $R_{\Sigma,e_{|H}}(E) \neq u$, then $R_{\Sigma,e_{|H}}(E) \subseteq R_{\Sigma,e_{|H'}}(E)$; otherwise, it is called* Δ-nonmonotonic.

Δ-monotonicity means: the larger (with respect to $\subseteq$) the hypothesis investigated, the larger the diagnostic solution. Note that from Δ-monotonicity, it follows that if $H \subseteq H'$, then $e(R_{\Sigma,e_{|H}}(E)) \subseteq e(R_{\Sigma,e_{|H'}}(E))$ if e is monotonically increasing. In the following example, a Δ-nonmonotonic notion of diagnosis is discussed.

Example 3.17. Consider a notion of diagnosis O for an interaction-free evidence function e, which admits a defect $d \in H$ to a diagnosis $O_{\Sigma,e_{|H}}(E)$ iff $e(d) \cap E \neq \varnothing$, and for each $d' \in H$ with $e(d') \subset e(d)$ it holds that $e(d') \cap E = \varnothing$. Hence, this notion of diagnosis selects the most specific defects from the hypothesis H. Now, consider an interaction-free evidence function e with $e(d_1) = \{f_1, f_2\}$, $e(d_2) = \{f_1\}$. Then, $O_{\Sigma,e_{|\{d_1\}}}(\{f_1\}) = \{d_1\}$, but $O_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_1\}) = \{d_2\}$. Hence, this notion of diagnosis O is Δ-nonmonotonic. ◊

The following proposition states that any notion of diagnosis satisfying the independence assumption is Δ-monotonic.

Proposition 3.9. *A notion of R-diagnosis is Δ-monotonic if the independence assumption is satisfied.*

Proof. Let R be a notion of diagnosis, then for every diagnostic specification $\Sigma \in \mathcal{S}$: if $H \subseteq H'$, with consistent $H, H' \subseteq \Delta$, and $R_{\Sigma,e_{|H}}(E) \neq u$, then $R_{\Sigma,e_{|H}}(E) \subseteq R_{\Sigma,e_{|H'}}(E)$, because

$$R_{\Sigma,e_{|H'}}(E) = R_{\Sigma,e_{|H}}(E) \cup R_{\Sigma,e_{|H'\setminus H}}(E)$$

◊

Independence and monotonicity were introduced as properties of abductive diagnosis for the first time in [Bylander et al., 1992].

3.2 Properties of notions of diagnosis

In the next two chapters, various notions of diagnosis are compared, and their diagnostic characteristics explored. In this section, several properties of notions of diagnosis and diagnostic components are discussed, from which we will benefit in the following chapters.

3.2.1 Ordering of notions of diagnosis

The two orderings to be defined below, shall be employed frequently in the comparison of notions of diagnosis.

Definition 3.26 (*restriction*)**.** *Let R and R' be two notions of diagnosis. Then, R is called a* restriction *of R', denoted by*

$$R \sqsubseteq R'$$

if for each $\Sigma \in \mathcal{S}$, $H \subseteq \Delta$, $E \subseteq \Phi$ it holds that: if $R_{\Sigma,e_{|H}}(E) = H'$, $H' \neq u$, then $R'_{\Sigma,e_{|H}}(E) = H'$.

Thus, if the restriction relation between two notions of diagnosis R and R' holds, then a diagnosis produced by R will also be a diagnosis by R'.

The notion of subdiagnostic relation is useful for characterizing the relative strictness in admitting defects to a diagnostic solution of notions of diagnosis.

Definition 3.27 (*subdiagnostic relation*)**.** *Let R and R' be two notions of diagnosis. The notion of diagnosis R is called* subdiagnostic *to R', denoted by*

$$R \trianglelefteq R'$$

if

$$R_{\Sigma,e_{|H}}(E) \subseteq R'_{\Sigma,e_{|H}}(E)$$

given that $R_{\Sigma,e_{|H}}(E), R'_{\Sigma,e_{|H}}(E) \neq u$, for each $\Sigma \in \mathcal{S}$, $H \subseteq \Delta$ and $E \subseteq \Phi$.

We shall employ the same symbol $\trianglelefteq$ to denote that the diagnostic solutions of some diagnostic component are a subset of those of another diagnostic component applied to the same diagnostic specification, i.e.

$$R_{\Sigma,e_{|H}} \trianglelefteq R'_{\Sigma,e_{|H'}}$$

iff $R_{\Sigma,e_{|H}}(E) \subseteq R'_{\Sigma,e_{|H'}}(E)$, for each set of observed findings E.

3.2.2 Accounting function and completeness

One frequently encountered means in formal diagnostic theories to determine the preferences among different diagnoses is the extent to which a diagnosis accounts for a given set of observed findings. Given that the set of defects D is a diagnosis, i.e. $D = R_{\Sigma,e_{|H}}(E)$, then $e(D)$ is the set of observable findings for D. Usually, only some of those observable findings have actually been observed, and are, therefore, accounted for by the diagnosis D. This aspect of diagnosis is formalized by means of a function, called the accounting function, that gives the set of observed findings for which a set of defects accounts.

Definition 3.28 (*accounting function*)**.** *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem, and let $D \subseteq \Delta$ be a consistent set of defects. The* set of observed findings accounted for by D, *denoted by $A(D, E)$, is defined as*

$$A(D, E) = e(D) \cap E$$

Furthermore, it is defined that $A(u, E) = \varnothing$, where u denotes the undefined defects symbol.

Note that the set $A(R_{\Sigma,e_{|H}}(E), E)$ denotes the part of the set of observed findings E accounted for by the diagnosis $R_{\Sigma,e_{|H}}(E)$. A disadvantage of this definition is that it relates a diagnosis to a set of observed findings E in terms of the applied evidence function e only. It tells nothing about the diagnostic interpretation of a given evidence function e and set of observed findings E. In the literature, the term 'explained by' is often employed, instead of 'accounted for', to refer to similar concepts. As the term 'explained by' seems more appropriate for clarification of a diagnosis in terms of a notion of diagnosis employed, we prefer the more neutral term used in the definition. An example of the use of the accounting function follows.

Example 3.18. Consider the nonmonotonic evidence function e that is given by a bottom-up partial specification $\tilde{e}$, where $\tilde{e}(d_1) = \{f_1\}$, $\tilde{e}(d_2) = \{f_2, f_3\}$, $\tilde{e}(\{d_1, d_2\}) = \{f_1, f_4\}$, and $\tilde{e}(\neg d_i) = \varnothing$, $i = 1, 2$. Using the notion of diagnosis introduced in Example 3.14, but now assumed to be applicable to any possible diagnostic specification, the following diagnosis is obtained for the diagnostic problem $\mathcal{P} = (\Sigma, E)$, where $E = \{f_1, f_2, f_3\}$:

$$S_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$$

Intuitively, this result can be interpreted as follows. According to the evidence function e, if d_1 and d_2 co-occur, the expected set of observable findings is $\{f_1, f_4\}$. This is an example of augmented cancellation. However, the set of observed findings is equal to $\{f_1, f_2, f_3\}$. The notion of diagnosis S attempts to interpret this set. (Some other notion of diagnosis may refuse to interpret this set of observed findings, and give the undefined defects symbol u as a result.) The interpretation adopted is that the knowledge base (evidence function) is not completely accurate, and that d_1 and d_2 may sometimes lack interaction. The notion of diagnosis is capable of accounting for every observed finding by assuming e to be interaction free, because $e(\{d_1\}) \cup e(\{d_2\}) = \{f_1, f_2, f_3\}$. However, $A(\{d_1, d_2\}, \{f_1, f_2, f_3\}) = \{f_1\}$, i.e. in terms of the evidence function e, the diagnostic solution $\{d_1, d_2\}$ accounts only for the finding f_1 in E. For $E' = \{f_1, f_2\}$, the set of observed findings accounted for by the diagnosis $S_{\Sigma,e_{|\{d_1,d_2\}}}(E') = \{d_1\}$ is equal to $A(\{d_1\}, \{f_1, f_2\}) = \{f_1\}$; hence, in this case, f_2 cannot be accounted for. $\Diamond$

This example is quite distinct from the usual examples in the literature on diagnosis; it shows that it may still be possible to come up with a diagnosis when notions of diagnosis as discussed in Chapter 2 fail, using a more flexible notion of diagnosis. From the discussion above, it is tempting to consider designing an accounting function in which the notion of diagnosis is incorporated. However, this would require defining an accounting function for every notion of diagnosis, an approach which shall not be pursued.

For a diagnostic specification with an interaction-free set of defects, it is useful to isolate individual defects that are essential in their contribution to a diagnosis. This information is obtained by the accounting function A.

Definition 3.29 (*essentially accounting defect*)**.** *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem with inferaction-free set of defects Δ, and let $D \subseteq \Delta$, be a consistent set of defects. Then, a defect $d \in D$ is called* essentially accounting (with respect to D and E) *if $A(D', E) \neq A(D' \cup \{d\}, E)$, for each $D' \subseteq D \backslash \{d\}$.*

In the next chapter, in the analysis of several different notions of diagnosis, one of the differences among notions of diagnosis will be the extent to which observed findings must be accounted for by a diagnosis. Consider, for example, a problem $\mathcal{P} = (\Sigma, E)$ with an exhaustive diagnostic specification, such that $e(D) = \Phi$, for some $D \subseteq \Delta$. Then, the equality $A(D, E) = E$ is always satisfied for some $D \subseteq \Delta$, i.e. all sets of observed findings can, in principle, be accounted for by some set of defects. As discussed in Section 2.2.3, this property holds for the set-covering theory of diagnosis (the 'explanation existence theorem', [Peng & Reggia, 1990]). Generally speaking, we consider this condition to be unnecessarily restrictive. A more liberal constraint on a notion of diagnosis is yielded by the concept of Φ-completeness.

Definition 3.30 (Φ-*completeness*). *A notion of diagnosis R is called* Φ-complete *if for each diagnostic specification $\Sigma \in \mathcal{S}$ for which R_Σ is defined, and each set of observed findings $E \subseteq \Phi$, it holds that*

$$e(R_{\Sigma,e_{|H}}(E)) \supseteq E$$

if $R_{\Sigma,e_{|H}}(E) \neq u$, for some $H \subseteq \Delta$; otherwise, it is called Φ-incomplete.

Φ-completeness does not require the existence of a diagnosis for every set of observed findings; a diagnosis may also be undefined. But if a diagnosis is defined, all observed findings must be accounted for. Φ-completeness is similar to the covering condition in the theory of abductive diagnosis by Console and Torasso (cf. Section 2.2.2).

From the definition above, the next proposition follows.

Proposition 3.10. *A notion of diagnosis R is Φ-complete iff for each set of observed findings $E \subseteq \Phi$ there exists a set of defects $H \subseteq \Delta$, such that $E = A(R_{\Sigma,e_{|H}}(E), E)$ if $R_{\Sigma,e_{|H}}(E) \neq u$.*

Proof. Observe that $e(R_{\Sigma,e_{|H}}(E)) \supseteq E$ iff $e(R_{\Sigma,e_{|H}}(E)) \cap E = E$. $\Diamond$

As illustrated in Example 3.18, Φ-completeness as a requirement may be too strong for many perfectly acceptable notions of diagnosis. In some domains, for example medicine, only a small fraction of the entire domain is represented in a knowledge base. Suppose that a finding in a set of observed findings cannot be accounted for, although there is a defect in the knowledge base associated with the observed finding. This may mean that some defects outside the represented domain must be responsible for the unaccounted findings. In addition, no real-world knowledge base is perfect; the failure to account for some observed finding may be caused by inaccurately modelled interactions among defects. In both cases, the notion of diagnosis may still attempt to establish a diagnosis, for example, by deriving as many defects as possible.

Φ-incompleteness occurs in hypothetico-deductive diagnosis (See Chapter 4), for which it is generally accepted that not all findings can be accounted for.

The notion of Φ-completeness is a property with regard to the set of findings Φ. A similar property can be formulated for the set of defects Δ.

Definition 3.31 (Δ-*completeness*). *A notion of diagnosis R is called* Δ-complete *if for each diagnostic specification $\Sigma \in \mathcal{S}$ for which R_Σ is defined, it holds that for each consistent set $D \subseteq \Delta$, there exist a set of defects $H \subseteq \Delta$ and a set of observed findings $E \subseteq \Phi$, such that $R_{\Sigma,e_{|H}}(E) = D$; otherwise, R is called Δ-incomplete.*

This means that the notion of diagnosis is capable of diagnosing any meaningful combination of defects, dependent on the diagnostic specification given. In Example 3.14, a notion of diagnosis is discussed which is Δ-incomplete for any interaction-free diagnostic specification, because such specifications may have evidence functions e with $e(d) = \{f, \neg f\}$, which cannot be applied for establishing a diagnosis using the notion of diagnosis discussed in the example.

3.2.3 Similarity of diagnostic components

In Section 3.1.4, the mathematical framework for defining notions of diagnosis was introduced. A notion of diagnosis was considered to consist of a collection of diagnostic components, where each component could be viewed as an inference relation between arbitrary observed findings and a set of defects, taken as a diagnosis. In this section, the possible relationships between different diagnostic components will be investigated. Diagnostic components that account for the same set of observed findings are of particular interest, because such components can be considered equivalent. Under certain conditions, it might be acceptable to use only one of the diagnostic components from a set of equivalent diagnostic components. Furthermore, it might be possible to restrict attention to a set of equivalent diagnostic components for computing diagnoses that satisfy certain preference conditions, such as subset minimality. In the next section, relationships between individual diagnostic solutions will be investigated instead of diagnostic components as a whole.

In the literature on formal theories of diagnosis, comparisons with regard to the results of applying a notion of diagnosis to a set of observed findings, is only made with regard to the resulting sets of defects, interpreted as a diagnosis (cf. [Console et al., 1989; Peng & Reggia, 1990; Reiter, 1987]). Our approach, in which testing a diagnostic hypothesis H involves examining part of a knowledge base (the evidence function $e_{|H}$ in the framework), makes it possible to examine relationships between the functions $R_{\Sigma,e_{|H}}$, thus lifting diagnostic problem solving to a more abstract level. The kind of relationships examined in the other literature on diagnosis, will be introduced in the following section, but again generalized, because the implicit assumption of monotonicity of the 'diagnostic' knowledge base (evidence function e) is not taken to hold in general in our framework.

In the following definition, an important condition under which diagnostic components can be considered equivalent is introduced.

Definition 3.32 (*component similarity*). *Let R be a notion of diagnosis, and let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. Let $P_{R_\Sigma} = \{R_{\Sigma,e_{|H}} \mid H \subseteq \Delta,\ H$ consistent$\}$. The* component similarity relation $\equiv$ *with respect to R is then defined as follows*

$$\equiv\ \subseteq P_{R_\Sigma} \times P_{R_\Sigma}$$

where $R_{\Sigma,e_{|H}} \equiv R_{\Sigma,e_{|H'}}$ *iff* $A(R_{\Sigma,e_{|H}}(E), E) = A(R_{\Sigma,e_{|H'}}(E), E)$ *for each set of observed findings* $E \subseteq \Phi$.

The component similarity relation is an equivalence relation because it is symmetric and transitive. Hence, the relation $\equiv$ partitions the set of diagnostic components for a given diagnostic specification Σ into equivalence classes, denoted by

$$[R_{\Sigma,e_{|H}}]_{\equiv} = \{R_{\Sigma,e_{|H'}} \mid R_{\Sigma,e_{|H}} \equiv R_{\Sigma,e_{|H'}}\}$$

The set of equivalence classes of P_{R_Σ} with respect to $\equiv$ is called the *quotient set* of P_{R_Σ} and will be denoted by $P_{R_\Sigma}/\equiv$.

Example 3.19. Consider the notion of diagnosis I, which expresses that every defect that has at least one finding in common with the set of observed findings should be included in a diagnosis; function values $e(D)$, with D a non-singleton set of defects are ignored by this notion of diagnosis (i.e. it only considers diagnoses obtained by examining findings associated with individual defects). A consequence of this definition is that the independence assumption is satisfied. Here we assume that I is restricted to diagnostic specifications with an evidence function that is defined for an interaction-free set of defects Δ. A diagnosis $I_{\Sigma,e_{|H}}(E)$ is then defined as follows:

$$I_{\Sigma,e_{|H}}(E) = \{d \in H \mid e(d) \cap E \neq \varnothing\}$$

Now, consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$ with $\Delta_P = \{d_1, d_2, d_3\}$, $\Phi_P = \{f_1, f_2, f_3\}$, and the bottom-up partial specification $\tilde{e}$ for e:

$$\tilde{e}(d) = \begin{cases} \{f_1, f_2\} & \text{if } d = d_1 \\ \{f_1, f_3\} & \text{if } d = d_2 \\ \{f_2\} & \text{if } d = d_3 \\ \varnothing & \text{if } d = \neg d_i, i = 1, \ldots, 3 \end{cases}$$

The definition for the notion of diagnosis I yields, for example, the following diagnostic components (we limit our attention to components $I_{\Sigma,e_{|H}}$ for which $H \subseteq \Delta_P$):

$$\begin{aligned} I_{\Sigma,e_{|\{d_1\}}} = \{&(\{f_1\}, \{d_1\}), \\ &(\{f_2\}, \{d_1\}), \\ &(\{f_1, f_2\}, \{d_1\}), \\ &(\{f_1, f_3\}, \{d_1\}), \\ &(\{f_2, f_3\}, \{d_1\}), \\ &(\{f_1, f_2, f_3\}, \{d_1\})\} \end{aligned}$$

and

$$\begin{aligned} I_{\Sigma,e_{|\{d_1,d_3\}}} = \{&(\{f_1\}, \{d_1\}), \\ &(\{f_2\}, \{d_1, d_3\}), \\ &(\{f_1, f_2\}, \{d_1, d_3\}), \\ &(\{f_1, f_3\}, \{d_1\}), \\ &(\{f_2, f_3\}, \{d_1, d_3\}), \\ &(\{f_1, f_2, f_3\}, \{d_1, d_3\})\} \end{aligned}$$

These two diagnostic components are equivalent with respect to $\equiv$, because, for example

$$\begin{aligned}A(I_{\Sigma,e_{|\{d_1\}}}(\{f_1, f_3\}), \{f_1, f_3\}) &= \{f_1\} \\ &= A(I_{\Sigma,e_{|\{d_1,d_3\}}}(\{f_1, f_3\}), \{f_1, f_3\})\end{aligned}$$

The following equivalence classes can be distinguished to constitute the quotient set $P_{I_\Sigma}/{\equiv}$:

$$\begin{array}{l}\{I_{\Sigma,\varnothing}\} \\ \{I_{\Sigma,e_{|\{d_1\}}}, I_{\Sigma,e_{|\{d_1,d_3\}}}\} \\ \{I_{\Sigma,e_{|\{d_2\}}}\} \\ \{I_{\Sigma,e_{|\{d_3\}}}\} \\ \{I_{\Sigma,e_{|\{d_1,d_2\}}}, I_{\Sigma,e_{|\{d_2,d_3\}}}, I_{\Sigma,e_{|\{d_1,d_2,d_3\}}}\}\end{array}$$

Although similar diagnostic components do not always yield the same diagnosis, they account for the same set of observed findings. The following examples are illustrative in this respect: $I_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_1\}) = \{d_1, d_2\}$ and $I_{\Sigma,e_{|\{d_2,d_3\}}}(\{f_1\}) = \{d_2\}$. This effect is caused by the fact that diagnostic components are restricted with respect to the set of defects that possibly can constitute a diagnosis. $\Diamond$

The example above suggests that some additional structure may be revealed among the elements of an equivalence class by considering the defects in the diagnostic solution obtained by a diagnostic component for a given set of observed findings. As has been defined above, diagnostic components may be partially ordered by the relation $\trianglelefteq$, i.e. by set inclusion of diagnostic solutions.

Example 3.20. Consider the quotient set $P_{I_\Sigma}/\equiv$ computed in the previous example. The subdiagnostic relation $\trianglelefteq$ consists of $I_{\Sigma,e_{|\{d_1\}}} \trianglelefteq I_{\Sigma,e_{|\{d_1,d_3\}}}$, because $I_{\Sigma,e_{|\{d_1\}}}(E) \subseteq I_{\Sigma,e_{|\{d_1,d_3\}}}(E)$, for each $E \subseteq \Phi$. In addition, $I_{\Sigma,e_{|\{d_1,d_2\}}} \trianglelefteq I_{\Sigma,e_{|\{d_1,d_2,d_3\}}}$ and $I_{\Sigma,e_{|\{d_2,d_3\}}} \trianglelefteq I_{\Sigma,e_{|\{d_1,d_2,d_3\}}}$. Both diagnostic components $I_{\Sigma,e_{|\{d_1,d_2\}}}$ and $I_{\Sigma,e_{|\{d_2,d_3\}}}$ are minimal elements (with respect to $\trianglelefteq$) of the equivalence class $[I_{\Sigma,e_{|\{d_1,d_2,d_3\}}}]_\equiv$. This means that a diagnosis obtained from these components is, for any set of observed findings, minimal with respect to set inclusion. However, note that there may be some other diagnostic component included in another equivalence class, that can produce a diagnosis that is still smaller. For example, for the set of observed findings $E = \{f_1\}$, the following are among the possible diagnoses: $I_{\Sigma,e_{|\{d_1\}}}(E) = \{d_1\}$ and $I_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$, while

$$A(I_{\Sigma,e_{|\{d_1\}}}(E), E) = A(I_{\Sigma,e_{|\{d_1,d_2\}}}(E), E)$$

The minimal diagnostic component $I_{\Sigma,e_{|\{d_1,d_2\}}}$ produces a minimal diagnosis with respect to the equivalence class to which it belongs, but there exist diagnostic components in I_Σ that produce smaller diagnoses. $\Diamond$

3.3 Selection of diagnostic solutions

In the previous section, individual diagnoses were taken as a basis for the comparison of diagnostic components. However, as demonstrated in Example 3.20, comparison of individual diagnoses is justified by its own rights, because under certain conditions not all defects constituting a diagnosis may be required to account for the observed findings. In this section, the existence of several alternative diagnoses for a diagnostic problem is therefore investigated, taking the diagnostic notions introduced in the previous section as a starting point. Here, the accounting function A is taken as a basis for comparison of diagnoses; other functions could be employed in a similar fashion.

3.3.1 Equal-accountability

Basically, alternative diagnoses are obtained by applying different diagnostic components from a notion of R-diagnosis to a diagnostic problem $\mathcal{P}$, and selecting only those diagnoses that satisfy a given selection criterion. This process of diagnosis selection has been introduced in Chapter 1. Depending on the notion of diagnosis R chosen, the result is a set of alternative diagnoses for a given diagnostic problem $\mathcal{P}$. A suitable starting point for selecting those diagnoses that are of interest is to impose structure on this set by investigating possible relationships among the individual diagnoses. One of those relationships is the extent to which observed findings are accounted for.

Definition 3.33 (*equal-accountability*)**.** *Let R be a notion of diagnosis, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem for which R is defined. Furthermore, let*

$$P_{R_\Sigma}(E) = \{R_{\Sigma,e_{|H}}(E) \mid H \subseteq \Delta,\ R_{\Sigma,e_{|H}}(E) \neq u\}$$

denote the set of all diagnoses from the notion of diagnosis R. An equal-accountability relation $\doteq$ *is then defined as a binary relation $\doteq\, \subseteq P_{R_\Sigma}(E) \times P_{R_\Sigma}(E)$, such that $D \doteq D'$ iff $A(D,E) = A(D',E)$, with $D, D' \in P_{R_\Sigma}(E)$.*

An equal-accountability relation defines an equivalence relation on the set of all possible diagnoses for a given diagnostic problem $\mathcal{P}$. In contrast to the previous section, where equivalence of diagnostic components was investigated, here equivalence of diagnostic solutions for a specific diagnostic problem is defined. Note that the following, straightforward proposition holds.

Proposition 3.11. *Let R be a notion of diagnosis, and let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification for which R is defined. Then, $R_{\Sigma,e_{|H}} \equiv R_{\Sigma,e_{|H'}}$ iff for each set of observed findings $E \subseteq \Phi$: $R_{\Sigma,e_{|H}}(E) \doteq R_{\Sigma,e_{|H'}}(E)$.*

Proof. Straight from the definitions. $\Diamond$

The set of all equivalence classes, called the *quotient set* of $P_{R_\Sigma}(E)$ with respect to $\doteq$, will be denoted by $P_{R_\Sigma}(E)/\doteq$; an equivalence class is denoted by $[R_{\Sigma,e_{|H}}(E)]_{\doteq}$. All equivalence classes in the quotient set have a similar structure; each diagnosis in an equivalence class accounts for the same set of observed findings as the other diagnoses in the same equivalence class. In the literature on diagnosis, often only one of the equivalence classes is considered, viz. the class $[R_{\Sigma,e_{|H}}(E)]_{\doteq}$, where for each $D \in [R_{\Sigma,e_{|H}}(E)]_{\doteq}$ it holds that $A(D,E) = E$ (without necessarily assuming that R is Φ-complete). Moreover, the diagnoses in this particular equivalence class are often sets of positive defects only, a consequence of the restriction to (general) Horn clause logic in many theories of abductive diagnosis. The following definition provides a generalization of the notion of multiple diagnosis as defined in the literature.

Definition 3.34 (*multiple diagnosis*). *Let R be a notion of diagnosis, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. A* multiple diagnosis *with respect to sets of observed findings E and E', where $E' \subseteq E$, denoted by $\mathcal{D}_R(E, E')$, is defined by*

$$\mathcal{D}_R(E, E') = \{D \in P_{R_\Sigma}(E) \mid A(D, E) = E'\}$$

Note that, despite the use of the singular noun 'diagnosis' in the phrase 'multiple diagnosis', a multiple diagnosis is actually a collection of alternative diagnoses. Multiple diagnoses can, in fact, also be characterized in terms of equal-accountability equivalence classes. Hence, a multiple diagnosis can be characterized in two ways. In the first place, it can be determined by the set of observed findings $E' = A(D, E)$, as defined above. In the second place, it can be determined by means of a single hypothesis $H \subseteq \Delta$ for $R_{\Sigma, e_{|H}}(E) \in [R_{\Sigma, e_{|H}}(E)]_{\doteq}$, i.e. the following proposition holds.

Proposition 3.12. *Let R be a notion of diagnosis, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. Then,*

$$\mathcal{D}_R(E, E') = [R_{\Sigma, e_{|H}}(E)]_{\doteq}$$

for some $H \subseteq \Delta$.

Proof. Since every set of defects $D \in \mathcal{D}_R(E, E')$ must be a diagnosis $R_{\Sigma, e_{|H}}(E)$, for some $H \subseteq \Delta$, there is precisely one equal-accountability equivalence class in which this diagnosis $R_{\Sigma, e_{|H}}(E)$ is included. Hence, the other elements from the equivalence class are also included in $\mathcal{D}_R(E, E')$. ◊

By assuming certain properties for notions of diagnosis, such as Δ-monotonicity, some structure can be imposed on multiple diagnosis sets.

3.3.2 Minimal diagnosis

As discussed in Chapter 2, in the parsimonious covering theory by Peng and Reggia, a notion of 'parsimony' is used to select only the most plausible diagnoses from the set of all gathered diagnoses. Parsimony is thus interpreted as a notion of plausibility. Parsimony assumptions are considered important by researchers, because they provide a practical means for reducing the large number of different diagnoses produced by many notions of diagnosis. Several notions of parsimony have been defined. The most interesting of the domain-independent criteria seems to be minimality according to set inclusion. When an evidence function is interaction free, minimality according to set inclusion means that a minimal set of defects accounting for the observed findings is taken as a diagnosis; no redundant defects are allowed (cf. Chapter 2). This criterion seems acceptable when defects are at least partially interaction free, and when small sets of defects are more likely as diagnoses than large sets of defects. In the extreme case that an evidence function is injective (excluding inconsistent sets of defects), none of the defects will be redundant, because every set of observable findings $e(D)$ will be unique. Interestingly, it is not evident from the literature that many of the properties described fail to hold for the general case, where many interactions exist between defects. In our approach, minimality assumptions are less important, because there are no fixed notions of diagnosis, as in the diagnostic theories presented in the previous chapter. Just by choosing a more restricted notion of diagnosis, it might be possible to reduce the number of different diagnoses without employing minimality as a selection criterion. However, it is possible to augment the framework with the notion of minimal diagnosis; we do so below.

In the following definition the notion of minimal diagnosis is introduced.

Definition 3.35 (*minimal diagnosis*)**.** *Let R be a notion of diagnosis, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. Furthermore, let $S \subseteq P_{R_\Sigma}(E)$ be a set of diagnoses. A diagnosis $D \in S$ is called a* minimal diagnosis *with respect to S if for each $D' \in S$: $D' \not\subset D$; otherwise it is called* non-minimal *with respect to S.*

When taking for the set of diagnoses S some equal-accountability equivalence class $[R_{\Sigma,e_{|H}}(E)]_{\doteq}$, every minimal diagnosis that is a member of this equivalence class accounts for the same set of findings as $R_{\Sigma,e_{|H}}(E)$. For multiple diagnosis, the set of minimal diagnoses is called minimal multiple diagnosis, because only the minimal diagnoses from a multiple diagnosis are selected.

Definition 3.36 (*minimal multiple diagnosis*)**.** *Let R be a notion of diagnosis, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. Let $\mathcal{D}_R(E, E')$ be a multiple diagnosis for the notion of diagnosis R. Then, a* minimal multiple diagnosis*, denoted by $\mathcal{D}_R^{\subseteq}(E, E')$, is defined as follows:*

$$\mathcal{D}_R^{\subseteq}(E, E') = \{D \in \mathcal{D}_R(E, E') \mid D \textit{ is a minimal diagnosis w.r.t. } \mathcal{D}_R(E, E')\}$$

Since the notion of minimal diagnosis can be interpreted as expressing that a small diagnostic solution D accounting for the observed findings E is more plausible than a larger (w.r.t. set inclusion) diagnostic solution D', accounting for the same findings, minimal multiple diagnosis can be assumed to contain only plausible diagnoses. Note that none of the defects from a given diagnosis element in $\mathcal{D}_R^{\subseteq}(E, E')$ can be removed, without causing the resulting set of defects to fail to be a diagnosis. It is easily seen that each member of a minimal multiple diagnosis is the result of a $\trianglelefteq$-minimal diagnostic component $R_{\Sigma,e_{|H}}$ of an equivalence class $[R_{\Sigma,e_{|H}}]_{\equiv}$.

Proposition 3.13. *Let R be a notion of diagnosis, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem, and let $E' \subseteq E$. Then, for each $D \in \mathcal{D}_R^{\subseteq}(E, E')$ there exists a diagnostic component $R_{\Sigma,e_{|H}}$ with $D = R_{\Sigma,e_{|H}}(E)$, such that $R_{\Sigma,e_{|H}} \trianglelefteq R_{\Sigma,e_{|H'}}$ for each $R_{\Sigma,e_{|H'}} \in [R_{\Sigma,e_{|H}}]_{\equiv}$.*

Proof. Let $D = R_{\Sigma,e_{|H}}(E)$. Suppose that D is not the result of any $\trianglelefteq$-minimal diagnostic component $R_{\Sigma,e_{|H'}}$. There exists a $\trianglelefteq$-minimal diagnostic component $R_{\Sigma,e_{|H''}}$, with $R_{\Sigma,e_{|H''}} \trianglelefteq R_{\Sigma,e_{|H}}$, such that, either:

(a) $R_{\Sigma,e_{|H''}}(E) = D$, or

(b) $R_{\Sigma,e_{|H''}}(E) \subset D$.

In case (a), $R_{\Sigma,e_{|H''}}(E)$ is included in the minimal multiple diagnosis; a contradiction. In case (b), D is not a minimal diagnosis; again a contradiction. It is concluded that each $D \in \mathcal{D}^{\subseteq}_{R}(E, E')$ is the result of a $\trianglelefteq$-minimal diagnostic component. ◇

However, it is not true in general that each $\trianglelefteq$-minimal diagnostic component results in a minimal diagnosis in $\mathcal{D}^{\subseteq}_{R}(E, E')$, simply because such a diagnosis may not account for the same set of observed findings as the members of $\mathcal{D}^{\subseteq}_{R}(E, E')$.

If the notion of diagnosis R satisfies the independence assumption, we have the following result.

Proposition 3.14. *Let R be a notion of diagnosis that satisfies the independence assumption, and let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem, with Δ interaction free. Let $\mathcal{D}^{\subseteq}_{R}(E, E')$ be a minimal multiple diagnosis for $\mathcal{P}$. If $A(\{d\}, E) \subseteq A(\{d'\}, E)$ holds, for $d, d' \in \Delta$ and $d \neq d'$, such that $d' \in D$, $D \in \mathcal{D}^{\subseteq}_{R}(E, E')$, then $d \notin D$.*

Proof. Suppose that $d, d' \in D$, and $D \in \mathcal{D}^{\subseteq}_{R}(E, E')$. The diagnosis D accounts for the same set of observed findings as any $D' \in \mathcal{D}^{\subseteq}_{R}(E, E')$, and is minimal. By the independence assumption, it follows that $D = \bigcup_{d \in H} R_{\Sigma,e_{|\{d\}}}(E)$, for some $H \subseteq \Delta$. Therefore, if $R_{\Sigma,e_{|H\backslash\{d\}}}(E) \neq u$, it holds that $R_{\Sigma,e_{|H\backslash\{d\}}}(E) = D\backslash\{d\}$. From $A(\{d\}, E) \subseteq A(\{d'\}, E)$, $A(D, E) = A(D', E)$, and the fact that Δ is interaction free, it follows that

$$A(D\backslash\{d\}) = A(D', E)$$

implying that D is non-minimal; contradiction. Hence, $d \notin D$. ◇

This result may be interpreted in various ways. Two different defects d and d' for which $A(\{d\}, E) = A(\{d'\}, E)$ may be viewed as conflicting defects, because these defects account for the same findings in E. Consequently, one may say that in the notion of minimal diagnosis, diagnostic solutions from the corresponding multiple diagnosis that contain conflicting defects are not admitted. This interpretation, however, is only valid for interaction-free sets of defects, and notions of diagnosis for which the independence assumption holds. A similar result, called the competing disorders theorem, was obtained by Peng and Reggia for their specific form of set-covering diagnosis [Peng & Reggia, 1990].

There is one important difference between an element from a minimal multiple diagnosis and a multiple diagnosis. Although an entire minimal multiple diagnosis provides useful information about alternatives that completely account for the findings observed, this may lead to a loss of information, as the following example demonstrates.

Example 3.21. Reconsider the notion of diagnosis I defined in Example 3.19. Consider the following diagnostic problem $\mathcal{P} = (\Sigma, E)$, where $\Sigma = (\Delta, \Phi, e)$ with $\Delta_P = \{d_1, d_2\}$, $\Phi_P = \{f_1, f_2, f_3\}$, $E = \{f_2, f_3\}$, and the following bottom-up partial specification (again restricted to Δ_P):

$$\tilde{e}(d) = \begin{cases} \{f_1, f_2\} & \text{if } d = d_1 \\ \{f_2, f_3\} & \text{if } d = d_2 \end{cases}$$

Then, we have $\mathcal{D}_I(E, E) = \{\{d_2\}, \{d_1, d_2\}\}$ and $\mathcal{D}^{\subseteq}_{I}(E, E) = \{\{d_2\}\}$. These sets of diagnoses are the result of the diagnostic components $I_{\Sigma,e_{|\{d_2\}}}$ and $I_{\Sigma,e_{|\{d_1,d_2\}}}$ combined, for

$\mathcal{D}_I(E, E)$, and $I_{\Sigma,e_{|\{d_2\}}}$ separately, for $\mathcal{D}_I^{\subseteq}(E, E)$. Note that the useful information that d_1 may also be considered as a possible defect is lost in $\mathcal{D}_I^{\subseteq}(E, E)$. Although there may be application domains in which removal of the set $\{d_1, d_2\}$ may be acceptable, this will not be true in all domains. The reader should observe that in this case, even the single diagnostic component $I_{\Sigma,e_{|\Delta_P}}$ provides more information. Hence, a notion of diagnosis consisting of a single diagnostic component $I_{\Sigma,e_{|\Delta_P}}$ might be preferred in this case. ◊

More specific notions of diagnosis are required to acquire more insight into the suitability of the framework to describe notions of diagnosis. In the next two chapters, several specific notions of diagnosis will be studied, using the basic concepts introduced in this chapter.

3.4 Comparison to related work

Above we have introduced a quite general framework for diagnostic problem solving. The main components of the framework are: (1) evidence functions, which correspond to diagnostic interpretations of knowledge bases of expert systems, and (2) notions of diagnosis that are used to interpret an evidence-function interpretation of a knowledge base for the purpose of diagnosis. The framework supports two different views on designing expert systems. On the one hand, given an evidence-function interpretation, a notion of diagnosis can be designed (or selected) that follows the meaning of the evidence function as closely as possible. On the other hand, applying a particular notion of diagnosis to solve a diagnostic problem implies that a specific (diagnostic) meaning is given to the associated evidence function by the notion of diagnosis.

Several researchers have developed other frameworks for diagnostic problem solving, with purposes similar to the framework introduced in this chapter, namely to characterize various forms of diagnostic problem solving. Most frameworks take specific formal notions of abductive diagnosis and consistency-based diagnosis as a foundation. The most important other research will be compared to our work.

As discussed in Chapter 2, Console and Torasso, [Console & Torasso, 1991], have proposed a spectrum of logical definitions of diagnosis by varying one parameter in their basic definitions: the set of observed findings E that must be covered by a diagnosis. A notion of diagnosis is obtained by choosing a *fixed* subset E' of the entire set of observed findings E, that must be covered by a diagnosis H, i.e. $\mathcal{R} \cup H \vDash E'$ must hold, where $\mathcal{R}$ stands for the logical representation of a model of normal and abnormal behaviour. The set of findings E^c that must be consistent with $\mathcal{R} \cup H$ cannot be chosen, but is fixed as the set of absent findings (although alternative definitions of this set appear in the literature). The consistency condition ensures that no finding is predicted (entailed) that has not been observed. Thus, the set E^c is determined by the set of observed findings E. Abductive diagnosis in the sense of Console and Torasso is obtained by choosing $E' = E$; in contrast, by taking the empty set for E', consistency-based diagnosis in the sense of Reiter is obtained [Console et al., 1989].

Ten Teije and Van Harmelen, [Ten Teije & Van Harmelen, 1994], have extended the spectrum of Console and Torasso's logical definitions of diagnosis by leaving the choice of the relations for defining the covering and consistency conditions open (one trivial possibility would be to choose the logical entailment relation as a basis for both relations), and by making it possible to choose an arbitrary decomposition of the set of observed findings E into E^c, the set that must be consistent with a diagnosis, and E', the set that must be covered by a diagnosis. As a consequence, the resulting framework is more flexible than the original framework, although it is still in the spirit of the original spectrum of logical definitions of diagnosis by Console and Torasso.

The framework presented in this chapter, which draws its inspiration from the set-theoretic approach to abductive diagnosis proposed by Reggia et al. [Reggia et al., 1983; Peng & Reggia, 1990], differs in several respects from the diagnostic frameworks based on logic mentioned above. Firstly, in the other frameworks, notions of diagnosis have been designed in close connection with specific domain models, such as causal models or models of structure and behaviour. In contrast, in our framework, there is no intimate connection between the theory and any of the existing conceptual models of diagnosis. In fact, the meaning of a knowledge base, described by means of an evidence function e, is completely separated from its diagnostic use. The underlying idea of our framework of diagnosis is that any given knowledge base may be applicable in some way for diagnostic problem solving. Given a specific knowledge base, the main problem then becomes the design of a notion of diagnosis that makes it possible to apply that knowledge base to solve diagnostic problems. Of course, it is usually desirable to define notions of diagnosis that closely mirror the meaning of a knowledge base. Secondly, where in the other frameworks, the modelled behaviour is usually monotonic, due to the monotonicity of the employed logical entailment relation, monotonicity is not a prerequisite in our framework. In Section 3.1, several possible meanings of evidence functions were discussed; several of these, although intuitively appealing, were nonmonotonic. Hence, the assumption made in other frameworks that the domain knowledge is monotonic may be too strong. Actually, in [Console et al., 1991] this restriction has also been alleviated by resorting to general Horn formulae (normal logic programs), but the adopted nonmonotonicity is very specific, namely that obtained by the completion semantics of normal logic programs [Clark, 1978]. Thirdly, in the other notions of diagnosis, the set of observed findings that must be accounted for is fixed for specific notions of diagnosis. For example, in the spectrum of logical definitions of diagnosis by Console and Torasso, the set of findings E^c that must be consistent with a diagnosis and the set of findings E that must all be covered by a diagnosis are kept fixed for specific notions of diagnosis. This requirement was abandoned in our framework; thus, it is now possible to distinguish between notions of diagnosis for which it is not required that a fixed collection of observed findings is accounted for, and notions of diagnosis in which a fixed set of observed finding must be accounted for.

In [Zadrozny, 1993] a general (meta) logical framework for abduction is presented, which bears some resemblance to our work. Abduction is described by means of a general explanation relation $e\!\!-$, which relates sets of explanations P to sets of well-formed formulae F. By providing different rules of inference, various explanation systems are obtained. The principles underlying the explanation relation $e\!\!-$ are thus similar to our notion of diagnosis R. However, Zadrozny is not particularly concerned with abductive diagnosis. The consequences with regard with diagnosis are not explored.

In the next two chapters, we undertake the study of specific examples of notions of diagnosis, where some of the requirements imposed by the other frameworks reappear.

Chapter 4

Analysis of Formal Theories of Diagnosis

In the previous chapter, a set-theoretical framework for the description and analysis of notions of diagnosis was developed. In this chapter, the framework will be applied to analyse the formal theories of diagnosis introduced in Chapter 2. By mapping the theories of diagnosis to our framework, a deeper understanding of their principles will be gained. Moreover, the assumptions underlying these formal theories will be clarified. In particular, the uncoupling of knowledge base (evidence function) and notion of diagnosis, which is an important feature of our framework, will yield insight into the choices with respect to each of these two aspects of a diagnostic theory.

The following method is adopted in the analysis. Firstly, the meaning of the knowledge base (specification of knowledge) in a diagnostic theory is captured by an evidence function e. Secondly, the diagnostic relationships between sets of observed findings and the interpreted knowledge base is expressed as a notion of diagnosis. We begin with an analysis of those theories of diagnosis in which the notion of causality plays an important part.

4.1 Causality and diagnosis

In several of the formal theories of diagnosis presented in Chapter 2, diagnostic problem solving was accomplished using a description of a problem domain in terms of cause-effect relationships. This is, for example, true for the abductive theory of diagnosis due to L. Console and P. Torasso, [Console et al., 1989], and the set-covering theory due to J. Reggia et al. [Reggia et al., 1983; Peng & Reggia, 1990]. As previously discussed, in the abductive theory of diagnosis, two different types of causality are distinguished. In the first type of causality, it is assumed that when a collection of defects is present all its associated findings will be present. (Although one might imagine that not every present finding need be observed, a distinction between present and observed findings, however, is not made in the theory.) This notion of causality will be called *strong causality*. In the second type of causality, it is assumed that a collection of defects (a cause) *may* produce everything between none and all of its associated observable findings (effects).

This notion of causality will be called *weak causality*, because the causal relationships between collections of defects and findings are uncertain. However, the exact nature, or the amount of uncertainty associated with a causal relationship, is assumed to be unknown. Thus, it cannot be precisely quantified by means of conditional probabilities $P(f|D)$, where f is a finding and D is a collection of defects; in fact, it is only known that $0 < P(f|D) < 1$. In addition to abductive diagnosis, the notion of causality that underlies the set-covering theory of diagnosis by Reggia et al. will be shown to express weak causality.

In diagnostic theories built upon notions of causality, it is usually assumed that every possible cause, or combination of causes, of a set of observed findings in the real world is included in a diagnostic specification. A consequence of this assumption is that a diagnosis is only established for a set of observed findings when *every* observed finding can be accounted for. In other words, these notions of diagnosis are Φ-complete. This holds in particular for the abductive theory of diagnosis, and the set-covering theory of diagnosis. It is a rather strong assumption, because practical diagnostic expert systems, as the human beings on whose knowledge these systems are based, are imperfect with respect to the real world. Firstly, because not all known defects from the real world are incorporated in the knowledge base; hence, it will not always be possible to account for every observed finding. Secondly, because there may be defects that are as yet not known, and therefore not included in the knowledge base, but that may produce findings that are included in the knowledge base. Φ-completeness will be more readily accepted in technical domains, where detailed descriptions of devices can be used to diagnose malfunction, than in the medical domain, where much is still open to research.

In this chapter it is assumed that abductive diagnosis is restricted to Horn logic, but extension to non-Horn formulae is possible. Given a Horn-formula restriction, not every possible set of observed findings can be accounted for, because abnormal or normal findings represented in conclusions of clauses are always positive literals; similarly, absent defects cannot be represented. In recent literature on abduction, the restriction to Horn logic has been alleviated by allowing for general Horn formulae with negation as failure, thus providing for greater flexibility [Console et al., 1991; Preist et al., 1994]. When appropriate, we shall remark on using deduction with negation as failure in the following. There seems to exist no natural restriction in dealing with negative information in the set-covering theory of diagnosis, although the literature does not mention this possibility.

The notions of causality which form the basis of abductive theory of diagnosis, and set-covering diagnosis, will be next analysed in terms of the framework presented in the previous chapter.

4.2 Analysis of the abductive theory of diagnosis

As discussed in Section 2.2.2, the abductive theory of diagnosis amounts to diagnostic problem solving using a logical specification $\mathcal{R}$ of a causal domain model. Elements in $\mathcal{R}$ have been called abnormality axioms. Two types of abnormality axioms were distinguished. In the first type, strongly causal knowledge was expressed by logical implication. By adding an assumption literal α to the antecedent of abnormality axioms, the causality relation was weakened, yielding the second type of causal relationship: the weakly causal relation. Diagnostic problem solving using causal knowledge was accomplished by reasoning about the presence of defects (originally, state literals S) and assumption literals, given a set of observed findings.

In the analysis, part of the interpretation of causal knowledge for the purpose of diagnosis will be captured by means of an evidence function; the other part will be captured by various notions of diagnosis. Dependent on the type of knowledge that must be interpreted, strongly or weakly causal knowledge, the emphasis will lie on either evidence function or notion of diagnosis. First, the interpretation of strongly and weakly causal knowledge will be discussed separately, to be combined at the end of this section.

4.2.1 The representation of strong causality

Recall that an abductive diagnostic problem is a pair $\mathcal{A} = (\mathcal{C}, E)$, where $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ is a causal specification, such that:

- Δ is a set of defect literals; positive defect literals are denoted by d, and negative defect literals are denoted by $\neg d$.

- Φ is a set of finding literals, where positive finding literals are denoted by f and negative finding literals by $\neg f$.

- $\mathcal{R}$ is a set of abnormality axioms, where each abnormality axiom is a Horn formula (extension to non-Horn formulae is possible) of one of the following two forms:

  $$d_1 \wedge \cdots \wedge d_n \rightarrow f$$
  $$d_1 \wedge \cdots \wedge d_n \rightarrow d$$

  with d_i, $i = 1, \ldots, n$, d and f ground atoms in predicate logic (or propositional symbols in propositional logic); the axioms are assumed to be acyclic, i.e. no cyclic dependencies between defects occur in $\mathcal{R}$ [Console et al., 1991]. The abnormality axioms represent strongly causal relationships; they are called *strong-causality axioms*.

- $E \subseteq \Phi$ is the set of observed finding literals.

Furthermore, the set E^c of complements of non-observed findings, based on the set of observed findings E, as discussed in Section 2.2.2, is required. Usually, we simply speak of a set of defects, instead of defect *literals*. Similarly, the phrases 'set of observable findings Φ' and 'set of observed findings E' will be employed.

Recall that a set of defects $H \subseteq \Delta$ is an *abductive diagnosis* iff for each $f \in E$: $\mathcal{R} \cup H \vDash f$ (covering condition), and $\mathcal{R} \cup H \cup E^c \nvDash \bot$ (consistency condition) (Definition 2.4). The notions of solution and abductive diagnosis coincide when the abnormality axioms express strongly causal relations only.

4.2.2 The interpretation of strong causality

For the purpose of the analysis, an abductive diagnostic problem $\mathcal{A}$ must be mapped to a diagnostic problem $\mathcal{P}$ in the diagnostic framework. The result of mapping strongly causal knowledge in the abductive theory of diagnosis to an evidence function is by definition an interpretation of that knowledge for the purpose of diagnosis. In this section, such a mapping τ, for which $\tau(\mathcal{A}) = \mathcal{P}$, will be designed.

To distinguish between the elements of a diagnostic problem $\mathcal{P}$, and the elements of an abductive diagnostic problem $\mathcal{A}$, the subscripts $\mathcal{P}$ and $\mathcal{A}$, respectively, will be attached to elements. The mapping τ is assumed to be bijective; it maps $\Delta_{\mathcal{A}}$ to $\Delta_{\mathcal{P}}$, $\Phi_{\mathcal{A}}$ to $\Phi_{\mathcal{P}}$ and $E_{\mathcal{A}}$ to $E_{\mathcal{P}}$. It is assumed that positive defects and findings are mapped to positive literals; similarly, negative defects and findings are mapped to negative literals. The set of axioms $\mathcal{R}$ is captured by an evidence function e with domain $\wp(\Delta_{\mathcal{P}})$, as follows. For each $D_{\mathcal{A}} \subseteq \Delta_{\mathcal{A}}$:

(1) if $\mathcal{R} \cup D_{\mathcal{A}}$ is satisfiable, then $e(D_{\mathcal{P}}) = \{\tau(f) \mid \mathcal{R} \cup D_{\mathcal{A}} \vDash f, f \in \Phi_{\mathcal{A}}\}$;

(2) otherwise, $e(D_{\mathcal{P}}) = \bot$.

where $D_{\mathcal{P}} = \tau(D_{\mathcal{A}})$. Condition (1) above is closely related to the covering condition. One possible interpretation of this condition is that it amounts to using causal knowledge to predict observable findings (cf. Definition 2.2). This aspect of the theory of abductive diagnosis is therefore interpreted in terms of an evidence function. Observe, however, that the consistency condition is not encoded within the evidence function. It is viewed in our framework as a condition that restricts possible diagnoses. Thus, it will be encoded within the notions of diagnosis yet to be discussed. If $\mathcal{R}$ includes *general* Horn formulae, i.e. $\mathcal{R}$ is a normal logic program, then instead of logical entailment, deduction with the negation as failure rule could be employed, as is done in [Console et al., 1991] and [Preist et al., 1994]. However, it is also possible to assume negative defects explicitly (they would be included in $D_{\mathcal{A}}$). The latter is more in agreement with the intended meaning of the notion of evidence function than the former.

Recall that different sets of abnormality axioms $\mathcal{R}$ may give rise to the same evidence function e.

Example 4.1. As an example, consider the following sets of abnormality axioms $\mathcal{R}$ and $\mathcal{R}'$:

$$\begin{aligned} \mathcal{R} &= \{d_1 \rightarrow d_2, d_2 \rightarrow f_1\} \\ \mathcal{R}' &= \{d_1 \rightarrow f_1, d_2 \rightarrow f_1\} \end{aligned}$$

For both sets, the resulting evidence function e can be given by the bottom-up partial specification $\tilde{e}$, where

$$\tilde{e}(D_{\mathcal{P}}) = \begin{cases} \{f_1\} & \text{if } D_{\mathcal{P}} = \{d_i\},\ i = 1, 2 \\ \varnothing & \text{if } D_{\mathcal{P}} = \{\neg d_i\},\ i = 1, 2 \end{cases}$$

◊

The evidence function e resulting from the transformation τ is monotonically increasing, and can always be represented by means of a bottom-up partial specification. This is a consequence of the monotonicity of logical entailment, and of the fact that the empty set of defects $D = \varnothing$ predicts nothing (in our framework: $e(\varnothing) = \varnothing$). The monotonicity would be lost if deduction with negation as failure is used in case $\mathcal{R}$ contains general Horn formulae. The empty set of defects would then not necessarily predict nothing. However, if negative defects are explicitly taken into account, monotonicity would be preserved, which will be assumed in the following. Note that if a Horn-formula restriction is adopted, it holds that for each $f \in e(D_{\mathcal{P}})$, with consistent $D_{\mathcal{P}} \subseteq \Delta_{\mathcal{P}}$, f is a positive finding ($f \in \Phi_{\mathcal{P},P}$). If general Horn or non-Horn formulae are allowed, (assumed) negative findings will be included in $e(D_{\mathcal{P}})$.

For ease of exposition, in the following, defects $\tau(d) \in \Delta_{\mathcal{P}}$ and defect literals $d \in \Delta_{\mathcal{A}}$ will not explicitly be distinguished; similarly, no difference is made between findings $\tau(f) \in \Phi_{\mathcal{P}}$ and finding literals $f \in \Phi_{\mathcal{A}}$.

4.2.3 Diagnostic notions of strong causality

As discussed in Section 2.2.2, Console and Torasso have presented two different versions of the consistency condition in the definition of abductive diagnosis. As far as known to the author, this difference has never been analysed. In the first version, it is assumed that all observable findings associated with a set of defects will be observed if all of the defects are present. If some finding is not included in the set of observed findings, it is assumed to be negative. The second version of the consistency condition is similar to the first version, except that predicate symbols are employed to represent test results. If no results of a particular test are available, the test results are assumed to be unknown; no negative findings are added to E^c with respect to the predicate symbol concerned. The interpretation adopted in Section 2.2.2 was that although the finding was predicted to be present, it was not observed, and was, therefore, assumed to be unknown. The two versions of the consistency condition give rise to two slightly different notions of diagnosis that will discussed subsequently.

We start by giving a motivating example.

Example 4.2. Consider the abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$, with causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, where $\Phi_P = \{f_1, f_2, f_3\} = E$, $\Delta_P = \{d_1, d_2\}$, and the set of abnormality axioms $\mathcal{R}$ comprises the following formulae:

$$\begin{array}{l} d_1 \to f_1 \\ d_1 \to f_2 \\ d_2 \to f_3 \\ d_2 \to d_1 \end{array}$$

The abnormality axioms are depicted as a causal net in Figure 4.1. The evidence function e resulting from $\tau(\mathcal{A})$ is defined by the following bottom-up partial specification:

$$\tilde{e}(D) = \begin{cases} \{f_1, f_2\} & \text{if } D = \{d_1\} \\ \{f_1, f_2, f_3\} & \text{if } D = \{d_2\} \\ \bot & \text{if } D = \{\neg d_1, d_2\} \\ \varnothing & \text{if } D = \{\neg d_i\}, i = 1, 2 \end{cases}$$

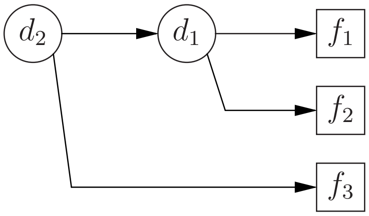


Figure 4.1: Set of abnormality axioms.

Recall that if D is syntactically inconsistent, then $e(D) = \perp$ by definition. When adopting the first version of the consistency condition, the diagnosis would be $\{d_1, d_2\}$, because

$$\mathcal{R} \cup \{d_1, d_2\} \vDash \{f_1, f_2, f_3\}$$

i.e. the covering condition is satisfied, and since $E^c = \varnothing$, the consistency condition is satisfied as well. Let SC denote the notion of diagnosis corresponding to the abductive theory of diagnosis for strong causality, adopting the first version of the consistency condition. Then, it must hold that

$$\mathrm{SC}_{\Sigma, e_{|\{d_1,d_2\}}}(\{f_1, f_2, f_3\}) = \{d_1, d_2\}$$

For $E \neq \{f_1, f_2, f_3\}$, however, no abductive diagnosis exists. Hence, it must hold that $\mathrm{SC}_{\Sigma, e_{|\{d_1,d_2\}}}(E) = u$ for $E \neq \{f_1, f_2, f_3\}$. ◊

It seems straightforward to formulate diagnosis based on causal knowledge in terms of our framework.

The first notion of diagnosis to be defined interprets observed findings by means of causal knowledge represented as an evidence function e. It takes the first version of the consistency condition into account. This notion of diagnosis accepts a diagnostic hypothesis H as a diagnosis iff every finding that causally follows from H, i.e. $e(H)$, has been observed. The notion is called strong-causality diagnosis, because the causal relations are interpreted as strongly causal relations.

Definition 4.1 (*strong-causality diagnosis*)**.** *The notion of* strong-causality diagnosis*, denoted by* SC*, is defined as follows:*

$$\mathrm{SC}_{\Sigma, e_{|H}}(E) = \begin{cases} H & \text{if } e_{|H}(H) = E \\ u & \text{otherwise} \end{cases}$$

for each diagnostic specification $\Sigma \in \mathcal{S}$ with monotonically increasing evidence function e, each set of observed findings $E \subseteq \Phi$, and each set of defects $H \subseteq \Delta$.

This notion of diagnosis is framed after abductive diagnosis. Obviously, in the exceptional situation that the restriction of the evidence function e to consistent sets is bijective, the notion of diagnosis SC would respect e (cf. Definition 3.22). Furthermore, since for any diagnosis $\mathrm{SC}_{\Sigma, e_{|H}}(E) \neq u$ it holds that

$$e(\mathrm{SC}_{\Sigma, e_{|H}}(E)) \supseteq E$$

SC is Φ-complete. This is a consequence of the fact that for any abductive diagnosis the covering condition must be satisfied. However, strong-causality diagnosis is Δ-incomplete, which is easily shown by a counter-example; a set of defects D, for which $e(D) = \{f, \neg f\}$ can never be a diagnosis.

Finally, note that the notion of strong-causality diagnosis does not fulfil the independence assumption, even if it is restricted to diagnostic specifications with interaction-free defects. The reason is that every diagnosis $\mathrm{SC}_{\Sigma, e_{|H}}(E)$ must account for every finding in E, which only would be possible if the function value $e(D)$ would be the same for each consistent $D \subseteq \Delta$. This notion of diagnosis is not Δ-monotonic, because it may hold that $\mathrm{SC}_{\Sigma, e_{|H}}(E) = H$, but $\mathrm{SC}_{\Sigma, e_{|H'}}(E) = u$, for $H \subset H'$ (which is again easily shown by means of a counter-example).

In the following proposition, abductive diagnosis using strong causality is proved equal to the notion of strong-causality diagnosis.

Proposition 4.1. *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem with strong-causality axioms $\mathcal{R}$, and let E^c be the set of findings for $\mathcal{A}$ according to the first version of the consistency condition. Let $\mathcal{P} = (\Sigma, E) = \tau(\mathcal{A})$ be the diagnostic problem obtained by the transformation τ, and let* SC *be the notion of strong-causality diagnosis. Then, a set of defects H is an abductive diagnosis for $\mathcal{A}$ iff $\mathrm{SC}_{\Sigma, e_{|H}}(E) = H$. Furthermore, a set of defects H is not an abductive diagnosis for $\mathcal{A}$ iff $\mathrm{SC}_{\Sigma, e_{|H}}(E) = u$.*

Proof. ($\Rightarrow$): Let H be an abductive diagnosis for $\mathcal{A}$. Then, according to the transformation τ and the covering condition, it holds that $e(H) \supseteq E$. Satisfaction of the consistency condition implies that for each finding $f \notin E$ (or $\neg f \in E^c$): $\mathcal{R} \cup H \nvDash f$, hence $f \notin e(H)$. It can be concluded that $e(H) = E$; therefore, $\mathrm{SC}_{\Sigma, e_{|H}}(E) = H$.

If H is not an abductive diagnosis, then: (a) the covering condition, or (b) the consistency condition fails to hold (possibly both). If the covering condition fails to hold, case (a), it holds by the definition of the mapping τ that $e(H) \neq E$. If the consistency condition fails to hold, case (b), $\mathcal{R} \cup H$ is unsatisfiable, i.e. – $\mathcal{R} \cup H$ does not consist of definite Horn formulae – it holds that $e(H) = \bot$, or there exists a finding $f \notin E$ such that $\mathcal{R} \cup H \cup \{\neg f\}$ is unsatisfiable, i.e. $\mathcal{R} \cup H \vDash f$. In both cases, $e(H) \neq E$. Therefore, $\mathrm{SC}_{\Sigma, e_{|H}}(E) = u$.

($\Leftarrow$): Let $\mathrm{SC}_{\Sigma, e_{|H}}(E) = H$, then by definition $e_{|H}(H) = E$. From the definition of the mapping τ it follows that $\mathcal{R} \cup H \vDash E$. Since $e(H)$ contains all findings f that are logically entailed by $\mathcal{R} \cup H$, it follows that for each $f \in \Phi$, $f \notin E$: $\mathcal{R} \cup H \nvDash f$. Therefore, $\mathcal{R} \cup H \cup E^c$ is satisfiable.

Now, let $\mathrm{SC}_{\Sigma, e_{|H}}(E) = u$. Then, $e_{|H}(H) \neq E$. Thus, either (a) there exists a finding $f \notin E$ such that $f \in e(H)$, (b) there exists a finding $f \in E$, such that $f \notin e(H)$, or (c) $e(H) = \bot$. Consider case (a). By the definition of τ, it follows that $\mathcal{R} \cup H \vDash f$, hence $\mathcal{R} \cup H \cup E^c$ is unsatisfiable; the consistency condition fails to hold. For case (b), it holds that $\mathcal{R} \cup H \nvDash f$, for $f \in E$, hence the covering condition fails to hold. Finally, for case (c) it holds that $\mathcal{R} \cup H \vDash \bot$. In all three cases, H is not an abductive diagnosis. ◊

Next, the framework is used to elucidate the nature of the abductive theory of diagnosis using the second version of the consistency condition. Recall that the main feature of the second version of the consistency condition is that predicate symbols are used to denote test results; the set of findings E^c used in the consistency condition contains only the negative counterparts of positive findings not included in the set of observed findings iff at least one finding concerning the test – i.e. finding atom with a predicate symbol denoting the test – is included in the set of observed findings.

Example 4.3. Reconsider the set of abnormality axioms $\mathcal{R}$ in Example 2.4:

$$\begin{aligned} d_1 &\rightarrow p(a) \\ d_1 &\rightarrow q(b) \\ d_2 &\rightarrow r(c) \end{aligned}$$

Let E be the following set of observed findings:

$$E = \{q(b), r(c)\}$$

representing results from two different tests. Using the second version of the consistency condition, the set

$$E^c = \{\neg q(e), \neg r(f)\}$$

is obtained. The first version of the consistency condition would have yielded:

$$E^c = \{\neg p(a), \neg p(d), \neg q(e), \neg r(f)\}$$

As a consequence, there exists no diagnosis in the first version of the theory (equivalently, it holds that $\mathrm{SC}_{\Sigma, e_{|H}}(E) = u$, for each $H \subseteq \Delta_P$), where application of the second version of the consistency condition yields as a diagnosis $D = \{d_1, d_2\}$. ◊

To enable encoding the results of tests, modelled in the abductive theory of diagnosis by means of predicate symbols, a special function is introduced.

Definition 4.2 (*subject function*)**.** *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let S be a set of elements, called* subjects. *A* subject function

$$s : \Phi \rightarrow S$$

assigns to each element $f \in \Phi$ a subject $s(f)$, such that $s(\neg f) = s(f)$.

A subject function can now be used to partition the set of observed findings into subsets of related findings, e.g. findings that are possible results of the same diagnostic test.

Definition 4.3 (*strong-causality with prediction diagnosis*)**.** *The notion of* strong-causality with prediction diagnosis, *denoted by* SCP, *is defined as follows:*

$$\mathrm{SCP}_{\Sigma, e_{|H}}(E) = \begin{cases} H & \text{if } e_{|H}(H) \supseteq E, \text{ and } \forall f, f' \in e_{|H}(H): \\ & \text{if } f \in E \text{ and } s(f) = s(f') \text{ then } f' \in E \\ u & \text{otherwise} \end{cases}$$

for each diagnostic specification $\Sigma \in \mathcal{S}$ with monotonically increasing evidence function e, each set of observed findings $E \subseteq \Phi$, and each set of defects $H \subseteq \Delta$, where s is a subject function.

This notion of diagnosis expresses that when one test result f is included in E, then

E should contain all other results of the same test, i.e. f', with $s(f) = s(f')$, if f' is included in $e_{|H}(E)$; otherwise, the set of observed findings is invalid, and no diagnosis can be determined. The notion of diagnosis SCP is Φ-complete (every diagnosis accounts for all observed findings). In contrast to SC, the notion of strong-causality with prediction diagnosis is Δ-complete. This result follows from the fact that for each consistent set of defects $D \subseteq \Delta$, it is satisfied that $e(D) \supseteq \varnothing$.

As for SC, the independence assumption does not hold for SCP, because a diagnosis is always undefined if some of the findings cannot be accounted for. For the same reason, SCP is not Δ-monotonic.

The transformation τ introduced above needs to be extended in order to correctly define a subject function for the diagnostic problem $\mathcal{P}$ that corresponds to the predicate representation of tests in an abductive diagnostic problem $\mathcal{A}$. The resulting transformation is denoted by τ'. The subject function is defined as follows. For each $f, f' \in \Phi_{\mathcal{A}}$: if $f = p(c)$, and $f' = p(d)$, then $s(\tau'(f)) = s(\tau'(f'))$.

The correspondence between the notion of strong-causality diagnosis with prediction and the second version of the abductive theory of diagnosis is established by the following proposition.

Proposition 4.2. *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem with strong-causality axioms $\mathcal{R}$, and let E^c be the set of finding literals for $\mathcal{A}$ according to the second version of the consistency condition. Let $\mathcal{P} = (\Sigma, E) = \tau'(\mathcal{A})$ be the diagnostic problem yielded by the transformation τ' with subject function s. Then, a set of defects H is an abductive diagnosis for $\mathcal{A}$ iff $\mathrm{SCP}_{\Sigma,e_{|H}}(E) = H$. A set of defects H is not an abductive diagnosis iff $\mathrm{SCP}_{\Sigma,e_{|H}}(E) = u$.*

Proof. ($\Rightarrow$): The only difference of this proof compared with the proof of Proposition 4.1 consists of the set E^c that is used in the consistency condition of abductive diagnosis. The remainder of the proof stays the same. The consistency condition expresses that not necessarily for each positive finding $f \notin E$: $\mathcal{R} \cup H \not\vDash f$; $\neg f \notin E^c$ will hold if $f = p(c)$ and for each d: $p(d) \notin E$. Suppose that $\mathcal{R} \cup H \vDash f$, $f \notin E$, and that H is a diagnosis. Then, by the transformation τ' it holds that $f \in e(H)$, but $f \notin E$, hence $e(H) \supseteq E$. The transformation τ' ensures that there exists no $f' \in E$, such that $s(f) = s(f')$.

($\Leftarrow$): The other side of the proof is a similar adaptation of the second part of the proof of Proposition 4.1. ◊

Above, two different notions of diagnosis have been defined for knowledge that is interpreted as expressing strongly causal relationships between defects and findings. The following proposition clarifies the relationships between these notions of diagnosis.

Proposition 4.3. *Let* SC *be the notion of strong-causality diagnosis and* SCP *be the notion of strong-causality diagnosis with prediction, then*

$$\mathrm{SC} \sqsubseteq \mathrm{SCP}$$

Proof. Simply observe that it may hold that $f \in e_{|H}(H)$, but $f \notin E$. Then always $\mathrm{SC}_{\Sigma,e_{|H}}(E) = u$, but it is possible that $\mathrm{SCP}_{\Sigma,e_{|H}}(E) \neq u$. Furthermore, if $\mathrm{SC}_{\Sigma,e_{|W}}(E) = H$ then $e_{|H}(H) = E$, thus $\mathrm{SCP}_{\Sigma,e_{|H}}(E) = H$. ◊

The essential difference between the abductive theories of diagnosis used as a basis for the notions of diagnosis SC and SCP was the consistency condition. As mentioned before, these semantical definitions of diagnosis place the covering condition at the level of evidence functions; the consistency condition is lifted to the level of notions of diagnosis. This seems to be the proper place for both conditions. The covering condition can be viewed as the causal interpretation of knowledge for the purpose of prediction; the consistency condition can be viewed as a constraint on possible diagnoses.

It is interesting to note that strong-causality diagnosis with prediction could serve as a basis for a theory of information gathering for causal notions of diagnosis, by means of which unknown, predicted, findings (i.e. findings $f \notin E$, but $f \in e(\mathrm{SCP}_{\Sigma, e_{|H}}(E))$ are requested from the user. Such a theory has been developed in the context of consistency-based diagnosis, (cf. [Reiter, 1987; Hou, 1994]), but not yet for abductive diagnosis.

4.2.4 The representation of weak causality

As discussed above, and in Section 2.2.2, the abductive theory of diagnosis also incorporates a notion of weak causality. This is obtained by the addition of assumption literals α to the individual abnormality axioms. This way, it can be expressed that a causal relation is uncertain. Hence, the abnormality axioms $\mathcal{R}$ of an abductive diagnostic problem $\mathcal{A}$ are of one of the following two forms:

$$\begin{array}{l} d_1 \wedge \cdots \wedge d_n \wedge \alpha_f \rightarrow f \\ d_1 \wedge \cdots \wedge d_n \wedge \alpha_d \rightarrow d \end{array}$$

These abnormality axioms are called *weak-causality axioms*.

To simplify matters, it is first assumed that assumption literals α_k are unique in every abnormality axiom. This simplification does not change the essential nature of the abductive, diagnostic theory of weak causality, but, as shall become clear, when this assumption is dropped it is necessary to include assumption literals in the evidence function e. Although Console and Torasso have nowhere in their papers explicitly stated that assumption literals are unique, it is likely (suggested by their examples) that they are supposed to be unique.[1] An advantage of the uniqueness of assumption literals is that the difference with respect to computed diagnoses between the two versions of the abductive theory, caused by two different definitions of the consistency condition, vanishes. This is illustrated by the following example.

Example 4.4. Consider the abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$, with causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, and abnormality axioms $\mathcal{R}$, in which five assumption literals α_i, $1 \leq i \leq 5$, are employed:

$$\begin{array}{lcl} d_1 \wedge \alpha_1 & \rightarrow & d_2 \\ d_1 \wedge \alpha_2 & \rightarrow & p(a) \\ d_1 \wedge \alpha_3 & \rightarrow & p(g) \\ d_1 \wedge \alpha_4 & \rightarrow & q(b) \\ d_2 \wedge \alpha_5 & \rightarrow & r(c) \end{array}$$

[1] Personal communication confirmed this impression.

Suppose that $q(e)$ and $r(f)$ are two other possible findings that may be observed. As discussed in Section 2.2.2, a solution H that satisfies the covering and consistency conditions consists of a set of defects and assumption literals. Recall that the set $D \subseteq H$, consisting of all defect literals in H, is called an abductive diagnosis. Now, consider the following set of observed findings

$$E = \{p(a), r(c)\}$$

Using the first version of the consistency condition for the construction of the set E^c, the set

$$E^c = \{\neg p(g), \neg q(b), \neg q(e), \neg r(f)\}$$

is obtained. A solution H satisfying both covering and consistency condition is $H = \{d_1, \alpha_1, \alpha_2, \alpha_5\}$; an alternative is $H' = \{d_1, d_2, \alpha_2, \alpha_5\}$. Hence, $D = \{d_1\}$ and $D' = \{d_1, d_2\}$ are two possible diagnoses. In contrast to the work by Console and Torasso, special initial 'states' are not distinguished, because this can be viewed as a minor refinement of the basic theory. Adopting the second version of the consistency condition:

$$E^c = \{\neg p(g), \neg r(f)\}$$

and as a possible solution $H = \{d_1, \alpha_1, \alpha_2, \alpha_4, \alpha_5\}$. An alternative solution is

$$H' = \{d_1, d_2, \alpha_2, \alpha_4, \alpha_5\}$$

The abductive diagnoses are: $D = \{d_1\}$ and $D' = \{d_1, d_2\}$. $\Diamond$

The diagnoses resulting from the two versions of the diagnostic theory of Console and Torasso are the same. This is no coincidence. It will always be possible to remove as many assumption literals (when the assumption literals are unique) from a solution H to restore consistency; therefore, the diagnoses, but not necessarily the solutions, will always coincide.

4.2.5 The interpretation of weak causality

The transformation τ introduced at the beginning of this chapter must be extended in order to deal with the assumption literals expressing weak causality. There are two possibilities. First, the abnormality axioms $\mathcal{R}$ could be translated to an evidence function e, where the assumption literals in a solution H are taken as defects, i.e. if for $f \in E$

$$\mathcal{R} \cup H \vDash f$$

and $\mathcal{R} \cup H$ is satisfiable, then $f \in e(H)$, where H is a set of defects, possibly including assumption literals α, i.e. $d = \tau''(\alpha)$, with extended transformation τ'', and d is a defect. Next, the notions of diagnosis introduced in the previous section for strong causality could be employed for diagnostic interpretation of the resulting evidence function e. Obviously, weak causality is then expressed at the level of the knowledge base; it is not a particular *diagnostic* interpretation of causal knowledge that leads to the concept of weak causality.

The second possibility would be to lift the notion of weak causality to the level of a notion of diagnosis, i.e. a special notion of diagnosis is designed that amounts to interpreting a knowledge base containing causal knowledge as being weakly causal in nature. The transformation τ''' that results from this approach is a simple adaptation of the transformation τ defined in Section 4.2.2. Let A denote the set of assumption literals in $\Delta_{\mathcal{A}}$. Then, $\mathcal{R}'$ is a set of abnormality axioms obtained by removing each assumption literal $\alpha \in A$ from each axiom in $\mathcal{R}$. The transformation τ''' is then defined in the same way as τ, except that $\mathcal{R}'$ replaces $\mathcal{R}$.

The abductive theory of diagnosis follows the first approach, because the same covering and consistency conditions are employed to define diagnosis for weakly causal knowledge, as well as for strongly causal knowledge. The second approach shall be studied in the next section; the resulting notion of diagnosis will then be compared to the first approach.

4.2.6 The notion of weak-causality diagnosis

The notion of diagnosis that corresponds to the abductive diagnosis using weak causality is called weak-causality diagnosis. It is defined as follows.

Definition 4.4 (*weak-causality diagnosis*). *The notion of* weak-causality diagnosis*, denoted by* WC*, is defined as follows:*

$$\mathrm{WC}_{\Sigma,e_{|H}}(E) = \begin{cases} H & \text{if } e_{|H}(H) \supseteq E \\ u & \text{otherwise} \end{cases}$$

for each diagnostic specification Σ with monotonically increasing evidence function e, each set of observed findings $E \subseteq \Phi$, and set of defects $H \subseteq \Delta$.

In the following example, the application of the notion of weak-causality diagnosis WC is illustrated.

Example 4.5. Consider the following abductive problem $\mathcal{A} = (\mathcal{C}, E)$, with causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, where the abnormality axioms $\mathcal{R}$ are as follows:

$$\begin{array}{lcl} d_1 \wedge \alpha_1 & \rightarrow & d_2 \\ d_2 \wedge \alpha_2 & \rightarrow & f_1 \\ d_1 \wedge \alpha_3 & \rightarrow & f_2 \end{array}$$

and $\Delta_{\mathcal{A},P} = \{d_1, d_2, \alpha_1, \alpha_2, \alpha_3\}$, $\Phi_{\mathcal{A},P} = \{f_1, f_2\}$, $A = \{\alpha_1, \alpha_2, \alpha_3\}$. The set of observed findings E is equal to $E = \{f_2\}$. The resulting evidence function conforms to the following bottom-up partial specification:

$$\tilde{e}(D) = \begin{cases} \{f_1, f_2\} & \text{if } D = \{d_1\} \\ \{f_1\} & \text{if } D = \{d_2\} \\ \bot & \text{if } D = \{d_1, \neg d_2\} \\ \varnothing & \text{if } D = \{\neg d_i\},\ i = 1, 2 \end{cases}$$

with $\Delta_{\mathcal{P},P} = \{d_1, d_2\} = \tau(\Delta_{\mathcal{A},P} \backslash A)$. The set $H = \{d_1, \alpha_3\}$ is a solution to $\mathcal{A}$, because the covering and consistency conditions are satisfied; a diagnosis is $D = \{d_1\}$. On the other hand, $\mathrm{WC}_{\Sigma,e_{|\{d_1\}}}(E) = \{d_1\}$. Observe that $\mathrm{WC}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$ is also a diagnosis, which corresponds to the solutions $H = \{d_1, d_2, \alpha_3\}$, and $H' = \{d_1, d_2, \alpha_1, \alpha_3\}$. ◊

In the following proposition, the correspondence between abductive diagnosis with weak-causality axioms and the notion of weak-causality diagnosis is established.

Proposition 4.4. *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem with weak-causality axioms $\mathcal{R}$, and let E^c be the set of finding literals according to the first version of the consistency condition. Let $\mathcal{P} = (\Sigma, E) = \tau'''(\mathcal{A})$ be the diagnostic problem corresponding to $\mathcal{A}$. A set of defects H is an abductive diagnosis for $\mathcal{A}$ iff $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$. Furthermore, the set of defects H is not an abductive diagnosis for $\mathcal{A}$ iff $\mathrm{WC}_{\Sigma,e_{|H}}(E) = u$.*

Proof. ($\Rightarrow$): Let $H' = \{d_1, \ldots, d_n, \alpha_1, \ldots, \alpha_m\}$ be a solution to the abductive diagnostic problem $\mathcal{A}$. From the fact that the covering and consistency conditions are satisfied, it follows that $e_{|H}(H) \supseteq E$, where H is the τ''' function value of the set of all defects in H'. Hence, by definition, it holds that $H = \mathrm{WC}_{\Sigma,e_{|H}}(E)$.

($\Leftarrow$): Let $H = \mathrm{WC}_{\Sigma,e_{|H}}(E)$, then from the mapping τ''' it follows that it is always possible to add sufficiently many assumption literals to H such that the covering and consistency conditions are satisfied. Hence, $H' = H \cup \{\alpha_1, \ldots, \alpha_m\}$ is a solution to $\mathcal{A}$ and H is an abductive diagnosis.

For undefined diagnoses, the proof is along similar lines. ◊

As may be expected, the notion of weak-causality diagnosis is Φ-complete. Again, a diagnosis accounts for all observed findings, although not every set of observed findings can be accounted for. The notion of weak-causality diagnosis is Δ-complete, because for each consistent set $D \subseteq \Delta$, it holds that $e(D) \supseteq E$, for some $E \subseteq \Phi$. It is Δ-monotonic, since if $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$ and $H' \supset H$ then $\mathrm{WC}_{\Sigma,e_{|H'}}(E) = H'$, because if $e_{|H}(H) \supseteq E$ then $e_{|H'}(H') \supseteq E$, due to the fact that the evidence function e is monotonically increasing.

Weak causality can be viewed as yet another way of interpreting an evidence function e. It is therefore possible to compare weak-causality and strong-causality diagnosis to each other. This is done in the following proposition.

Proposition 4.5. *Let* SCP *be the notion of strong-causality diagnosis with prediction and* WC *be the notion of weak-causality diagnosis with prediction, then*

$$\mathrm{SCP} \sqsubseteq \mathrm{WC}$$

Proof. Let $\mathcal{P} = (\Sigma, E)$ be any diagnostic problem with monotonically increasing evidence function e. If $\mathrm{SCP}_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) \supseteq E$, and for each $f, f' \in e_{|H}(H)$: if $f \in E$ and $s(f) = s(f')$ then $f' \in E$. However, the last condition is not required for weak-causality diagnosis, thus $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$ by definition. ◊

Summarized, the following now holds: SC $\sqsubseteq$ SCP $\sqsubseteq$ WC.

If the condition of the uniqueness of assumption literals is dropped, assumption literals must be represented in the framework in a way similar to defects. In essence, this means that the assumption literals introduce certain interactions between defects, which must be represented by means of an evidence function.

Example 4.6. Consider the abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$, where the set of abnormality axioms is equal to

$$\begin{aligned} d_1 \wedge \alpha_1 &\rightarrow f_1 \\ d_2 \wedge \alpha_1 &\rightarrow f_2 \\ d_2 \wedge \alpha_2 &\rightarrow f_3 \end{aligned}$$

and $E = \{f_1, f_3\}$, $E^c = \{\neg f_2\}$. The consistency condition fails to hold for this problem, because

$$\mathcal{R} \cup \{d_1, d_2, \alpha_1, \alpha_2\} \cup E^c \vDash \bot$$

Hence, there exists no solution to $\mathcal{A}$. However, if the abnormality axioms $\mathcal{R}$ are transformed by τ''', the following bottom-up partial specification of an evidence function e is obtained:

$$\tilde{e}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\} \\ \{f_2, f_3\} & \text{if } D = \{d_2\} \\ \varnothing & \text{if } D = \{\neg d_i\},\ i = 1, 2 \end{cases}$$

Note that now $\mathrm{WC}_{\Sigma, e_{|\{d_1,d_2\}}}(\{f_1, f_3\}) = \{d_1, d_2\}$, because

$$e_{|\{d_1,d_2\}}(\{d_1, d_2\}) = \{f_1, f_2, f_3\} \supseteq \{f_1, f_3\}$$

The reason for the difference in result is that the assumption literal α_1 introduces an interaction between the defects d_1 and d_2, which is neither expressed by means of evidence function e, nor by the notion of diagnosis WC. In order to express this interaction, it is necessary to add assumption literals as defects to the domain of the evidence function, as follows:

$$\tilde{e}'(D) = \begin{cases} \varnothing & \text{if } D = \{x_i\},\ D = \{\neg x_i\},\ x_i = \alpha_i \text{ or } x_i = d_i,\ i = 1, 2 \\ \{f_1\} & \text{if } D = \{d_1, \alpha_1\} \\ \{f_2\} & \text{if } D = \{d_2, \alpha_1\} \\ \{f_3\} & \text{if } D = \{d_2, \alpha_2\} \end{cases}$$

Observe the interaction introduced by the assumption literal α_1; for example, it holds that $e'(\{d_1, d_2\}) = \varnothing$, and $e'(\{\alpha_1\}) = \varnothing$, but $e(\{d_1, d_2, \alpha_1\}) = \{f_1, f_2\}$. Using strong-causality diagnosis SC with e' yields an undefined diagnosis. The result corresponds to that obtained by the theory of abductive diagnosis. $\Diamond$

It turns out that the notion of weak causality in its most general variant must be encoded by means of an evidence function e, instead of expressed at the level of a notion of diagnosis.

4.2.7 Combining weak and strong causality

Until now, the notions of weak and strong causality in abductive diagnosis have been studied separately. The abductive theory of diagnosis, however, allows for the combined application of these two forms of causality in the specification of the abnormality axioms

$\mathcal{R}$. In this section, the resulting combination will be studied. From now on, it is supposed that assumption literals are unique.

The notion of strong-causality diagnosis, as well as the notion of weak-causality diagnosis, were defined for the same evidence functions. However, if part of a causal theory is to be interpreted using a notion of strong-causality diagnosis, where another part of the causal theory is interpreted using a notion of weak-causality diagnosis, it is necessary to decompose e into two separate functions; one function for each notion of diagnosis. For this purpose, the transformation τ is slightly extended to a transformation τ^{iv}.

Firstly, the evidence function e in a diagnostic specification is split up into two evidence functions:

$$\nu : \wp(\Delta) \to \wp(\Phi) \cup \{\bot\}$$

called the *strong evidence function*, and

$$\alpha : \wp(\Delta) \to \wp(\Phi) \cup \{\bot\}$$

called the *weak evidence function*, and the functions ν and α are defined such that

$$e(D) = \begin{cases} \nu(D) \cup \alpha(D) & \text{if } \nu(D), \alpha(D) \neq \bot \\ \bot & \text{otherwise} \end{cases}$$

for each $D \subseteq \Delta$. Secondly, the translation scheme proposed in Section 4.2.2 is now modified by decomposing the set of abnormality axioms $\mathcal{R}$ into two disjoint sets $\mathcal{R}'$ and $\mathcal{R}''$, where the set $\mathcal{R}'$ contains only logical implications with every implication supplied with a unique assumption literal, and $\mathcal{R}'' = \mathcal{R} \backslash \mathcal{R}'$ is the set of all other logical implications without assumption literals.

To capture the combined result of strong and weak causality on diagnostic problem solving, the result of two separate diagnostic components must be combined. However, a diagnostic component that captures abductive diagnosis using strong causality and a diagnostic component capturing abductive diagnosis using weak causality operate each on part of a diagnostic specification. To describe a diagnostic specification as a collection of diagnostic specifications, the notion of modularization appears to be convenient.

Definition 4.5 (*modularization*). *A* modularization $\mathcal{M}_\Sigma$ *of a diagnostic specification* $\Sigma = (\Delta, \Phi, e)$ *is a finite set of diagnostic specifications* $\mathcal{M}_\Sigma = \{\Sigma_1, \ldots, \Sigma_n\}$, *where* $\Sigma_i = (\Delta, \Phi, e_i)$, $1 \leq i \leq n$, $n \geq 1$, *such that for each* $D \subseteq \Delta$*:*

$$e(D) = \begin{cases} \bigcup_{i=1}^{n} e_i(D) & \text{if } e_i(D) \neq \bot,\ 1 \leq i \leq n \\ \bot & \text{otherwise} \end{cases}$$

Modularization of a diagnostic specification is now employed to define the composition of two diagnostic components.

Definition 4.6 (*composition of diagnostic components*). *Let* P, Q *and* R *be three notions of diagnosis, and let* $\mathcal{M}_\Sigma = \{\Sigma', \Sigma''\}$ *be a modularization of the diagnostic specification* Σ. *Then, the diagnostic component* $P_{\Sigma, e_{|H}}$ *is called the* composition *of* $Q_{\Sigma', e'_{|H}}$ *and* $R_{\Sigma'', e''_{|H}}$, *denoted by*

$$P_{\Sigma, e_{|H}} = Q_{\Sigma', e'_{|H}} \| R_{\Sigma'', e''_{|H}}$$

if it holds that

$$P_{\Sigma,e_{|H}}(E) = Q_{\Sigma',e'_{|H}}(E') \cup R_{\Sigma'',e''_{|H}}(E'')$$

for each set of observed findings $E \subseteq \Phi$, and each decomposition $E = E' \cup E''$ for which $Q_{\Sigma',e'_{|H}}(E')$, $R_{\Sigma'',e''_{|H}}(E'') \neq u$; otherwise $P_{\Sigma,e_{|H}}(E) = u$.

Observe that the sets E' and E'' resulting from a decomposition of the set of observed findings E are neither necessarily disjoint nor unique. Note also that the hypothesis H is the same for all diagnostic components in a composition. This prerequisite ensures that possible dependencies among the respective evidence functions e' and e'' are dealt with adequately.

Using the translation scheme and the composition of diagnostic components, the following notion of diagnosis fully captures the abductive theory of diagnosis. The resulting notion of diagnosis is called weak-and-strong causality diagnosis, abbreviated to WSC.

Definition 4.7 (*weak-and-strong causality diagnosis*). *Let $\mathcal{M}_\Sigma = \{\Sigma', \Sigma''\}$ be a modularization of a diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Sigma' = (\Delta, \Phi, e')$ and $\Sigma'' = (\Delta, \Phi, e'')$. The notion of* weak-and-strong causality diagnosis*, denoted by* WSC*, is defined as follows:*

$$\mathrm{WSC}_{\Sigma,e_H} = \mathrm{SC}_{\Sigma',e'_{|H}} \| \mathrm{WC}_{\Sigma'',e''_{|H}}$$

where SC *is the notion of strong-causality diagnosis, and* WC *is the notion of weak-causality diagnosis.*

Note that another notion of weak-and-strong causality diagnosis is obtained if the second version of the consistency condition in the definition of abductive diagnosis is adopted. Then,

$$\mathrm{WSC}_{\Sigma,e_H} = \mathrm{SCP}_{\Sigma',e'_{|H}} \| \mathrm{WC}_{\Sigma'',e''_{|H}}$$

Since the notion of weak-and-strong-causality diagnosis is based on the notions of strong-causality diagnosis and weak-causality diagnosis, it is as rigorous as the notions of diagnosis on which it is based. The following lemma illustrates this point.

Lemma 4.1. *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem with modularization $\mathcal{M}_\Sigma = \{\Sigma', \Sigma''\}$, where $\Sigma' = (\Delta, \Phi, e')$ and $\Sigma'' = (\Delta, \Phi, e'')$. Then, if $\mathrm{WSC}_{\Sigma,e_{|H}}(E) = D$, $D \neq u$, then:*

(1) $\mathrm{SC}_{\Sigma,e'_{|H}}(E') = D$,

(2) $\mathrm{WC}_{\Sigma,e''_{|H}}(E'') = D$, *and*

(3) $D = H$,

for $E = E' \cup E''$.

Proof. From the definition of the composition of diagnostic components, and from the definitions of strong-causality and weak-causality diagnosis, the equalities (1) and (2)

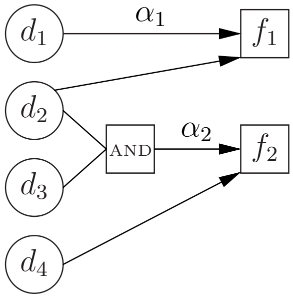


Figure 4.2: Causal net corresponding to $\mathcal{C}$.

follow with $E = E' \cup E''$. From these, it follows that $D = H$. ◊

In the following proposition, it is established that the notion of abductive diagnosis fully expresses the abductive, diagnostic theory by Console and Torasso.

Proposition 4.6. *Let $\mathcal{A} = (\mathcal{C}, E)$ be an abductive diagnostic problem. Let $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ be transformed to the modularization $\mathcal{M}_\Sigma = \{\Sigma', \Sigma''\}$ by the mapping τ^{iv}, such that:*

(1) $\Sigma' = (\Delta, \Phi, \nu)$, *with* $\nu = \tau^{\text{iv}}(\mathcal{R}_{strong})$

(2) $\Sigma'' = (\Delta, \Phi, \alpha)$, *with* $\alpha = \tau^{\text{iv}}(\mathcal{R}_{weak})$

where $\mathcal{R} = \mathcal{R}_{strong} \cup \mathcal{R}_{weak}$ is a partition of $\mathcal{R}$, with $\mathcal{R}_{strong}$ the set of abnormality axioms for strong causality, and $\mathcal{R}_{weak}$ the set of abnormality axioms for weak causality. Then, H is an abductive diagnosis of $\mathcal{A}$ iff $H = \text{WSC}_{\Sigma, e_{|H}}(E)$.

Proof. An abductive diagnosis is the result of the application of both sets of abnormality axioms $\mathcal{R}_{strong}$ and $\mathcal{R}_{weak}$. It has been proven in propositions 4.1 and 4.4 that the notions of strong-causality and weak-causality diagnosis capture abductive diagnosis using strong and weak causality, respectively. Combining these two notions of diagnosis yields, according to Lemma 4.1, a corresponding diagnosis. ◊

The following example illustrates the proposition above.

Example 4.7. Consider the following abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$, with causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, where $\mathcal{R}$ is equal to:

$$\begin{array}{l} d_1 \wedge \alpha_1 \rightarrow f_1 \\ d_2 \wedge d_3 \wedge \alpha_2 \rightarrow f_2 \\ d_2 \rightarrow f_1 \\ d_4 \rightarrow f_2 \end{array}$$

$\Delta_P = \{d_1, d_2, d_3, d_4\}$, and $\Phi_P = \{f_1, f_2\}$. The causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ is graphically depicted in Figure 4.2. The following modularization $\mathcal{M}_\Sigma = \{\Sigma', \Sigma''\}$ can be constructed: $\Sigma' = (\Delta, \Phi, \alpha)$, where the bottom-up partial specification $\tilde{\alpha}$ is defined as follows:

$$\tilde{\alpha}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2, d_3\} \\ \varnothing & \text{if } D = \{d_i\},\ i = 2, 3, 4, \text{ or } D = \{\neg d_i\},\ i = 1, \ldots, 4 \end{cases}$$

Furthermore, $\Sigma'' = (\Delta, \Phi, \nu)$, where the bottom-up partial specification of ν is defined as

$$\tilde{\nu}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_2\} \\ \{f_2\} & \text{if } D = \{d_4\} \\ \varnothing & \text{if } D = \{d_i\},\ i = 1, 3, \text{ or } D = \{\neg d_i\},\ i = 1, \ldots, 4 \end{cases}$$

Since every observable finding in $e(D)$ is positive, only positive findings will be dealt with.

An example of a diagnostic component for weak-causality diagnosis is:

$$\begin{aligned} \mathrm{WC}_{\Sigma',\alpha_{|\{d_1,d_2\}}} = \{ & (\varnothing, \{d_1, d_2\}), \\ & (\{f_1\}, \{d_1, d_2\}), \\ & (\{f_2\}, u), \\ & (\{f_1, f_2\}, u)\} \end{aligned}$$

This diagnostic component expresses the part of abductive diagnosis concerned with weak causality. An example of a diagnostic component for strong-causality diagnosis is:

$$\begin{aligned} \mathrm{SC}_{\Sigma'',\nu_{|\{d_1,d_2\}}} = \{ & (\varnothing, u), \\ & (\{f_1\}, \{d_1, d_2\}), \\ & (\{f_2\}, u), \\ & (\{f_1, f_2\}, u)\} \end{aligned}$$

which expresses abductive diagnosis by strong causality. An example of a diagnostic component from the notion of weak-and-strong causality diagnosis WSC is

$$\begin{aligned} \mathrm{WSC}_{\Sigma,e_{|\{d_1,d_2\}}} &= \mathrm{WC}_{\Sigma',\alpha_{|\{d_1,d_2\}}} \| \mathrm{SC}_{\Sigma'',\nu_{|\{d_1,d_2\}}} \\ &= \{(\varnothing, u\}), \\ &\quad\ (\{f_1\}, \{d_1, d_2\}), \\ &\quad\ (\{f_2\}, u), \\ &\quad\ (\{f_1, f_2\}, u)\} \end{aligned}$$

Note that, for example, $\mathrm{WSC}_{\Sigma,e_{|\{d_1,d_2\}}}(\varnothing) = u$, because $\mathrm{SC}_{\Sigma'',\nu_{|\{d_1,d_2\}}}(\varnothing) = u$, although $\mathrm{WC}_{\Sigma',\alpha_{|\{d_1,d_2\}}}(\varnothing) = \{d_1, d_2\}$. Observe also that the set of observed findings E may be decomposed among diagnostic components WC and SC in several ways. For example,

$$\begin{aligned} \mathrm{WSC}_{\Sigma,e_{|\{d_2,d_3,d_4\}}}(\{f_1, f_2\}) &= \mathrm{WC}_{\Sigma',\alpha_{|\{d_2,d_3,d_4\}}}(\varnothing) \cup \mathrm{SC}_{\Sigma'',\nu_{|\{d_2,d_3,d_4\}}}(\{f_1, f_2\}) \\ &= \mathrm{WC}_{\Sigma',\alpha_{|\{d_2,d_3,d_4\}}}(\{f_2\}) \cup \mathrm{SC}_{\Sigma'',\nu_{|\{d_2,d_3,d_4\}}}(\{f_1, f_2\}) \\ &= \mathrm{WC}_{\Sigma',\alpha_{|\{d_2,d_3,d_4\}}}(\{f_1\}) \cup \mathrm{SC}_{\Sigma'',\nu_{|\{d_2,d_3,d_4\}}}(\{f_1, f_2\}) \\ &= \mathrm{WC}_{\Sigma',\alpha_{|\{d_2,d_3,d_4\}}}(\{f_1, f_2\}) \cup \\ &\quad\ \mathrm{SC}_{\Sigma'',\nu_{|\{d_2,d_3,d_4\}}}(\{f_1, f_2\}) \end{aligned}$$

$\Diamond$

The following proposition brings weak, strong and weak-and-strong causality diagnosis in relation to each other.

Proposition 4.7. *Let* WC, SC *and* WSC *be the notions of weak, strong and weak-and-strong causality diagnosis, respectively. Then,*

$$\mathrm{SC} \sqsubseteq \mathrm{WSC} \sqsubseteq \mathrm{WC}$$

Proof. Let Σ be a diagnostic specification. If $\mathrm{SC}_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) = E$. Since, $e_{|H}(H) = \nu_{|H}(H) \cup \alpha_{|H}(H) = E$, by definition, for any choice of the evidence functions ν and α for the modularization $\mathcal{M}_\Sigma$, it holds that $\nu_{|H}(H) = E' \subseteq E$ and $E'' \subseteq \alpha_{|H}(H) \subseteq E$, $E'' = E \backslash E'$. Hence, $\mathrm{SC}_{\Sigma,\nu_{|H}}(E') = H$ and $\mathrm{WC}_{\Sigma,\alpha_{|H}}(E'') = H$. Furthermore, if $\mathrm{WSC}_{\Sigma,e_{|H}}(E) = H$, then, by Lemma 4.1, it follows that $\nu_{|W}(H) \cup \alpha_{|H}(H) = e_{|H}(H) \supseteq E$. Therefore, $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$. ◊

4.2.8 Remaining issues

There remain some issues that have not been dealt with above, although they are part of the abductive theory of diagnosis. In particular, the theory includes criteria for selecting most plausible diagnoses from the set of all diagnoses for a set of observable findings. In the literature, plausible diagnoses are usually selected from the equivalence class $\mathcal{D}_{\mathrm{WSC}}(E, E)$, if it exists.

The notion of minimal multiple diagnosis $\mathcal{D}_R^{\subseteq}(E, E)$ can immediately be applied by taking for R the notion of weak-and-strong causality diagnosis WSC (or WC, SC). This completes our analysis of the theory of abductive diagnosis.

4.3 Analysis of the set-covering theory of diagnosis

In Section 2.2.3, the basic principles underlying the set-covering theory of diagnosis have been discussed in detail. As with the abductive theory of diagnosis, fixed causal associations between defects and observable findings are the central elements of this diagnostic theory [Peng & Reggia, 1990; Reggia et al., 1983; Tuhrim et al., 1991]. However, the analysis of the set-covering theory of diagnosis in terms of our framework of diagnosis is much easier, because both theories have a foundation in set theory.

4.3.1 Representation and interpretation

Recall from Section 2.2.3 that a diagnostic problem in the set-covering theory of diagnosis is a pair $\mathcal{D} = (\mathcal{N}, E)$, where $\mathcal{N} = (\Delta, \Phi, C)$ is a causal net. For reasons of convenience, in Section 2.2.3, the causation relation $C \subseteq \Delta \times \Phi$ was represented by means of an 'effects function' e, which was defined in terms of pairs (d, f) appearing in the relation C.

In defining a transformation τ^{v} to map a diagnostic problem $\mathcal{D}$ in set-covering theory to a diagnostic problem $\mathcal{P}$ in our framework, i.e. $\tau^{\mathrm{v}}(\mathcal{D}) = \mathcal{P}$, its effects function e can immediately be taken as the evidence function e' of a diagnostic specification $\Sigma = (\Delta', \Phi', e')$. Then, Δ' and Φ' correspond to Δ and Φ from $\mathcal{D}$, respectively. The sets Φ and Δ will consist of positive defects and findings only, when only knowledge concerning the presence of defects is represented. In this case, the set Δ'_P is defined as $\Delta'_P = \tau^{\mathrm{v}}(\Delta)$ and $\Delta'_N = \{\neg\tau^{\mathrm{v}}(d) \mid d \in \Delta\}$; similar definitions apply to the set of findings Φ. The set of observed findings E' in a diagnostic problem $\mathcal{P} = (\Sigma, E')$ is just the set of observed findings E in $\mathcal{D}$. The resulting evidence function e' can be partially specified by means of a bottom-up partial specification, because it is interaction free. (Actually, this was already done in Section 2.2.3.)

The intended meaning of an effects function e in the set-covering theory of diagnosis is that each function value $e(d)$, $d \in \Delta$, provides a description of the observable findings for the defect d; hence, an additional property of the corresponding evidence function e' is that for each $d \in \Delta'$: $e(d) \neq \varnothing$.

4.3.2 Notion of diagnosis

In Definition 2.6, a set-covering diagnosis for a diagnostic problem $\mathcal{D} = (\mathcal{N}, E)$ was defined as a set of defects $D \subseteq \Delta$ for which

$$e(D) \supseteq E$$

The same condition must be incorporated in a notion of diagnosis R in our framework. In fact, the corresponding notion of diagnosis has already been introduced; it is the notion of weak-causality diagnosis WC (cf. Definition 4.4). The notion of weak-causality diagnosis requires that the evidence function e is monotonically increasing. This condition is fulfilled, because any function that is interaction free is also monotonically increasing (cf. Proposition 3.2). Hence, the notion of weak-causality diagnosis for interaction-free evidence functions corresponds to the notion of set-covering diagnosis. For ease of exposition, most of the elements of $\mathcal{P}$ will be identified with the corresponding elements of $\mathcal{D}$.

Proposition 4.8. *Let $\mathcal{D} = (\mathcal{N}, E)$ be a diagnostic problem in the set-covering theory of diagnosis. Let $\mathcal{P} = \tau^{\mathrm{v}}(\mathcal{D})$ be the diagnostic problem obtained by the transformation τ^{v}. Then, $H \subseteq \Delta$ is a set-covering diagnosis of $\mathcal{D}$ iff $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$.*

Proof. ($\Rightarrow$): If H is a set-covering diagnosis, then $e(H) \supseteq E$. According to the transformation τ^{v} it holds that $e'_{|H}(H) \supseteq E$, where $e' = \tau^{\mathrm{v}}(e)$. Therefore, $\mathrm{WC}_{\Sigma,e'_{|H}}(E) = H$.

($\Leftarrow$): Simply note that if $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) \supseteq E$. The correspondence of $\mathcal{D}$ and $\mathcal{P}$ ensures that H is a set-covering diagnosis of $\mathcal{D}$. ◊

Of particular importance in the set-covering theory of diagnosis are ‘criteria of parsimony’, which explains the alternative name of the theory: ‘parsimonious covering theory’. These criteria are used to select the most plausible diagnoses from the entire set of diagnoses. The criterion that is usually taken as the most useful domain-independent measure of plausibility is that of irredundant diagnosis. In our framework the notion of minimal multiple diagnosis with respect to weak-causality diagnosis, i.e. $\mathcal{D}^{\subseteq}_{\mathrm{WC}}(E, E)$, corresponds to the set of irredundant diagnoses of a diagnostic problem $\mathcal{D}$.

Proposition 4.9. *Let $\mathcal{P} = (\Sigma, E) = \tau^{\mathrm{v}}(\mathcal{D})$ be a diagnostic problem obtained from the diagnostic problem $\mathcal{D}$ in the set-covering theory of diagnosis. Then, the set of irredundant diagnoses of $\mathcal{D}$ corresponds to $\mathcal{D}^{\subseteq}_{\mathrm{WC}}(E, E)$.*

Proof. The multiple diagnosis $\mathcal{D}_{\mathrm{WC}}(E, E)$ is according to Proposition 4.8 equal to the set of all set-covering diagnosis. Since, a minimal multiple diagnosis is just the set of all minimal sets with respect to set inclusion in $\mathcal{D}_{\mathrm{WC}}(E, E)$, and all irredundant diagnoses are also minimal with respect set inclusion, the two sets must correspond to each other. ◊

4.4 Analysis of consistency-based diagnosis

The theory of consistency-based diagnosis, as originally developed by R. Reiter, [Reiter, 1987], and later extended by de Kleer et al., [De Kleer et al., 1992], differs from the formal theories discussed in the previous sections by having the notion of (un)satisfiability as its core concept. This very general logical notion does not have an immediate relationship with the interpretation of observed findings in the process of diagnosis. Therefore, in order to translate consistency-based diagnosis in terms of our framework, it is necessary to be more precise about the intended meaning of the framework of consistency-based diagnosis than in Section 2.2.1. First, the logical representation used in consistency-based diagnosis will be translated into the terminology of our framework. Next, the set-theoretical analogue of consistency-based diagnosis will be developed and explored.

4.4.1 Representation and interpretation

Recall that the starting point for consistency-based diagnosis is the notion of a system $\mathcal{S}$. Following [De Kleer et al., 1992], a system $\mathcal{S}$ was defined in Section 2.2.1 as a triple $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS}, \mathrm{OBS})$, where

- SD is a *system description*, represented by means of formulae in first-order logic;
- COMPS is a finite set of *components*, represented by means of constants in first-order logic;
- OBS is a finite set of *observations*, represented by means of formulae in first-order logic.

In order to create a sensible translation of a system $\mathcal{S}$ to a diagnostic problem, it is desirable to make a distinction between formulae in the set of observations OBS representing the inputs I to the system, and the formulae in OBS representing the outputs O. Furthermore, for a system description SD to correspond to our notion of diagnostic specification Σ, it is necessary to know which sets constitute acceptable sets of observations OBS. This information is actually hidden in a system description SD. Henceforth, it is assumed that it is possible to derive from a system description SD, information concerning the set of legal inputs, denoted by IN, and the set of legal outputs, denoted by OUT. The sets IN and OUT are not necessarily disjoint. Accordingly, the set of observations OBS is defined as

$$\mathrm{OBS} = I \cup O$$

where $I \subseteq \mathrm{IN}$ and $O \subseteq \mathrm{OUT}$. The bijective transformation τ^{vi} that maps a system $\mathcal{S}$ to a diagnostic problem $\mathcal{P} = (\Sigma, E)$, with $\Sigma = (\Delta, \Phi, e)$, and $\tau^{\mathrm{vi}}(\mathcal{S}) = \mathcal{P}$, can now be defined as follows. Let

$$C_P = \{\mathrm{Abnormal}(c) \mid c \in \mathrm{COMPS}\}$$

and

$$C_N = \{\neg\mathrm{Abnormal}(c) \mid c \in \mathrm{COMPS}\}$$

then, a diagnostic problem $\mathcal{P} = (\Sigma, E)$, with $\Sigma = (\Delta, \Phi, e)$, can be chosen, such that:

- $\tau^{\mathrm{vi}}(C_P) = \Delta'_P$, $\tau^{\mathrm{vi}}(C_N) = \Delta'_N$, and $\tau^{\mathrm{vi}}(c) = \neg\tau^{\mathrm{vi}}(\neg c)$, for each $c \in C_P \cup C_N$ ($\neg c$ denotes a formula in first-order logic, where $\neg$ stands for logical negation)
- $\tau^{\mathrm{vi}}(\mathrm{IN} \cup \mathrm{OUT}) = \Phi$, and $\tau^{\mathrm{vi}}(o) = \neg\tau^{\mathrm{vi}}(\neg o)$, for each $o \in \mathrm{IN} \cup \mathrm{OUT}$
- $\tau^{\mathrm{vi}}(\mathrm{OBS}) = E$
- for each $C \subseteq C_N \cup C_P$, and for each $I \subseteq \mathrm{IN}$:

  (1) if $\mathrm{SD} \cup I \cup C$ is satisfiable, then

  $$\begin{aligned} Q = \{\tau^{\mathrm{vi}}(\{l_1, \ldots, l_n\}) \mid\ & \mathrm{SD} \cup I \cup C \vDash O, O = l_1 \vee \cdots \vee l_n, \\ & l_i \in \mathrm{OUT}, \text{for } i = 1, \ldots, n, n \geq 1, \\ & \mathrm{SD} \cup I \cup C \nvDash O', O' \vDash O, O \nvDash O'\} \end{aligned}$$

  $e(D) = \bigcup_{S \in Q} S$

  (2) otherwise ($\mathrm{SD} \cup I \cup C$ is unsatisfiable), $e(D) = \bot$

  where $D = \tau^{\mathrm{vi}}(I \cup C), D \subseteq \Delta$.

The resulting evidence function e is a function

$$e : \wp(\Phi' \cup \Delta') \to \wp(\Phi'') \cup \{\bot\}$$

where $\Delta = \Phi' \cup \Delta'$, $\Delta' = \tau^{\mathrm{vi}}(C)$, $\Phi' = \tau^{\mathrm{vi}}(\mathrm{IN})$, $\Phi'' = \tau^{\mathrm{vi}}(\mathrm{OUT})$, and $\Phi = \Phi' \cup \Phi''$. Hence, the set of observable findings Φ and the set of defects Δ need not be disjoint. If a system with a fixed set of observations OBS, as in the definition of $\mathcal{S}$, is mapped to a diagnostic problem $\mathcal{P}$, the sets Φ and Δ are again disjoint.

Given a function value $e(D) = F$, the set D consists of defects and observable findings. Now, assume that functions φ and δ are defined, such that findings and defects are extracted from each set $D \subseteq \Delta$, as follows:

- $\varphi(D) = D \backslash \Delta'$
- $\delta(D) = D \backslash \Phi'$

Note that $D = \varphi(D) \cup \delta(D)$. Outside the traditional domain of electronic circuits, usually $\delta(D) = D$ and $\varphi(D) = \varnothing$. These two functions will be required in the next section.

The construction of a diagnostic problem $\mathcal{P}$ from a system $\mathcal{S}$ is illustrated by means of the following example.

Example 4.8. Reconsider the full-adder system described in Section 2.2.1. Part of the evidence function e resulting from the mapping $\mathcal{P} = \tau^{\mathrm{vi}}(\mathcal{S})$ is:

$$\begin{aligned} e(\{i_1, \neg i_2, i_3, \neg x_1, \neg x_2, \neg a_1, \neg a_2, \neg r_1\}) &= \{\neg o_1, o_2\} \\ e(\{i_1, \neg i_2, i_3, x_1, \neg x_2, \neg a_1, \neg a_2, \neg r_1\}) &= \varnothing \\ e(\{i_1, \neg i_2, i_3, \neg x_1, x_2, \neg a_1, a_2, \neg r_1\}) &= \varnothing \\ e(\{i_1, \neg i_2, i_3, \neg x_1, x_2, \neg a_1, \neg a_2, r_1\}) &= \varnothing \\ e(\{i_1, \neg i_2, i_3, \neg x_1, x_2, \neg a_1, \neg a_2, \neg r_1\}) &= \{o_2\} \end{aligned}$$

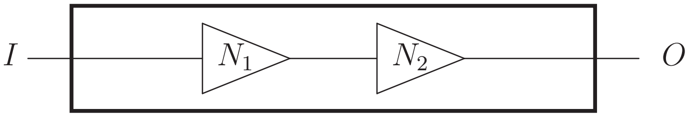


Figure 4.3: Two NOT gates in series (repeated).

where $i_1, \neg i_2, i_3$ represent the input to the system (being $1, 0, 1$), x_1 and x_2 express that the exclusive OR gates X_1 and X_2 are defective, etc; r_1 denotes that the OR gate R_1 is defective. The function value $e(D)$ for $\{i_1, \neg i_2, i_3, x_1, \neg x_2, \neg a_1, \neg a_2, \neg r_1\}$ states that when only X_1 is defective, none of the findings is observable. This is due to the fact that the presence of a defect is simulated by removing the corresponding component from the system, rendering part of the simulated system non-functional. Note that if X_2 is assumed to be defective, part of the system is still functioning because the output $e(D) = \{o_2\}$ is produced for $D = \{i_1, \neg i_2, i_3, \neg x_1, x_2, \neg a_1, \neg a_2, \neg r_1\}$.

If the evidence function above is assumed to stand for a top-down partial specification, it holds, for example, that

$$e(\{i_1, \neg i_2, i_3, \neg x_2, \neg a_1, \neg a_2, \neg r_1\}) = \{\neg o_1, o_2\}$$

Similar function values can be obtained for the other sets of defects D. ◊

If $e(D) = \varnothing$ this may indicate that the specification of the system $\mathcal{S}$ on which the evidence function e is based, is incomplete. No complete information about the output of the system for certain defects and inputs is available. This will happen when the system description SD only comprises a specification of the expected normal behaviour and not of abnormal behaviour as well. Such abnormal behaviour can be incorporated in our framework easily, by providing function values for $e(D) = \varnothing$, when $D \subseteq \Delta_P$. The resulting evidence function resembles the evidence functions in the abductive theory of diagnosis with respect to sets of positive defects. The process is illustrated by the evidence function e for the logical circuit consisting of two NOT gates in series (cf. Figure 4.3), introduced in Example 3.7.

Example 4.9. We freely follow [De Kleer et al., 1992] in the description of the normal behaviour of a logical circuit consisting of two NOT gates (inverters), denoted by N_1 and N_2, which is represented in logic as follows:

$$\begin{array}{l} \forall x(\neg \text{Abnormal}(x) \to (in(x) = 0 \leftrightarrow out(x) = 1)) \\ out(N_1) = in(N_2) \end{array}$$

where the equality axioms are supposed to be available, and $0 \neq 1$; this formula constitutes the system description SD of a system $\mathcal{S} = (\text{SD}, \text{COMPS}, \text{OBS})$, where furthermore

- the set of components COMPS is equal to COMPS $= \{N_1, N_2\}$;
- the given set of observations is equal to

  $$\text{OBS} = \{in(N_1) = 0, out(N_2) = 1\}$$

  which indicates that something is wrong with the circuit (because $out(N_2) = 0$ would be expected for a correctly functioning circuit).

The corresponding diagnostic problem $\mathcal{P} = (\Sigma, E)$, with $\Sigma = (\Delta, \Phi, e)$, resulting from $\tau^{\rm vi}(\mathcal{S})$, is defined as follows:

- $\Delta' = \{n_1, n_2, \neg n_1, \neg n_2\} = \tau^{\rm vi}(C)$, where

  $C = \{\text{Abnormal}(N_1), \text{Abnormal}(N_2), \neg\text{Abnormal}(N_1), \neg\text{Abnormal}(N_2)\}$

  $n_1 = \tau^{\rm vi}(\text{Abnormal}(N_1))$, and $n_2 = \tau^{\rm vi}(\text{Abnormal}(N_2))$
- $\Phi = \{i, \neg i, o, \neg o\} = \tau^{\rm vi}(\{in(N_1) = 1, in(N_1) = 0, out(N_2) = 1, out(N_2) = 0\})$, where $i = \tau^{\rm vi}(in(N_1) = 1)$, and $o = \tau^{\rm vi}(out(N_2) = 1)$
- $\Phi' = \{i, \neg i\} = \tau^{\rm vi}(\{in(N_1) = 1, in(N_1) = 0\})$, $\Delta = \Phi' \cup \Delta'$
- $E = \{\neg i, o\} = \tau^{\rm vi}(\text{OBS})$
- the corresponding evidence function e is defined as follows:

$$e(D) = \begin{cases} \{o\} & \text{if } D = \{i, \neg n_1, \neg n_2\} \\ \{\neg o\} & \text{if } D = \{\neg i, \neg n_1, \neg n_2\} \\ \bot & \text{if } D \text{ is syntactically inconsistent} \\ \varnothing & \text{otherwise} \end{cases}$$

Observe that only information of normal behaviour has been represented by means of the evidence function e. In [De Kleer et al., 1992], the system description SD is next extended by including information about abnormal behaviour, yielding SD$'$, by means of the following additional logical formulae:

$$\begin{aligned} \forall x(\text{Abnormal}(x) &\rightarrow (\text{SA0}(x) \vee \text{Short}(x))) && (4.1) \\ \forall x(\text{SA0}(x) &\rightarrow out(x) = 0) && (4.2) \\ \forall x(\text{Short}(x) &\rightarrow out(x) = in(x)) && (4.3) \end{aligned}$$

They mean that the output of an abnormal NOT gate is either stuck at 0 or shorted unmodified to its input. Among others, these three formulae produce the following changes to function values of the evidence function e, yielding e':

$$\begin{aligned} e'(\{i, n_1, n_2\}) &= \{o, \neg o\} \\ e'(\{\neg i, n_1, n_2\}) &= \{\neg o\} \end{aligned}$$

This reveals that the logical circuit exhibits nondeterministic abnormal behaviour if $D = \{i, n_1, n_2\}$, because both o and $\neg o$ are possible outputs. ◊

Nondeterministic behaviour in model-based diagnosis has received little attention in the literature on diagnostic problem solving.

4.4.2 The notion of consistency-based diagnosis

Recall that a consistency-based diagnosis C of a system $\mathcal{S}$ is a certain assignment of either a positive literal Abnormal(c) or a negative literal ¬Abnormal(c) to each $c \in \text{COMPS}$, such that

$$\text{SD} \cup \text{OBS} \cup C$$

is satisfiable (the *consistency condition*), where

$$C = \{\text{Abnormal}(c) \mid c \in D\} \cup \{\neg\text{Abnormal}(c) \mid c \in \text{COMPS}\backslash D\}$$

with $D \subseteq \text{COMPS}$ (cf. Definition 2.1).

Establishing a diagnosis in consistency-based diagnosis comprises two steps. Firstly, it is observed that the assumption that all components are functioning correctly contradicts with the observations made. This means that something must be wrong with the system. Secondly, having observed this discrepancy, it is attempted to discover which components are responsible for the malfunctioning. Only the second aspect of diagnosis is expressed in the following definition.

Definition 4.8 (*consistency-based diagnosis*). *The notion of* consistency-based diagnosis*, denoted by* CB*, is defined as follows:*

$$\text{CB}_{\Sigma,e_{|H}}(E) = \begin{cases} \delta(H) & \text{if } \forall f \in E : f \in e_{|H}(H) \vee \neg f \notin e_{|H}(H), \varphi(H) \subseteq E \\ u & \text{otherwise} \end{cases}$$

for each diagnostic specification $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$.

This notion of diagnosis is appropriate for dealing with nondeterministic output of the form $\{o, \neg o\}$; however, if such outputs need not be described, the condition simplifies to

$$e_{|H}(H) \subseteq E$$

because, if this condition holds, then for each $f \in E$, $f \notin e_{|H}(H)$, it holds that $\neg f \notin e_{|H}(H)$, otherwise (assuming $\neg f \in e_{|H}(H)$ holds) $e_{|H}(H) \nsubseteq E$ would hold, contradicting the assumption.

It is interesting to note that the first aspect of the description of consistency-based diagnosis, viz. that a discrepancy exists between the expected, observable, and observed findings, which is expressed by unsatisfiability, is not reflected in the definition of the notion of diagnosis CB. Although this condition could be expressed by adding the expression

$$\exists f \in e_{|F \cup \Delta'_N}(F \cup \Delta'_N) : \neg f \in E, \neg f \notin e_{|F \cup \Delta'_N}(F \cup \Delta'_N)$$

$F \subseteq \tau^{\text{vi}}(\text{IN})$, to the if-part of the definition above, this requirement is lacking from the original definitions proposed by Reiter, [Reiter, 1987], and de Kleer et al., [De Kleer et al., 1992].

The use of the notion of diagnosis CB is illustrated by means of an example.

Example 4.10. Consider again Example 4.8. For

$$E = \{i_1, \neg i_2, i_3, o_1, \neg o_2\}$$

and

$$H = \{i_1, \neg i_2, i_3, x_1, \neg x_2, \neg a_1, \neg a_2, \neg r_1\}$$

a possible diagnosis, using the notion of consistency-based diagnosis CB, is:

$$\mathrm{CB}_{\Sigma, e_{|H}}(E) = \{x_1, \neg x_2, \neg a_1, \neg a_2, \neg r_1\}$$

Following [Reiter, 1987], this diagnosis would be denoted as the set $\{x_1\}$, containing only the positive elements from $\mathrm{CB}_{\Sigma, e_{|H}}$. The set of findings accounted for by this diagnosis is equal to

$$A(\mathrm{CB}_{\Sigma, e_{|H}}(E) \cup \varphi(H), E) = \varnothing$$

This is, of course, a consequence of the lack of information concerning the abnormal behaviour of the circuit. Now, suppose that

$$H' = \{i_1, \neg i_2, i_3, \neg x_1, x_2, \neg a_1, \neg a_2, \neg r_1\}$$

then

$$\mathrm{CB}_{\Sigma, e_{|H'}}(E) = u$$

because $\neg o_2 \in E$ but $\neg o_2 \notin e_{|H'}(H')$ and $o_2 \in e_{|H'}(H')$. $\Diamond$

As discussed above, models of abnormal behaviour can be incorporated easily in our diagnostic framework. In fact, the notion of diagnosis CB need not be adapted to deal with models of abnormal behaviour. It depends, of course, on the nature of the evidence function e whether or not the notion of consistency-based diagnosis handles knowledge of abnormal behaviour adequately. For example, if the knowledge of abnormal behaviour is causal in nature, for instance similar to the kind of causal knowledge represented in the abductive theory of diagnosis, CB may produce diagnoses that appear intuitively incorrect. For example, consistency-based diagnosis would accept as a diagnosis a hypotheses H for which $e_{|H}(H) \subset E$, which would not be a valid diagnosis in the abductive theory of diagnosis.

In [De Kleer et al., 1992] the notion of *partial diagnosis* is introduced, which is a satisfiable subset C of $C_P \cup C_N$, such that the consistency condition is fulfilled for every satisfiable $C' \supseteq C$, $C' \subseteq C_P \cup C_N$. A *kernel diagnosis* is a partial diagnosis that is minimal with respect to set inclusion. To prepare for dealing with partial diagnosis, partial information concerning defects has already been taken into account in defining the evidence function e. If the set of defects is a not a maximally syntactically consistent set of defects, the function value $e(F \cup D)$, where F represents the set of input findings and D defective components, can be handled by the notion of diagnosis CB. We give an example.

Example 4.11. Reconsider Example 4.9. The diagnostic problem $\mathcal{P} = (\Sigma, E)$, with set of observed findings $E = \{\neg i, o\}$, corresponds to the system $\mathcal{S}$, without the specification of abnormal behaviour (formulae (4.1) – (4.3)). Then, for $H = \{\neg i, n_1, n_2\}$ the CB diagnosis is equal to

$$\mathrm{CB}_{\Sigma, e_{|H}}(E) = \{n_1, n_2\}$$

because $e(\{\neg i, n_1, n_2\}) = \varnothing$, and since

$$\mathrm{SD} \cup \{\mathrm{Abnormal}(N_1), \mathrm{Abnormal}(N_2)\} \cup \{\mathit{in}(N_1) = 0, \mathit{out}(N_2) = 1\}$$

is satisfiable, the diagnoses correspond to each other. However, if the logical formulae (4.1) – (4.3) are added to SD, yielding the system description SD′, then

$$\mathrm{SD}' \cup \{\mathrm{Abnormal}(N_1), \mathrm{Abnormal}(N_2)\} \cup \{\mathit{in}(N_1) = 0, \mathit{out}(N_2) = 1\}$$

is unsatisfiable. Similarly, $\mathrm{CB}_{\Sigma', e'_{|H}}(E) = u$, because $o \in E$, but $o \notin e'_{|H}(H)$ and $\neg o \in e'_{|H}(H)$, where e' represents the evidence function e after incorporation of the information on abnormal behaviour.

A partial diagnosis for the hypothesis $H' = \{\neg i, n_1\}$ is

$$\mathrm{CB}_{\Sigma, e_{|H'}}(E) = \{n_1\}$$

(again precluding the knowledge concerning abnormal behaviour), because $e_{|\{\neg i, n_1\}}(\{\neg i, n_1\}) = \varnothing \subseteq E$. ◊

In the following proposition, the equivalence of the notion of consistency-based diagnosis introduced in Definition 2.1, [De Kleer et al., 1992], and our notion of diagnosis CB is proven, for system descriptions SD with nondisjunctive (nondeterministic) outputs.

Proposition 4.10. *Let $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS}, \mathrm{OBS})$ be a system and let $\mathcal{P} = (\Sigma, E)$ be the corresponding diagnostic problem with $\mathcal{P} = \tau^{\mathrm{vi}}(\mathcal{S})$. Then, C is a consistency-based diagnosis iff $\mathrm{CB}_{\Sigma, e_{|H}}(E) = H'$, for some $H \subseteq \Delta$ and $H' = \tau^{\mathrm{vi}}(C)$.*

Proof. (⇒): If C is a (possibly partial) diagnosis, then the consistency condition must be satisfied. Let

$$O = \{o \in \mathrm{OUT} \mid \mathrm{SD} \cup I \cup C \vDash o\}$$

where $C \subseteq C_P \cup C_N$, $I \subseteq \mathrm{IN}$, and $I \subseteq \mathrm{OBS}$. For each output $o \in \mathrm{OBS}$: $o \in O$ or $\neg o \notin O$ (consistency condition). If $\tau^{\mathrm{vi}}(I \cup C) = H$, then by definition $e(H) = \tau^{\mathrm{vi}}(O) = F$. Then, for each output $o \in \mathrm{OBS}$: $\tau^{\mathrm{vi}}(o) = f \in F$ or $\tau^{\mathrm{vi}}(\neg o) = \neg f \notin F$, and $\varphi(H) \subseteq \tau^{\mathrm{vi}}(\mathrm{OBS}) = E$. Hence, $\delta(H) = H'$ is a diagnosis.

(⇐): If $\mathrm{CB}_{\Sigma, e_{|H}}(E) = H'$, with $H' \neq u$, and $e(H) = F$, then for each $f \in E$: $f \in F$ or $\neg f \notin F$, and $\varphi(H) \subseteq E$. Let $\tau^{\mathrm{vi}}(O) = F$, and $\tau^{\mathrm{vi}}(I \cup C) = H$, $\tau^{\mathrm{vi}}(\mathrm{OBS}) = E$. Then, it holds that

$$\mathrm{SD} \cup I \cup C \vDash O$$

and for each $o \in \mathrm{OBS}$: $o \in O$ or $\neg o \notin O$. Therefore, $\mathrm{SD} \cup \mathrm{OBS} \cup C \nvDash \perp$, i.e. the consistency condition is satisfied. ◊

Clearly, the notion of diagnosis CB is Φ-incomplete, because it is not necessary that every observed finding is accounted for, as is shown by choosing an evidence function e with $e(D) = \varnothing$ for each syntactically consistent set $D \subseteq \Delta$. Furthermore, CB diagnosis is Δ-complete, because a diagnosis exists for any consistent set of defects with the empty set of observed findings (or possibly with a set of observed findings, containing only inputs). Without further restrictions with regard to the evidence functions e, the notion of CB diagnosis is neither Δ-monotonic nor is the independence assumption satisfied. However, the independence assumption is satisfied if CB diagnosis is restricted to diagnostic specifications that are monotonically increasing.

Proposition 4.11. *If the notion of consistency-based diagnosis* CB *is restricted to diagnostic specifications* $\Sigma = (\Delta, \Phi, e)$ *with monotonically increasing evidence function* e*, then the independence assumption is satisfied for* CB*.*

Proof. If $\mathrm{CB}_{\Sigma, e_{|H}}(E) = H'$, then for each $f \in E$: $f \in e_{|H}(H)$ or $\neg f \notin e_H(H)$ and $\varphi(E) = \varphi(H)$. Because e is monotonically increasing, it holds that for each $f \in E$: $f \in e_{|\{d\}}(\{d\})$ or $\neg f \notin e_{|\{d\}}(\{d\})$, hence $\mathrm{CB}_{\Sigma, e_{|H}}(E) = \bigcup_{d \in H} \mathrm{CB}_{\Sigma, e_{|\{d\}}}(E)$. $\Diamond$

However, as discussed in Chapter 3, evidence functions representing system descriptions are typically monotonically decreasing. If the notion of diagnosis CB is defined for such functions, the independence assumption fails to hold, as can be shown by a simple counter-example. However, the following useful proposition holds in case the evidence function is monotonically decreasing.

Proposition 4.12. *Let* $\mathcal{P} = (\Sigma, E)$*,* $\Sigma = (\Delta, \Phi, e)$*, be a diagnostic problem with monotonically decreasing evidence function* e*, and let* $H \supseteq H'$*, with* $H, H' \subseteq \Delta$*, then if* $\mathrm{CB}_{\Sigma, e_{|H}}(E) = D$*, then* $\mathrm{CB}_{\Sigma, e_{|H'}}(E) = D'$ *with* $D' \subseteq D$*.*

Proof. If $\mathrm{CB}_{\Sigma, e_{|H}}(E) = D$ then for each $f \in E$: (1) $f \in e_{|H}(H)$ or (2) $\neg f \notin e_{|H}(H)$. If condition (1) holds then $f \in e_{|H'}(H')$, because $e_{|H}(H) \subseteq e_{|H'}(H')$; for the same reason from condition (2) it follows that $\neg f \notin e_{|H'}(H')$. $\Diamond$

In terms of the approach by de Kleer et al., [De Kleer et al., 1992], from this proposition the existence of a partial diagnosis can be derived.

Corollary. *Let* $\mathcal{P} = (\Sigma, E)$ *be a diagnostic problem, with monotonically decreasing evidence function* e*, then if* $\mathrm{CB}_{\Sigma, e_{|H \cup \{d\}}}(E) = H \cup \{d\}$ *and* $\mathrm{CB}_{\Sigma, e_{|H \cup \{\neg d\}}}(E) = H \cup \{\neg d\}$*, then also* $\mathrm{CB}_{\Sigma, e_{|H}}(E) = H$*.*

In [De Kleer et al., 1992], the notion of partial diagnosis is provided as a basic definition; it is not derived from the notion of diagnosis for exhaustive consistent sets of defects, as done above. It is easily shown, by means of a counter-example, that the reverse of the proposition above does not hold in general. Consider an evidence function e with $\tilde{e}(\{d_1, d_2\}) = \{f\}$, $\tilde{e}(\{d_1, \neg d_2\}) = \{\neg f\}$, and thus $e(\{d_1\}) = \{\neg f, f\}$, assuming that $\tilde{e}$ constitutes a top-down partial specification. For $E = \{\neg f\}$ there exists no CB diagnosis with $H = \{d_1, d_2\}$, but there exists one for $H = \{d_1\}$.

Since the notion of diagnosis CB is defined for any diagnostic specification, it is possible to compare this notion to the causal notions of diagnosis defined at the beginning of this chapter, which were only defined for monotonically increasing evidence functions. We assume now that a special notion of weak-causality diagnosis WC^g is defined for all diagnostic specification. It is also assumed that the original evidence function e, which does not take extra input findings among the set of defects as an argument, is considered.

Proposition 4.13. *Let* WC^g *and* CB *be the notions of weak-causality and consistency-based diagnosis, respectively, then*

$$\mathrm{WC}^g \sqsubseteq \mathrm{CB}$$

Proof. If $\mathrm{WC}^g_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) \supseteq E$. Hence, for each $f \in E$: $f \in e_{|H}(H)$, and $\mathrm{CB}_{\Sigma,e_{|H}}(E) = H$ holds. ◊

4.5 Analysis of hypothetico-deductive diagnosis

As discussed in Section 2.2.4, the hypothetico-deductive approach to diagnostic problem solving originates from the MYCIN system [Shortliffe, 1976]. In MYCIN-like expert systems, diagnostic problem solving is modelled as the process of accepting or rejecting the elements of a finite set of (diagnostic) hypotheses by means of input data and production rules.

4.5.1 Representation and interpretation

Recall from Section 2.2.4, that a hypothetico-deductive diagnostic problem is defined as a pair $\mathcal{H} = (\mathcal{B}, E)$, where $\mathcal{B} = (\Delta, \Phi, \mathcal{R})$ is an associational specification, with:

- $\Delta_{\mathcal{H}}$ a set of ground literals, called *defects* (or disorders).
- $\Phi_{\mathcal{H}}$ a set of ground literals, called *observable findings*.
- $\mathcal{R}$ a *rule base*, i.e. a set of ground implications in predicate logic of the form

  $$c_1 \wedge \cdots \wedge c_m \to d$$

  called *rules*, where each c_i, $1 \leq i \leq m$, $m \geq 1$, is called a *condition* and d is called a *conclusion*, $d \in \Delta_{\mathcal{H}}$; if for a condition c_i there exists a conclusion d in a rule $r \in \mathcal{R}$, with $c_i = d$, then $c_i \in \Delta_{\mathcal{H}}$, else $c_i \in \Phi_{\mathcal{H}}$;
- $E_{\mathcal{H}} \subseteq \Phi_{\mathcal{H}}$ the *set of observed findings*.

Note the difference between the logical implications in the abductive theory of diagnosis and rules in hypothetico-deductive diagnosis. In the abductive theory of diagnosis, findings follow from defects; logical implication has the meaning of a causal relation (cf. Chapter 2). In the theory of hypothetico-deductive diagnosis, defects follow from findings; logical implication has the meaning of a classification relation. In the sequel, the following simplifying assumptions are made until explicitly stated otherwise. It is assumed that rules are definite Horn formulae, i.e. in rules $c_1 \wedge \cdots \wedge c_m \to d$, conditions c_i and conclusions d are positive literals. This is not an unrealistic restriction, because it is adopted in many practical systems. Furthermore, it is assumed that for each positive literal $d \in \Delta_P$, there exists at most one rule $r \in \mathcal{R}$ such that d is a conclusion of r. We will drop these restrictions later on in this section.

A hypothetico-deductive diagnostic problem $\mathcal{H} = (\mathcal{B}, E)$ is mapped to the diagnostic problem $\mathcal{P}$, i.e. $\mathcal{P} = \tau(\mathcal{H})$, by the bijective transformation τ^{vii}, which will be defined presently. To prevent ambiguity, identical symbols from $\mathcal{H}$ and $\mathcal{P}$ will be supplied with a subscript, if necessary. In the following, a collection of rules and observable findings is mapped to an evidence function e. The transformation τ^{vii} is defined as follows:

- $\tau^{\text{vii}}(\Delta_{\mathcal{H}}) = \Delta_{\mathcal{P},P}$ and $\tau^{\text{vii}}(\Delta_{\mathcal{H}}) = \Phi_{\mathcal{P},P}$;
- $\tau^{\text{vii}}(E_{\mathcal{H}}) = E_{\mathcal{P}}$;
- for each $D \subseteq \Delta_{\mathcal{H}}$, and each $E \subseteq \Phi_{\mathcal{H}}$:

  (1) if $\mathcal{R} \cup E \vDash D$, such that for each $E' \subset E$: $\mathcal{R} \cup E' \nvDash D$, then $e(D_{\mathcal{P}}) = F_{\mathcal{P}}$, $F_{\mathcal{P}} = \tau^{\text{vii}}(E)$;

  (2) otherwise, $e(D_{\mathcal{P}}) = \perp$.

  with $D_{\mathcal{P}} = \tau^{\text{vii}}(D)$.

Furthermore, $e(D) = \perp$, for each $D \subseteq \Delta_{\mathcal{P}}$, with $D \cap \Delta_N \neq \varnothing$. Observe that the set of observed findings E will be a *unique* minimum with respect to set inclusion, because $\mathcal{R} \cup E \vDash D$ implies that for each $d \in D$: $\mathcal{R} \cup E_d \vDash d$, $E = \bigcup_{d \in D} E_d$, for some sets $E_d \subseteq \Phi_{\mathcal{H}}$, and if $\mathcal{R} \cup E' \vDash D$ would hold for another minimal set (with respect to set inclusion) E', then for each $d \in D$: $\mathcal{R} \cup E'_d \vDash d$, $E' = \bigcup_{d \in D} E'_d$. This indicates that at least two different rules have the same conclusion d, which contradicts the previously made assumptions. Hence, E constitutes a unique minimal set. Note that this condition would fail to hold if rules were allowed in which non-unique conclusions could appear, as often is encountered in practical knowledge bases.

The resulting evidence function e will be monotonically increasing, because if $D \subseteq D'$, $\mathcal{R} \cup E \vDash D$ and $\mathcal{R} \cup E' \vDash D'$, then $E \subseteq E'$ holds, assuming that E and E' are minimal with respect to set inclusion, because from $\mathcal{R} \cup E' \vDash D'$, it follows that $\mathcal{R} \cup E' \vDash D$. In fact, the resulting evidence function is almost interaction free, which is a consequence of the fact that for positive sets of defects D: $\mathcal{R} \cup E \vDash D$ iff for each $d \in D$, there exists a (minimal) set $E_d \subseteq E$, such that $\mathcal{R} \cup E_d \vDash d$, where $E = \bigcup_{d \in D} E_d$. However, if a set of defects D contains a negative defect, $e(D) = \perp$. From the fact that the evidence function e is almost interaction free, it follows that it can be defined by means of a bottom-up partial specification $\tilde{e}$, which need only be specified for singleton sets with a positive defect, and equals $\perp$ if a set of defects contains a negative defect. We give an example of the application of the transformation.

Example 4.12. Consider the hypothetico-deductive diagnostic problem $\mathcal{H} = (\mathcal{B}, E)$, with $\mathcal{B} = (\Delta, \Phi, e)$, where $\Delta_P = \{d_1, d_2, d_3\}$, $\Phi_P = \{f_1, f_2, f_3\}$, and $\mathcal{R}$ is equal to:

$$f_1 \wedge f_2 \;\to\; d_1$$

$$\begin{aligned} f_2 &\rightarrow d_2 \\ f_3 \wedge d_2 &\rightarrow d_3 \end{aligned}$$

Finally, let $E_{\mathcal{H}} = \{f_1, f_2\}$. The corresponding diagnostic problem $\mathcal{P} = \tau^{\text{vii}}(\mathcal{H})$ is defined as indicated above, where the following bottom-up partial specification $\tilde{e}$ yields the corresponding evidence function e:

$$\tilde{e}(D) = \begin{cases} \{f_1, f_2\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2\} \\ \{f_2, f_3\} & \text{if } D = \{d_3\} \\ \bot & \text{if } \neg d_i \in D, i = 1, 2, 3 \end{cases}$$

Note that $e(\{d_1, d_2\}) = \{f_1, f_2\}$, since e is given by a bottom-up partial specification. $\Diamond$

This function e will be used as a basis for the analysis of hypothetico-deductive diagnosis in terms of our framework.

4.5.2 The notion of associational diagnosis

Recall from Definition 2.7, that a set of defects $D \subseteq \Theta$, with $\Theta \subseteq \Delta_{\mathcal{H}}$ a diagnostic hypothesis, is a *hypothetico-deductive diagnosis* of a diagnostic problem $\mathcal{H} = (\mathcal{B}, E)$ iff $\mathcal{R} \cup E \vDash D$, where D is the maximal subset of Θ with respect to set inclusion that is logically entailed by $\mathcal{R} \cup E$.

Given a diagnostic problem, hypothetico-deductive diagnosis undertakes to derive as many defects as possible from the set of rules to account for the observed findings. The following notion of diagnosis captures this view on diagnostic problem solving for monotonically increasing evidence functions.

Definition 4.9 (*notion of associational diagnosis*). *The notion of* associational diagnosis*, denoted by* AD*, is defined as follows:*

$$\mathrm{AD}_{\Sigma, e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent} \\ u & \text{otherwise} \end{cases}$$

for each diagnostic specification $\Sigma \in \mathcal{S}$ with monotonically increasing evidence function e, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$.

The set of defects H in $\mathrm{AD}_{\Sigma, e_{|H}}$ corresponds to the diagnostic hypothesis Θ distinguished in solving a hypothetic-deductive diagnostic problem. This notion of diagnosis differs from all previously defined notions of diagnosis in that a diagnosis can never be undefined if the given hypothesis is consistent. The notion of associational diagnosis always tries to determine the maximal subset of defects from the hypothesis H that is capable of accounting for a given set of observed findings.

Note that the notion of diagnosis AD satisfies the independence assumption, as is shown in the following proposition.

Proposition 4.14. *The independence assumption holds for the notion of diagnosis* AD*.*

Proof. Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem with monotonically increasing evidence function e. Let $V \subseteq H$ be a subset of the hypothesis $H \subseteq \Delta$. The powerset $\wp(H)$ is partitioned into the set of sets P for which it holds that for each $U \in P$: $U \subseteq V$, and the set of sets P' for which it holds that for each $U \in P'$: $U \not\subseteq V$. Then, according to basic set theory, it holds that:

$$\mathrm{AD}_{\Sigma,e_{|H}}(E) = \bigcup_{\substack{H' \in P \\ e_{|V}(H') \subseteq E}} H' \cup \bigcup_{\substack{H' \in P' \\ e_{|H}(H') \subseteq E}} H'$$

The first component of this union can also be written as $\mathrm{AD}_{\Sigma,e_{|V}}(E)$ if H is consistent. Since e is monotonically increasing, the sets $H' \in P'$ may be changed to $H'' = H' \backslash V$, because if $e(H') \subseteq E$, then $e(H'') \subseteq E$, and because $H' \cap V \subseteq V$, the set $H' \cap V$ is considered in the diagnosis $\mathrm{AD}_{\Sigma,e_{|V}}(E)$. Hence,

$$\mathrm{AD}_{\Sigma,e_{|H}}(E) = \mathrm{AD}_{\Sigma,e_{|V}}(E) \cup \mathrm{AD}_{\Sigma,e_{|H\backslash V}}(E)$$

Since the set V has been selected arbitrarily, AD satisfies the independence assumption. $\Diamond$

Hence, associational diagnosis with respect to a diagnostic specification Σ can also be expressed by the following independent form AD^i:

$$\mathrm{AD}^i_\Sigma = \{(E, (d, F)) \mid E \subseteq \Phi, e(d) = F, F \subseteq E\}$$

Hence,

$$\mathrm{AD}_{\Sigma,e_{|H}}(E) = \{d \in H \mid e(d) \subseteq E\}$$

The independence assumption for subset diagnosis holds a fortiori for diagnostic specifications with a set of defects Δ that is interaction free. A consequence of this proposition is that the notion of diagnosis AD is Δ-monotonic. Observe that AD it is Φ-incomplete, i.e. not every finding in a given set of observed findings E need be accounted for. For example, if $\mathcal{R} = \{f_1 \wedge f_2 \rightarrow d\}$, then $e(\{d\}) = \{f_1, f_2\}$, and

$$A(\mathrm{AD}_{\Sigma,e_{|\{d\}}}(\{f_1\}), \{f_1\}) = \varnothing$$

because $\mathrm{AD}_{\Sigma,e_{|\{d\}}}(\{f_1\}) = \varnothing$. In addition, AD is Δ-incomplete, because $e(D) = \{f, \neg f\} \not\subseteq E$, if E is a set of observed findings. In the following proposition, it is shown that the notion of diagnosis AD is exactly the notion of hypothetico-deductive diagnosis.

Proposition 4.15. *Let $\mathcal{H} = (\mathcal{B}, E)$ be a hypothetico-deductive diagnostic problem and let $\mathcal{P} = (\Sigma, E)$, with $\mathcal{P} = \tau^{\mathrm{vii}}(\mathcal{H})$, $\Sigma = (\Delta, \Phi, e)$, be the corresponding diagnostic problem. Then, D is a diagnosis of $\mathcal{H}$ with hypothesis Θ iff $\mathrm{AD}_{\Sigma,e_{|H}}(E) = H'$, $H' \neq u$, for $H = \tau^{\mathrm{vii}}(\Theta)$ and $H' = \tau^{\mathrm{vii}}(D)$.*

Proof. ($\Rightarrow$): By definition, $D = \{d \in \Theta | \mathcal{R} \cup E \vDash d\}$, because D is a diagnosis. This implies that there exists a set $E' \subseteq E$ that is minimal with respect to set inclusion, such that $\mathcal{R} \cup E' \vDash D$. Using the transformation τ yields $H = \tau^{\mathrm{vii}}(\Theta)$, $H' = \tau(D)$, and $e(H') = E'_{\mathcal{P}}$, where $E'_{\mathcal{P}} = \tau^{\mathrm{vii}}(E'_{\mathcal{H}})$. Since τ is bijective, it follows that $E'_{\mathcal{P}} \subseteq E_{\mathcal{P}} = \tau(E_{\mathcal{H}})$, therefore

$e(H') \subseteq E_{\mathcal{P}}$. Because e is monotonically increasing, it follows that $e_{|H}(H'') \subseteq E_{\mathcal{P}}$, for each $H'' \subseteq H'$. Next, suppose that $H''' \supset H'$, $H''' \subseteq H$, then if $e(H''') \subseteq E_{\mathcal{P}}$, it holds that $H''' \subseteq \mathrm{AD}_{\Sigma,e_{|H}}(E_{\mathcal{P}})$. However, if $e(H''') \subseteq E_{\mathcal{P}}$, then $\mathcal{R} \cup E \vDash D'$, where $D' = \tau(H''')$ with $D' \supset D$. But, then, D would not have been a diagnosis of $\mathcal{H}$; contradiction. Hence, $\mathrm{AD}_{\Sigma,e_{|H}}(E) = H'$.

($\Leftarrow$): If $\mathrm{AD}_{\Sigma,e_{|H}}(E) = H'$, then $e_{|H'}(H') = E'_{\mathcal{P}} \subseteq E_{\mathcal{P}}$, because e is monotonically increasing. By the transformation τ^{vii}, let $\tau^{\mathrm{vii}}(\Theta) = H$, $\tau^{\mathrm{vii}}(D) = H'$ and $\tau^{\mathrm{vii}}(E'_{\mathcal{H}}) = E'_{\mathcal{P}}$, then it follows that $\mathcal{R} \cup E' \vDash D$, $D \subseteq \Theta$, and also $\mathcal{R} \cup E \vDash D$, because $E_{\mathcal{H}} \supseteq E'_{\mathcal{H}}$. From the fact that H' is maximal with respect to set inclusion, it follows that D is a diagnosis. $\Diamond$

Having described the notion of associational diagnosis, next the notion is brought into relation with the notions of diagnosis described in previous sections. It appears that the previously described ordering on notions of diagnosis cannot be extended in a straightforward way, because $\mathrm{WC} \sqsubseteq \mathrm{AD}$ does *not* hold. The reason is that for a set of defects H, it holds that $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$ if $e_H(H) \supseteq E$, whereas $\mathrm{AD}_{\Sigma,e_{|H}}(E) = H$ if $e_H(H) \subseteq E$. However, the following relationship is satisfied.

Proposition 4.16. *Let* SC *and* AD *be the notions of strong causality and associational diagnosis, respectively, then*

$$\mathrm{SC} \sqsubseteq \mathrm{AD}$$

Proof. If $\mathrm{SC}_{\Sigma,e_{|H}}(E) = H \neq u$, then $e_{|H}(H) = E$, therefore $e_{|H}(H) \subseteq E$, hence $H = \mathrm{AD}_{\Sigma,e_{|H}}(E)$. $\Diamond$

Note that the proposition only holds if in the corresponding hypothetico-deductive diagnostic problem every observed finding occurs as a condition of, at least, one successful rule.

The relationship between AD and CB deserves some attention. Neither of the two notions of diagnosis is a restriction of the other. Applying the diagnostic problem given above, where $e(\{d\}) = \{f_1, f_2\}$, yields $\mathrm{AD}_{\Sigma,e_{|\{d\}}}(\{f_1\}) = \varnothing$, where $\mathrm{CB}_{\Sigma,e_{|\{d\}}}(\{f_1\}) = \{d\}$. However, these two notions of diagnosis stand in subdiagnostic relation to each other.

Proposition 4.17. *Let* CB *and* AD *be the notions of consistency-based diagnosis and associational diagnosis, then*

$$\mathrm{AD} \trianglelefteq \mathrm{CB}$$

Proof. If $\mathrm{AD}_{\Sigma,e_{|H}}(E) = H' \neq u$, then $e_{|H}(H') \subseteq E$, hence, for each $f \in E$: $f \in e_{|H}(H)$ or $\neg f \notin e_{|H}(H)$, because if $\neg f \in e_{|H}(H)$, then $\mathrm{CB}_{\Sigma,e_{|H}}(E) = u$. From this the proposition follows. $\Diamond$

Until now, we have assumed that only positive literals were allowed in the logical representation of conditions and conclusions in rules, that conclusions in rules were unique. One of the characteristics of knowledge bases containing empirical knowledge is the availability of several different rules to express alternative empirical characterizations of a given defect. Furthermore, conclusions in rules may also be negative literals. The consequences of these extensions with respect to our framework are investigated next.

Basically, the association of several sets of findings with each set of defects, which is the result of the possible existence of more than one rule concerning the same defect, can be encoded in an evidence function of a diagnostic specification by introducing a unique defect symbol $d \in \Delta$ for every conclusion in a rule. A more natural representation would be to employ a structural evidence function e^s (cf. Section 3.1.2), which is obtained by adapting the transformation τ^{viii}, which maps a hypothetico-deductive diagnostic problem $\mathcal{H}$ to a diagnostic problem $\mathcal{P}$, as follows. For each $D \subseteq \Delta_{\mathcal{H}}$:

(1) for each $E \subseteq \Phi_{\mathcal{H}}$: if $D = \{d \in \Theta \mid \mathcal{R} \cup E \vDash d\}$, and for each $E' \subset E$: $\mathcal{R} \cup E' \nvDash D$, then $\tau^{\text{viii}}(E) \in e^s(H)$;

(2) otherwise, if for each $E \subseteq \Phi_{\mathcal{H}}$: $\mathcal{R} \cup E \nvDash D$, then $e^s(H) = \perp$,

with $\tau^{\text{viii}}(D) = H$. The function values $e^s(H)$ is assumed minimal with respect to set inclusion for each $H \subseteq \Delta_{\mathcal{P}}$. The resulting structural evidence function is called monotonically increasing, if for each $H \subseteq H'$:

$$\bigcup_{E \in e^s(H)} E \subseteq \bigcup_{E \in e^s(H')} E$$

A straightforward generalization of the previously given notion of associational diagnosis is presented in the following definition.

Definition 4.10 (*structural associational diagnosis*). *Let $\Sigma = (\Delta, \Phi, e^s)$ be a diagnostic specification with a structural, monotonically increasing evidence function e^s. The notion of* structural associational diagnosis, *denoted by* SAD, *is then defined as follows*

$$\text{SAD}_{\Sigma, e^s_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ \exists E' \in e^s_{|H}(H'),\, E' \subseteq E}} H' & \text{if } H \text{ is consistent} \\ u & \text{otherwise} \end{cases}$$

for each set of observed findings $E \subseteq \Phi$, and for each $D \subseteq \Delta$.

Structural associational diagnosis increases diagnostic flexibility in comparison with associational diagnosis using an unstructural, evidence function e. This is evident from the observation that if e is defined by e^s as follows:

$$e(D) = \bigcup_{E \in e^s(D)} E$$

for each consistent set $D \subseteq \Delta$, it holds that

$$\text{AD}_{\Sigma, e_{|H}}(E) \subseteq \text{SAD}_{\Sigma', e^s_{|H}}(E)$$

when both diagnoses are defined.

4.6 Discussion

In this chapter, several notions of diagnosis, as presented in the literature, have been interpreted in terms of our set-theoretical framework of diagnosis, rendering them amenable to a uniform analysis. Four basic notions of diagnosis have been analysed in this way, namely abductive diagnosis, set-covering diagnosis, consistency-based diagnosis and hypothetico-deductive diagnosis. The translation of these formal theories of diagnosis into our framework revealed implicit properties of problem representations used for diagnosis, and of the underlying notions of diagnosis. Based on an analysis of theories of diagnosis, several formal notions of diagnosis have been defined in terms of our framework: (predictive) weak and strong causality diagnosis, consistency-based diagnosis and associational diagnosis. The relationships between these notions of diagnosis have been investigated as well.

The analysis revealed that the knowledge used in the various notions of diagnosis differs in several respects. It was already known from the literature on diagnosis that the conceptual basis of the various theories of diagnosis differs considerably. In this chapter, differences and similarities have been identified in a precise, mathematical fashion.

Analysis of the notions of diagnosis underlying these formal theories indicated that it is not possible to construct a total order in terms of a restriction relation, as proposed in Chapter 3, in which all notions of diagnosis discussed participate. Although, some notions of diagnosis could indeed be viewed as restrictions of others, the relationships among notions of diagnosis is more complicated than is often suggested in the literature. For example, the often cited statement: "covering is a stronger requirement than consistency" (cf. [Console & Torasso, 1991] and [Benjamins & Jansweijer, 1994]), which states that the set of diagnoses produced by the notions of weak or strong causality is a subset of the set of diagnoses produced by consistency-based diagnosis, expresses just one relationship between two notions of diagnosis. There are many other features that need investigation. Any comparison between notions of diagnosis requires detailed information about the nature of the diagnostic interpretation of a knowledge base, or evidence function in our terminology, a subject that has not been given proper attention in the literature on knowledge acquisition and modeling.

In [Reiter, 1987], R. Reiter shows that abductive diagnosis can be mapped to his notion of consistency-based diagnosis (cf. Section 2.2.5). However, in the resulting theory of consistency-based diagnosis, no distinction is made between problem representation and diagnostic interpretation; in fact, it is not clear from his result that the notion of abductive diagnosis studied corresponds to our notion of weak causality diagnosis only, not to abductive diagnosis in general.

In their "spectrum of logical definitions of diagnosis", Console and Torasso argue that any notion of diagnosis can be expressed by varying in the consistency condition [Console & Torasso, 1991]. However, as has been shown in this chapter, some of the aspects of diagnosis may be hidden in the problem representation, and not be revealed using the consistency condition, while others cannot be expressed easily in terms of satisfiability. Our analysis indicates that, although satisfiability has been successfully used for defining notions of diagnosis, it is not a suitable concept for clarifying the characteristics of notions of diagnosis in general. The same can be said of logical entailment.

Chapter 5

Refinement Diagnosis

In Chapter 3, a set-theoretical framework of diagnosis was developed, and used in Chapter 4 for the analysis of several formal theories of diagnosis presented in the literature. As shown in Chapter 4, these notions of diagnosis have built-in assumptions about properties of problem domains, captured in our framework in terms of an evidence function and notion of diagnosis. In the present chapter, some of these assumptions will be relaxed, yielding notions of diagnosis that are more flexible than most of the notions of diagnosis discussed in the previous chapter.

5.1 Motivation

With the exception of associational diagnosis, all notions of diagnosis discussed in the previous chapter were defined in such a way that a given diagnostic hypothesis H was either accepted or rejected, yielding the undefined diagnosis in the latter case. In contrast, a principle of *refinement* of a diagnostic hypothesis was incorporated in the definition of associational diagnosis. Associational diagnosis expresses that the least upper bound of accepted subhypotheses of a given diagnostic hypothesis H will be accepted as the diagnostic solution of a diagnostic problem. Thus, the hypothesis may also be adjusted or refined, and is not simply rejected or accepted, although in the refinement process certain elements may be removed from the hypothesis. Generally speaking, refinement is the construction of a diagnosis that is the 'best' possible in some particular sense, given the domain knowledge, the set of observed findings and the hypothesis at hand. The resulting notions of diagnosis differ from the rigorous 'accept-or-reject' notions of diagnosis described in the previous chapter.

There are various reasons why refinement diagnosis may be a more appropriate basis for diagnostic problem solving than the rigorous notions of diagnosis previously described:

- Real-world knowledge bases are, almost without exception, incomplete, i.e. the modelled problem domain has not been fully described. For example, knowledge of certain interactions among defects may be missing.
- Real-world knowledge bases are not completely accurate, e.g. the meaning of the domain knowledge may not have been captured sufficiently precisely, or may have been specified incorrectly.

- The findings that may be observed, and interpreted by an expert system, are only part of what might have been collected without limitations, such as available time and money.

- Part of the observed findings may be unreliable, due to impediments to the observation process, such as limited available time.

The extent to which these problems are encountered in practical systems may depend on the conceptual approach adopted in developing a knowledge-based system, as reviewed in Chapter 2. Although a model-based approach is often thought to shield the developer from such problems (cf. [Reiter, 1987]), making simplifying assumptions will always be necessary in order to deal with real-world problems, whether a model-based or an associational diagnosis approach is followed. Yet, to manage the consequences of a problem, it must be solved first, i.e. it is often essential to establish a diagnosis, even when confronted with the imperfections mentioned above. In many domains, in particular medicine, it is usually better to arrive at a diagnosis that does not account for all observed findings, or that suggests findings that have not been observed, than to establish no diagnosis at all. It is sometimes said that such a diagnosis *underaccounts* or *overaccounts* for the set of observed findings. To emphasize the fact that a diagnosis accounts for all observed findings, we shall sometimes say that a diagnosis *strictly* accounts for the observed findings.

The impact of the above-mentioned problems on the applicability of a system may also be examined using an experimental approach. By validating such a system, valuable insight into its diagnostic quality is obtained. Validation offers some protection against the unjustified use of an imperfect knowledge base, but does not provide full guarantee of completeness and accuracy of an expert system. The issue of validation is dealt with in the second part of this thesis (Chapter 8), where the validation of expert systems in general, and the HEPAR system in particular, is discussed.

In the framework of diagnosis introduced in Chapter 3, adjustment of a diagnostic hypothesis was explicitly stated as a possibility. Notions of diagnosis in which adjustment is included as a basic mechanism will be referred to as notions of *refinement diagnosis*. As far as known to the author, similar notions of diagnosis have not appeared in the literature before.

The following question now arises: what can be taken as a basis for notions of diagnosis which incorporate certain principles of refinement? Obviously, there exists a wide range of possibilities. Which of the possible choices yields the most natural result depends, to a large extent, on the nature of the problem domain, which is partially expressed by the characteristics of the evidence functions e. Dependencies between a notion of diagnosis R, on the one hand, i.e. the interpretation of the set of observed findings given a specific knowledge base, and properties of a given evidence function e, on the other hand, exist. These dependencies will be encountered repeatedly in the subsequent sections.

Two classes of refinement diagnosis will be studied. Firstly, the class of notions of refinement diagnosis, called most general diagnosis, is examined, where the least upper bound of accepted hypotheses (with respect to set inclusion) is taken as a diagnostic solution. Secondly, the class of notions of refinement diagnosis, called most specific diagnosis, based on taking the greatest lower bound of accepted hypotheses is studied.

5.2 Most general diagnosis

In this section, the concept of refinement diagnosis is defined as the least upper bound with respect to set inclusion of accepted subhypotheses H' of a given hypothesis H. Recall that a subhypothesis H' is accepted if some predefined relationship between the set of observable findings $e_{|H}(H')$ and the set of observed findings E is satisfied. The instances of refinement diagnosis that will be considered, are called notions of *most general diagnosis*. These notions of diagnosis capture the idea that if a specific diagnostic hypothesis is not accepted, then the 'nearest' subhypothesis should be taken instead. The least upper bound with respect to set inclusion of the set of accepted subhypotheses is an example of such a nearest subhypothesis. If hypotheses are accepted, the following holds for notions of most general diagnosis G: if $G_{\Sigma,e_{|H}}(E) = H$ and $G_{\Sigma,e_{|H'}}(E) = H'$ then $G_{\Sigma,e_{|H\cup H'}}(E) = H \cup H'$. Thus, the notions of most general diagnosis enforce independence or compositionality of diagnostic components. In the following, several examples of notions of most general diagnosis will be discussed.

There are three notions of refinement diagnosis that follow more or less naturally from the basic principles of diagnosis introduced in Chapter 3, and the specific notions of diagnosis analysed in Chapter 4. These notions are: most general subset, superset and intersection diagnosis, where set inclusion and set intersection, respectively, are applied as a basis for the construction of a diagnosis. These notions are first defined, after which the consequences with respect to the possible meaning of an evidence function to which a notion of diagnosis is applied, are investigated. Each notion of diagnosis actually stands for several different notions of diagnosis, because the characteristics of these notions will be investigated with respect to the properties of various classes of evidence functions to which the particular notion of diagnosis can be restricted.

5.2.1 Most general subset diagnosis

In the first notion to be formally defined, every finding associated with a collection of defects must be included among the observed findings if the defects in the collection are to be admitted as part of a diagnosis. A similar notion of diagnosis, called associational diagnosis, denoted by AD, was discussed in the previous chapter. Associational diagnosis was shown to be a restrictive version of hypothetico-deductive diagnosis as employed in rule-based systems, captured by means of an unstructural evidence function. Here, the restriction to monotonically increasing evidence functions is relaxed, and the conditions under which a diagnosis can be undefined is extended.

Definition 5.1 (*most general subset diagnosis*). *The notion of* most general subset diagnosis*, denoted by* GS*, is defined as follows:*

$$\mathrm{GS}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \subseteq E \\ u & \text{otherwise} \end{cases}$$

for each $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$. The diagnostic solution $\mathrm{GS}_{\Sigma,e_{|H}}(E)$ *of a diagnostic problem $\mathcal{P} = (\Sigma, E)$ is called the* most general subset diagnosis of $\mathcal{P}$ with respect to H.

From the definition of the concept of evidence function (cf. Definition 3.4), it follows that the set of defects resulting from applying the notion of most general subset diagnosis to a diagnostic problem will be consistent, because the starting hypothesis H is assumed to be consistent. The most general subset diagnoses with respect to Δ_P and Δ_N are called the positive and negative most general subset diagnosis, respectively. In contrast with the notions of diagnosis, except associational diagnosis, in the previous chapter, in which a hypothesis was either accepted or rejected (yielding u as a result), it may be sufficient to establish a single diagnosis $\mathrm{GS}_{\Sigma,e_{|H}}(E)$ for a given hypothesis H, because each subhypothesis H' of H is investigated. The definition above indicates that when applying the notion of subset diagnosis, diagnostic problem solving amounts to finding the least upper bound with respect to set inclusion of sets of defects. Associated findings are all included in the set of observed findings. Intuitively, this means that most general subset diagnosis is the smallest set of defects that includes all accepted subhypotheses of a given hypothesis. It may be viewed as a refinement approach to diagnosis; of primary importance is the construction of a diagnosis as general as possible with respect to a given hypothesis, that includes as many of the defects suggested by the given hypothesis as possible. This may be a suitable approach in domains in which neglecting a particular defect may be dangerous.

The following straightforward lemma will be used below.

Lemma 5.1. *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. If $e_{|H}(H) \subseteq E$, then $\mathrm{GS}_{\Sigma,e_{|H}}(E) = H$, for each $H \subseteq \Delta$.*

Proof. Immediate from Definition 5.1. ◊

Although for the individual accepted subhypotheses $H' \subseteq H$ comprising the resulting diagnosis $\mathrm{GS}_{\Sigma,e_{|H}}(E) = H''$ it holds that $e_{|H}(H') \subseteq E$, it need not hold that H'' strictly accounts or underaccounts for E. This will only hold under certain conditions discussed below.

Example 5.1. Consider a diagnostic problem $\mathcal{P} = (\Sigma, E)$, where the evidence function e is defined as follows:

$$e(D) = \begin{cases} \varnothing & \text{if } D = \varnothing \\ \{f_1\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2\} \\ \{f_1, f_2, f_3\} & \text{if } D = \{d_1, d_2\} \\ \bot & \text{otherwise} \end{cases}$$

Furthermore, the set of observed findings is equal to $E = \{f_1, f_2, f_4\}$. Then, $\mathrm{GS}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$, although $e(\{d_1, d_2\}) \not\subseteq E$. Note that the observation of merely the findings f_1 and f_2, and f_4 instead of f_3, seems incorrect if e is interpreted as representing strong causality (cf. Definition 4.1), yielding $\mathrm{SC}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = u$ as a diagnosis, i.e. SC tells us that it is impossible to have observed the set of findings E. The diagnosis resulting from GS, $\{d_1, d_2\}$, does not account for every finding in E, and does predict a finding, f_3, that has not been observed. This diagnosis has been concluded by assuming the set $\{d_1, d_2\}$ to be interaction free, otherwise no diagnosis would have been established. It seems as if the 'knowledge base' has been adapted in order to solve the diagnostic problem. ♢

With regard to the real world, the basic assumption underlying the notion of most general subset diagnosis is that when a set of defects occurs (corresponding to a subhypothesis), all associated findings are assumed to be observed (ignoring the feature that an evidence function can be adapted to a diagnostic problem). Observe the similarity with the notion of strong-causality diagnosis SC introduced in Section 4.2.1 in connection with the abductive theory of diagnosis. In contrast with this notion of diagnosis, most general subset diagnosis need not account for all observed findings. Instead, as many observed findings are accounted for as possible, given the hypothesis at hand, possibly after adjustment. The resulting diagnosis, however, may also be associated with findings other than those observed, as demonstrated in the example above. It is a consequence of ignoring particular interactions among defects. This may not always be acceptable. However, there is a 'security measure': a diagnosis is the result of accepted subhypotheses H' for which the set-inclusion relationship between $e_{|H}(H')$ and E is satisfied. A set of findings associated with any collection of defects may not include complementary elements $f, \neg f$ because such a set of defects can never be diagnosed using subset diagnosis.

Recall that an evidence function that can be defined using a bottom-up partial specification, in particular, a monotonically increasing evidence function, can be used to represent a causal domain model. If a monotonically increasing evidence function is used for establishing a diagnosis $\mathrm{GS}_{\Sigma, e_{|H}}(E) = H' \neq u$, then, of course, for each set $H'' \subseteq H$ that contributes to H' it holds that $e_{|H}(H'') \subseteq E$. If the evidence function e is interaction free, then the set of defects H' is the unique largest subset of H for which $e_{|H}(H') \subseteq E$; the set of findings accounted for by the diagnosis H' is equal to $e_{|H}(H') = E'$. Any partitioning of the set H', including a partitioning into singleton sets, will account for the same set of findings E' as H'. More formally, let P be a partitioning of H', then $e(H') = \bigcup_{V \in P} e(V)$. Thus, the set of findings $e(H')$ may be interpreted as being collectively accounted for by the elements of the partitioning P. For monotonically increasing evidence functions, whether or not interaction free, nothing more specific can be said concerning the set H' than stated in Lemma 5.1 for any evidence function. If the evidence function is monotonically increasing, however, it is possible to obtain a more precise characterization of accepted subhypotheses H' of a given hypothesis H.

Proposition 5.1. *If e is a monotonically increasing evidence function from a diagnostic specification Σ, then*

$$e_{|H}(H') \subseteq A(\mathrm{GS}_{\Sigma, e_{|H}}(E), E)$$

if $\mathrm{GS}_{\Sigma, e_{|H}}(E) \neq u$, $H' \subseteq H$ *and* $e_{|H}(H') \subseteq E$.

Proof. Let $H'' = \mathrm{GS}_{\Sigma, e_{|H}}(E)$, $H'' \neq u$. For each $H' \subseteq H$ with $e_{|H}(H') \subseteq E$, it holds that $H' \subseteq H''$. By the fact that e is monotonically increasing, it follows that $e_{|H}(H') \subseteq e_{|H}(H'')$. Hence, $e_{|H}(H') \subseteq e_{|H}(H'') \cap E = A(H'', E)$. ♢

This means that when the evidence function is monotonically increasing, every subhypothesis H' that contributes to the most general subset diagnosis has a set of associated findings that is included in the set of observed findings $E' \subseteq E$ accounted for by $\mathrm{GS}_{\Sigma,e_{|H}}(E)$.

Recall that an evidence function that can be defined by a top-down partial specification, in particular, an evidence function that is monotonically decreasing, can be used to represent knowledge regarding normal and abnormal functional behaviour, such as regarding an electronic circuit, or some other device. The following property holds in this case.

Lemma 5.2. *Let Σ be a diagnostic specification with monotonically decreasing evidence function. If $\mathrm{GS}_{\Sigma,e_{|H}}(E) = H'$, with $H' \neq u$, then $H' = H$ and $e_{|H}(H') \subseteq E$.*

Proof. If there exists a set $H' \subseteq H$, such that $e_{|H}(H') \subseteq E$, then also $e_{|W}(H) \subseteq E$ by the fact that e is monotonically decreasing. ♢

This lemma indicates that this notion of refinement diagnosis behaves like the 'accept-or-reject' notions of diagnosis from the previous chapter. In general, most general subset diagnosis is not a suitable notion of diagnosis for such specifications of knowledge, because, as discussed in Section 3.1, the function values $e(D)$ may contain complementary findings, reflecting nondeterministic output. It may be acceptable in a domain to reject hypotheses H with nondeterministic findings in the function value $e(H)$. Then, most general subset diagnosis is a specific form of consistency-based diagnosis (CB), restricted to monotonically decreasing evidence functions, because any set of findings $e(D)$ that is included in a set of observed findings E is also consistent with such set. Under these conditions, there is little difference between consistency-based diagnosis restricted to monotonically decreasing evidence functions and most general subset diagnosis.

Example 5.2. Consider again the two-inverter problem from Example 4.9. From the evidence function value

$$e(\{i, n_1, n_2\}) = \{o\}$$

it follows that

$$\mathrm{GS}_{\Sigma,e_{|\{i,n_1,n_2\}}}(\{o\}) = \{i, n_1, n_2\}$$

which is identical to the consistency-based diagnosis

$$\mathrm{CB}_{\Sigma,e_{|\{i,n_1,n_2\}}}(\{o\}) = \{i, n_1, n_2\}$$

When a consistency-based diagnosis is undefined, for example

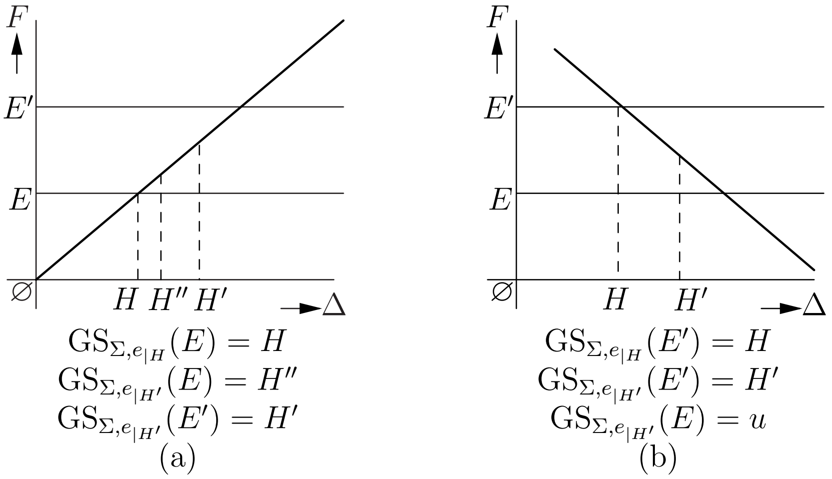
$$\mathrm{CB}_{\Sigma,e_{|\{i,n_1\}}}(\{o\}) = u$$

because $e(\{i, n_1\}) = \{o, \neg o\}$ and the set of observed findings is inconsistent with $\{o, \neg o\}$, then the most general subset diagnosis is undefined as well:

$$\mathrm{GS}_{\Sigma,e_{|\{i,n_1\}}}(\{o\}) = u$$

♢

This example illustrates the following general result.



$\mathrm{GS}_{\Sigma,e_{|H}}(E) = H$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E) = H''$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E') = H'$
(a)

$\mathrm{GS}_{\Sigma,e_{|H}}(E') = H$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E') = H'$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E) = u$
(b)

Figure 5.1: Monotonically increasing (a) and decreasing (b) evidence functions.

Proposition 5.2. *Let* CB *and* GS *be the notions of consistency-based and most general subset diagnosis, respectively. Let* Σ *be a diagnostic specification with monotonically decreasing evidence function* e*. If* $\mathrm{GS}_{\Sigma,e_{|H}}(E) = H$*,* $H \neq u$*, then* $\mathrm{CB}_{\Sigma,e_{|H}}(E) = H$*; if* $\mathrm{CB}_{\Sigma,e_{|H}}(E) = u$*, then* $\mathrm{GS}_{\Sigma,e_{|H}}(E) = u$*.*

Proof. From Lemma 5.2, it follows that if $\mathrm{GS}_{\Sigma,e_{|H}}(E) = H$, $H \neq u$, then $e_{|H}(H) \subseteq E$; hence, $\mathrm{CB}_{\Sigma,e_{|H}}(E) = H$. If $\mathrm{CB}_{\Sigma,e_{|H}}(E) = u$, then there must be at least one finding f with $f \in E$ and $\neg f \in e(H)$. Therefore, $e_{|H}(H) \not\subseteq E$. By the fact that e is monotonically decreasing, it follows that for none of the subsets $H' \subseteq H$ it holds that $e_{|H}(H') \subseteq E$. From this, the result follows. ◊

In Figure 5.1, the relationship between diagnostic hypothesis H, the set of observed findings E and the resulting diagnosis $\mathrm{GS}_{\Sigma,e_{|H}}(E)$ is summarized by schematically depicting these sets as if they were real numbers and by taking set inclusion as the $\leq$ total order on the real numbers. The independent variable is a set D of defects; sets of observable findings E act as the dependent variable, with $E = e(D)$. Hypotheses with associated diagnoses, designated by H, H' or H'', have been inserted into the figure with associated sets of findings $e(H)$, $e(H')$ and $e(H'')$, indicated by dashed lines. The diagnoses are the result of the notion of most general subset diagnosis applied to fixed sets of observed findings E and E', also included in the figure. As discussed above, if most general subset diagnosis is applied to a monotonically decreasing evidence function, the resulting diagnosis is either undefined or equal to the given hypothesis H. This contrasts with GS applied to a monotonically increasing evidence function, which may also yield subsets of the hypothesis as a diagnosis. $\mathrm{GS}_{\Sigma,e_{|H'}}(E) = H''$ in Figure 5.1.(a) is intended to illustrate that $e(H'')$ may even be a superset of E.

If the evidence function e is nonmonotonic, then the relationships between E and $e_{|H}(H')$ are investigated as before, but again, certain interactions between defects may be ignored, or modified. Consider the following real-life medical example.

Example 5.3. The diagnostic specification $\Sigma = (\Delta, \Phi, e)$, with $\Delta_P = \{d_1, d_2, d_3\}$ expresses knowledge concerning the following three disease processes:

$$\begin{aligned} d_1 &= \text{atherosclerosis} \\ d_2 &= \text{myocardial ischaemia} \\ d_3 &= \text{hypothyroidism} \end{aligned}$$

It is known that diminished function of the thyroid gland (called hypothyroidism), may produce atherosclerosis. Atherosclerosis may cause myocardial ischaemia (insufficient blood and oxygen supply to the heart muscle), producing retrosternal chest pain. Myocardial ischaemia may occur when the oxygen demand of the heart is increased, but, due to atherosclerosis, the vascular system cannot comply with this demand. Typically, patients with hypothyroidism display slow physical and mental activity; their hearts require less oxygen. As a consequence, myocardial ischaemia need not arise in these patients or is less severe, even when the atherosclerosis is considerable. Suppose that the set of positive observable findings Φ_P consists of the following elements:

$$\begin{aligned} f_1 &= \text{retrosternal chest pain} \\ f_2 &= \text{slow activity} \\ f_3 &= \text{coldness complaints} \end{aligned}$$

Then, the evidence function e expresses formally what has been stated informally above:

$$e(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\}, \{d_2\}, \{d_1, d_2\} \\ \{f_2, f_3\} & \text{if } D = \{d_3\}, \{d_2, d_3\}, \{d_1, d_3\}, \{d_1, d_2, d_3\} \\ \bot & \text{if } D \text{ is syntactically inconsistent} \\ \varnothing & \text{otherwise} \end{cases}$$

i.e. retrosternal chest pain (f_1) is a finding that may be observed in a patient with atherosclerosis (d_1) and also in atherosclerosis combined with myocardial ischaemia (d_1, d_2), etcetera. This evidence function expresses cancellation, because $e(\{d_1, d_2, d_3\}) \subset e(\{d_1\}) \cup e(\{d_2, d_3\})$ (cf. Section 3.1).

If $E = \{f_1\}$, then $\mathrm{GS}_{\Sigma, e_{|\{d_1,d_2,d_3\}}}(E) = \{d_1, d_2\}$. This seems correct because $e(\{d_1, d_2\}) = E$. From a medical point of view, there is no evidence for hypothyroidism (d_3); rightfully, d_3 is not included in the diagnosis. (Recall that this means that it is unknown whether the third defect is present, d_3, or absent, $\neg d_3$.) When $E' = \{f_1, f_2, f_3\}$ the most general subset diagnosis is equal to $\mathrm{GS}_{\Sigma, e_{|\{d_1,d_2,d_3\}}}(E') = \{d_1, d_2, d_3\}$. The diagnosis underaccounts for E (f_1 is not accounted for), but by assuming that cancellation of findings did not arise, every finding can be accounted for. From a medical point of view, it can be said that although hypothyroidism usually suppresses symptoms of myocardial ischaemia, which is represented in e, in the present patient the symptom did arise. The diagnosis correctly includes all disease processes as a diagnosis in this case, although the set of observed findings does not match the function value $e(\{d_1, d_2, d_3\})$. ◊

Again, this indicates that when for most general subset diagnosis a set of observed findings cannot be related in a simple way to a function value of an evidence function, a kind of adaptive interpretation of the knowledge base results by ignoring certain interactions among defects. In practical circumstances, this may be an acceptable approach, as has been motivated above.

In Chapter 4, the independence assumption was shown to be satisfied for associational diagnosis (Proposition 4.14). Likewise, the independence assumption is satisfied for GS if GS is restricted to diagnostic specifications with a monotonically increasing evidence function. (The proposition is almost identical, except that there are more situations in which GS is undefined.) However, if the evidence function e is not monotonically increasing, then the independence assumption is not satisfied. Hence, the independence assumption fails to hold in general for most general subset diagnosis, as shown by the following counter-example.

Example 5.4. Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, which includes, amongst others, the following function values $e(\varnothing) = \varnothing$, $e(d_1) = \{f_1\}$, $e(d_2) = \{f_2\}$ and $e(\{d_1, d_2\}) = \{f_3\}$. This nonmonotonic evidence function expresses 'cancellation and augmentation' (cf. Section 3.1). Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem with $E = \{f_3\}$. Obviously, because

$$\mathrm{GS}_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_3\}) \neq \mathrm{GS}_{\Sigma,e_{|\{d_1\}}}(\{f_3\}) \cup \mathrm{GS}_{\Sigma,e_{|\{d_2\}}}(\{f_3\})$$

the independence assumption fails to hold for most general subset diagnosis GS. $\Diamond$

Hence, it is not possible to decompose a knowledge base into separate components, and to apply GS to each component separately, yielding the same result as for the entire knowledge base. However, most general subset diagnosis is Δ-monotonic, as proven in the following proposition.

Proposition 5.3. *The notion of most general subset diagnosis* GS *is* Δ*-monotonic.*

Proof. If $H \subseteq H'$, then $\mathrm{GS}_{\Sigma,e_{|H}}(E) \subseteq \mathrm{GS}_{\Sigma,e_{|H'}}(E)$ given that $\mathrm{GS}_{\Sigma,e_{|H}}(E), \mathrm{GS}_{\Sigma,e_{|H'}}(E) \neq u$, because if $e_{|H}(H'') \subseteq E$, $H'' \subseteq H$, then $e_{|H'}(H'') \subseteq E$. $\Diamond$

It is not always possible to find a set of observed findings E and a set of defects H, for a consistent set of defects $D \subseteq \Delta$, such that $\mathrm{GS}_{\Sigma,e_{|H}}(E) = D$ – e.g. choose $e(D) = \{f, \neg f\}$; it follows that most general subset diagnosis is Δ-incomplete. Because associational diagnosis AD was shown to be Φ-incomplete, most general subset diagnosis is Φ-incomplete as well.

If strong-causality diagnosis SC (Definition 4.1) is generalized to SC^g, which differs from SC only by being defined for any diagnostic specification $\Sigma \in \mathcal{S}$, and not only for diagnostic specifications with monotonically increasing evidence functions (the implicit restriction of logical abductive notions of diagnosis when adopting standard logical entailment), the next result, which is a generalization of Proposition 4.16, follows immediately.

Proposition 5.4. *Let* SC^g *and* GS *be the notions of generalized strong-causality diagnosis and most general subset diagnosis, respectively, then*

$$\mathrm{SC}^g \sqsubseteq \mathrm{GS}$$

Proof. Simply note that if $\mathrm{SC}^g_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) \subseteq E$; hence, by Lemma 5.1, it follows that $\mathrm{GS}_{\Sigma,e_{|H}}(E) = H$. $\Diamond$

From the discussion above, it follows that the relations between most general subset diagnosis and other notions of diagnosis also depend on the properties of evidence functions.

For monotonically decreasing evidence functions, most general subset diagnosis is much like consistency-based diagnosis, whereas if the notion of most general subset diagnosis is restricted to monotonically increasing evidence functions, it resembles strong-causality diagnosis.

5.2.2 Most general superset diagnosis

Where most general subset diagnosis can be viewed as a more flexible version of strong-causality diagnosis, which for certain evidence functions is as little restrictive as consistency-based diagnosis, a similar, flexible notion of diagnosis can be designed for weak-causality diagnosis. This suggests replacing the subset relation in most general subset diagnosis by the superset relation, yielding the notion of most general superset diagnosis GO (the letter 'O' stands for 'cOntains').

Definition 5.2 (*most general superset diagnosis*). *The notion of* most general superset diagnosis*, denoted by* GO*, is defined as follows:*

$$\mathrm{GO}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \supseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \supseteq E \\ u & \text{otherwise} \end{cases}$$

for each $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$. The diagnostic solution $\mathrm{GO}_{\Sigma,e_{|H}}(E)$ of a diagnostic problem $\mathcal{P} = (\Sigma, E)$ is called the most general superset diagnosis of $\mathcal{P}$ with respect to *H.*

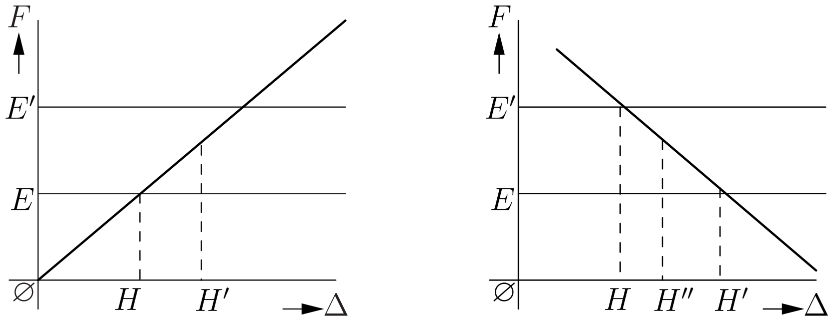
A most general superset diagnosis for a consistent hypothesis H is always consistent. This is a consequence of the manner in which the evidence function e has been defined (cf. Definition 3.4). The diagnostic solutions $\mathrm{GO}_{\Sigma,e_{|H}}(E)$, with $H = \Delta_P$ or $H = \Delta_N$, are simply called the positive and negative most general superset diagnosis, respectively. According to the definition above, most general superset diagnosis is taken as the smallest subset of defects that includes all satisfied subhypotheses H' of the hypothesis H accounting for all observed findings. As for most general subset diagnosis, the notion of most general superset diagnosis ignores certain interactions among defects if necessary to achieve a diagnosis. The following lemma concerns a basic property of most general superset diagnosis.

Lemma 5.3. *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. If $e_{|H}(H) \supseteq E$, then $\mathrm{GO}_{\Sigma,e_{|H}}(E) = H$, for each $H \subseteq \Delta$.*

Proof. The proof follows immediately from the definition above. ◊

Note that, in contrast to most general subset diagnosis, a set of defects D for which $f, \neg f \in e(D)$ can be included in a diagnosis. GO is Δ-complete, but is Φ-incomplete. The reason is that a diagnosis $H' = \mathrm{GO}_{\Sigma,e_{|H}}(E)$ need not satisfy the relation $e_{|H}(H') \supseteq E$ (see below). The independence assumption is not generally satisfied for most general superset diagnosis, but most general superset diagnosis is Δ-monotonic. Both results follow from straightforward modification of Example 5.4 and Proposition 5.3.

Next, the relationship between most general superset diagnosis and characteristics of the evidence function is investigated.



$\mathrm{GO}_{\Sigma,e_{|H}}(E) = H$
$\mathrm{GO}_{\Sigma,e_{|H'}}(E') = u$
$\mathrm{GO}_{\Sigma,e_{|H'}}(E) = H'$
(a)

$\mathrm{GO}_{\Sigma,e_{|H}}(E') = H$
$\mathrm{GO}_{\Sigma,e_{|H'}}(E') = H''$
$\mathrm{GO}_{\Sigma,e_{|H''}}(E) = H''$
(b)

Figure 5.2: Monotonically increasing (a) and decreasing (b) evidence functions.

Proposition 5.5. *Let e be a monotonically increasing evidence function from a diagnostic specification Σ. If $\mathrm{GO}_{\Sigma,e_{|H}}(E) = H'$, $H' \neq u$, then it holds that $H = H'$ and $e_{|H}(H) \supseteq E$.*

Proof. Simply note that if $\mathrm{GO}_{\Sigma,e_{|H}}(E) \neq u$, then there must exist a set $H'' \subseteq H$, such that $e_{|H}(H'') \supseteq E$. But then, from the monotonically increasing nature of e, it holds that $e_{|H}(H) \supseteq E$. $\Diamond$

Hence, most general superset diagnosis has much in common with weak-causality diagnosis WC discussed in the previous chapter. Where weak-causality diagnosis is only defined for monotonically increasing evidence functions, most general superset diagnosis is defined for any proper evidence function. If the notion of most general superset diagnosis is applied to evidence functions that are monotonically decreasing, or nonmonotonic, for the resulting diagnosis $\mathrm{GO}_{\Sigma,e_{|H}}(E) = H'$ it may even hold that $e(H') \subset E$, although for each of the diagnostic hypotheses $H'' \subseteq H$ that contribute to the diagnosis it holds that $e_{|H}(H'') \supseteq E$. Hence, the situation is the reverse of that for most general subset diagnosis discussed above, as might be expected from their respective definitions.

In Figure 5.2, the various possibilities are schematically depicted.

The notion of weak-causality diagnosis may be generalized, by defining it for each possible evidence function. If this generalized notion of weak-causality diagnosis is denoted by WC^g, then the following proposition holds.

Proposition 5.6. *Let WC^g and GO be the notions of generalized weak-causality diagnosis and most general superset diagnosis, respectively, then*

$$\mathrm{WC}^g \sqsubseteq \mathrm{GO}$$

Proof. Simply note that if $\mathrm{WC}^g_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) \supseteq E$; hence, by Lemma 5.3, $\mathrm{GO}_{\Sigma,e_{|H}}(E) = H$. $\Diamond$

If $\mathrm{WC}^g_{\Sigma,e_{|H}}(E)$ is undefined, then $\mathrm{GO}_{\Sigma,e_{|H}}(E)$ may be any proper subset of H or undefined.

As is true for weak-causality diagnosis WC, most general superset diagnosis restricted to monotonically increasing evidence functions is very unrestrictive, which is revealed by the fact that $\mathrm{GO}_{\Sigma,e_{|H}}(\varnothing) = H$ if $e(H) \neq \bot$, meaning that all defects constituting the hypothesis may occur if no findings have been observed. Note that the same diagnosis would have been produced by weak-causality diagnosis WC in this case. By adopting some criterion of parsimony, such as minimality according to set inclusion, the unrestrictiveness is alleviated; the empty diagnosis $\varnothing$ would then be produced. This is the approach adopted in the diagnostic theories reviewed in Chapter 2. Our framework also includes the possibility of selecting diagnoses in similar fashion (cf. Section 3.3).

The notion of most general superset diagnosis produces more interesting results if it is restricted to evidence functions that are monotonically decreasing. Then, the independence assumption is satisfied for GO. This is not difficult to see, because if $e_{|H}(D) \supseteq E$, then $e_{|\{d\}}(d) \supseteq E$, for each $d \in D$. Monotonically decreasing evidence functions have been investigated in Chapter 3 for describing the normal and abnormal behaviour of devices.

Example 5.5. Reconsider Example 4.9, where defects of an electronic circuit consisting of two inverters, referred to as N_1 and N_2, respectively, is discussed. Some relevant values for the evidence function e were as follows:

$$\begin{aligned} e(\{i,n_1,n_2\}) &= \{o\} \\ e(\{i,n_1,\neg n_2\}) &= \{\neg o\} \\ e(\{i,n_1\}) &= \{o,\neg o\} \end{aligned}$$

Now, assume that the observed findings are only output signals. For the set of observed findings equal to $E = \{o\}$, the most general superset diagnosis is equal to

$$\mathrm{GO}_{\Sigma,e_{|\{i,n_1,n_2\}}}(E) = \{i,n_1,n_2\}$$

which corresponds to the consistency-based diagnosis. However,

$$\mathrm{GO}_{\Sigma,e_{|\{i,n_1\}}}(E) = \{i,n_1\}$$

where, as we have seen in Example 5.2, the consistency-based diagnosis is undefined, because o in the set of observed findings E is inconsistent with $\{o,\neg o\}$. Hence, GO is not a suitable notion of diagnosis for partial diagnosis in the sense of [De Kleer et al., 1992]. The most general superset diagnosis $\{i,n_1\}$ expresses the very weak information that there may exist a superset of $\{i,n_1\}$ that accounts for $\{o\}$ (Here, we have that the superset is equal to $\{i,n_1,n_2\}$.) $\diamondsuit$

Most general superset diagnosis that is restricted to monotonically decreasing evidence functions, can also be defined using the independent form

$$\mathrm{GO}^i_\Sigma = \{(E,(d,F)) \mid E \subseteq \Phi, e(d) = F, F \supseteq E\}$$

Therefore,

$$\mathrm{GO}_{\Sigma,e_{|H}}(E) = \{d \in H \mid e_{|H}(d) \supseteq E\}$$

if $\mathrm{GO}_{\Sigma,e_{|H}}(E) \neq u$. When the set of defects in a diagnostic specification is externally described (cf. Definition 3.17), the independence assumption is satisfied as well. Finally, note that the notion of most general superset diagnosis is Δ-complete – any consistent set of defects can be diagnosed given the empty set of observed findings –, and Φ-incomplete, because not every observed finding need be accounted for.

5.2.3 Most general intersection diagnosis

As shown above, in the formalization of diagnostic problem solving by means of the notion of most general subset diagnosis, there may be findings in the set of observed findings that are not accounted for by any set of defects in a diagnosis. More precisely, there may be a finding $f \in E$ for which we have a set of defects $D \subseteq \Delta$ such that $f \in e(D)$, but $e(D) \not\subseteq E$. Such defects may not be included in the subset diagnosis (they may be, though, by some other set D' for which $e(D') \subseteq E$). An alternative to the definition of subset diagnosis is to consider all sets of defects D that have at least one finding f in common with the findings E observed. This leads to the following definition.

Definition 5.3 (*most general intersection diagnosis*)**.** *The notion of* most general intersection diagnosis*, denoted by* GI*, is defined as follows:*

$$\mathrm{GI}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ (E = \varnothing \vee e_{|H}(H') = \varnothing \vee \\ e_{|H}(H') \cap E \neq \varnothing)}} H' & \begin{array}{l} \text{if } H \text{ is consistent, and } (E = \varnothing \text{ or} \\ \quad \exists H' \subseteq H : e_{|H}(H') = \varnothing \text{ or} \\ \quad e_{|H}(H') \cap E \neq \varnothing) \end{array} \\ u & \text{otherwise} \end{cases}$$

for each $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$. The diagnostic solution $\mathrm{GI}_{\Sigma,e_{|H}}(E)$ of a diagnostic problem $\mathcal{P} = (\Sigma, E)$ is called the most general intersection diagnosis of $\mathcal{P}$ with respect to H*.*

The most general intersection diagnoses with respect to Δ_P and Δ_N are called the positive and negative most general intersection diagnosis, respectively. If the sets of observed and observable findings are nonempty, intersection diagnosis with respect to H stands for the least upper bound of subsets of defects of $H \subseteq \Delta$, where for each subset of defects H' admitted to the most general intersection diagnosis $\mathrm{GI}_{\Sigma,e_{|H}}(E)$, the associated set of observable findings $e_{|H}(H')$ is empty or has at least one finding in common with the set of observed findings E.

The independence assumption is not satisfied for most general intersection diagnosis, which is even true if GI is restricted to interaction-free evidence functions. The reason is that if $e_{|H}(H') \cap E \neq \varnothing$, then it need not be true that for all $d \in H'$: $e_{|\{d\}}(d) \cap E \neq \varnothing$. Only if $e(D) = e(D')$, for each consistent $D, D' \subseteq \Delta$ (every set of defects has the same set of associated findings) would the independence assumption hold. However, if e is interaction free, the notion of most general intersection diagnosis restricted to such interaction-free evidence functions is Δ-monotonic.

The advantage of most general intersection diagnosis over most general subset and superset diagnosis is that all defects that have at least one associated observable finding that has actually been observed, are included in the diagnosis. This will be an acceptable assumption in a domain where not all findings associated with a set of defects need be observed and not all observed findings need be accounted for. This notion of diagnosis is less restrictive than most general subset and most general superset diagnosis, but if considerable overlap exists between the findings associated with the defects, the diagnostic solution $\mathrm{GI}_{\Sigma,e_{|H}}(E)$ may consist of many elements from $H \subseteq \Delta$. Hence, in representing a domain, it may be required to restrict to those observable findings that are in some way 'typical' for the defects.

$$\begin{tikzcd}
& \mathrm{GS} & \\
\mathrm{SC}^g \arrow[ur, phantom, "\sqsubseteq"] \arrow[dr, phantom, "\sqsubseteq"] & & \mathrm{GO} \\
& \mathrm{WC}^g \arrow[ur, phantom, "\sqsubseteq"] \arrow[dr, phantom, "\sqsubseteq"] & \\
& & \mathrm{GI}
\end{tikzcd}$$

Figure 5.3: Restriction taxonomy of notions of diagnosis.

For trivial reasons, most general intersection diagnosis GI is Δ-complete (just take $E = \varnothing$). Furthermore, it is Φ-incomplete, because not every finding in a given set of observed findings needs to be accounted for. It seems as if most general intersection diagnosis defined for exhaustive diagnostic specifications, can account for any set of observed findings, given some hypothesis H. However, if $e(d) = \{f_1\}$ and $e(\neg d) = \{f_2\}$, then it is never possible to account for f_1 and f_2 at the same time (if no other values of the evidence function include these findings).

Most general intersection diagnosis can be viewed as a refinement version of a mixture of the notions of weak-causality and strong-causality diagnosis. If again WC^g denotes generalized weak-causality diagnosis, the following proposition follows immediately.

Proposition 5.7. *Let* WC^g *be the notion of generalized weak-causality diagnosis, and let* GI *be the notion of most general intersection diagnosis, then*

$$\mathrm{WC}^g \sqsubseteq \mathrm{GI}$$

Proof. Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. If $\mathrm{WC}^g_{\Sigma,e_{|H}}(E) = H$, then $e_{|H}(H) \supseteq E$. Assume that $E \neq \varnothing$, then $e_{|H}(H) \cap E \neq \varnothing$, hence $\mathrm{WC}^g_{\Sigma,e_{|H}}(E) = \mathrm{GI}_{\Sigma,e_{|H}}(E)$. Otherwise, if $E = \varnothing$, then $\mathrm{WC}^g_{\Sigma,e_{|H}}(\varnothing) = H$ and $\mathrm{WC}^g_{\Sigma,e_{|H}}(E) = \mathrm{GI}_{\Sigma,e_{|H}}(E)$. From this the proposition follows. $\Diamond$

Although the notion of most general intersection diagnosis seems less restrictive than the notions of subset and superset diagnosis, the latter two are not restrictions of most general intersection diagnosis.

Interestingly, when the set of defects Δ of a diagnostic specification is interaction free, most general intersection diagnosis is similar to relevance diagnosis in the set-covering theory of diagnosis (cf. Section 2.2.3) [Peng & Reggia, 1990]. The main difference is that a relevant diagnosis always accounts for all observed findings, where a most general intersection diagnosis need not account for all observed findings.

5.2.4 Comparison

The most general subset, superset and intersection diagnosis are three refinement approaches to diagnosis. The restriction relationships between these notions of diagnosis are reviewed in Figure 5.3. For most general subset diagnosis, all findings associated with a set of defects must be observed if the set of defects is to be included as part of the diagnosis. Most general superset diagnosis focusses on common findings of defects. For most general intersection diagnosis, at least one finding associated with a defect must be observed if the defect is to be included as part of the diagnosis. The three notions of diagnosis discussed above stand in a subdiagnostic relation to each other:

$$\begin{aligned}&\mathrm{GS} \trianglelefteq \mathrm{GI}\\&\mathrm{GO} \trianglelefteq \mathrm{GI}\end{aligned}$$

This follows from the fact that if a set of observed findings is included in the set of observable findings associated with a set of defects, or vice versa, the intersection of the set of observed findings and observable findings is nonempty, given that neither E nor $e_{|H}(H')$ is empty. For the empty cases, the most general intersection diagnosis is always equal to the largest result with respect to set inclusion of GO and GS. Hence, a most general intersection diagnosis will always contain at least as many elements as most general superset and subset diagnosis.

5.3 Most specific diagnosis

Rather than taking the least upper bound of a set of accepted subhypotheses of a given hypothesis, taking the greatest lower bound provides another approach to refinement diagnosis. We shall refer to notions of diagnosis based on taking the greatest lower bound as notions of *most specific diagnosis*. Where the concept of most general diagnosis formalizes notions of diagnosis that yield diagnoses that include every accepted subhypothesis, most specific diagnosis formalizes notions of diagnosis that yield diagnoses that are common to every accepted subhypothesis. In most general diagnosis, the smallest set of defects that includes every accepted subhypothesis, is considered most plausible; in contrast, in most specific diagnosis, the largest set of defects that is included in every accepted subhypothesis, is considered most plausible. Furthermore, in general it holds for a notion of most specific diagnosis S that if $S_{\Sigma,e_{|H}}(E) = \varnothing$ and $S_{\Sigma,e_{|H'}}(E) = H''$, then, by definition, $S_{\Sigma,e_{|H\cup H'}}(E) = \varnothing$. In medical diagnosis, it is often assumed that the occurrence of small sets of positive defects is more likely than the occurrence of large sets of positive defects. As for most general diagnosis, we focus on notions of diagnosis where every finding associated with a subhypothesis must have been observed, or every observed finding must have been accounted for by a subhypothesis, or a mixture of the two. This brings us to define notions of most specific subset, most specific superset and most specific intersection diagnosis.

5.3.1 Most specific subset diagnosis

As with the notion of most general subset diagnosis, in the notion of most specific subset diagnosis, subhypotheses are admitted to a diagnosis if their associated sets of findings are included in the set of observed findings of a diagnostic problem. However, of these accepted subhypotheses, only the defects the subhypotheses have in common constitute a diagnosis.

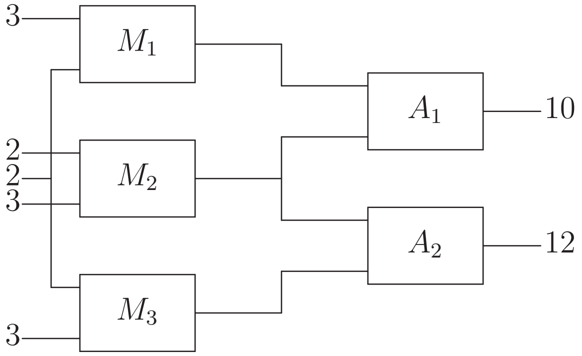


Figure 5.4: Multiplier-adder circuit.

Definition 5.4 (*most specific subset diagnosis*). *The notion of* most specific subset diagnosis, *denoted by* SS, *is defined as follows:*

$$\mathrm{SS}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcap_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \subseteq E \\ u & \text{otherwise} \end{cases}$$

for each $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$. The diagnostic solution $\mathrm{SS}_{\Sigma,e_{|H}}(E)$ of a diagnostic problem $\mathcal{P} = (\Sigma, E)$ is called the most specific subset diagnosis of $\mathcal{P}$ with respect to H.

This notion of diagnosis is extremely restrictive. For example, if an evidence function is interaction free, then the most specific subset diagnosis will be always equal to the empty set.

Proposition 5.8. *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. If $e_{|H}(\varnothing) = \varnothing$, then $\mathrm{SS}_{\Sigma,e_{|H}}(E) = \varnothing$, for each $H \subseteq \Delta$.*

Proof. Immediate from the definition. ◊

Hence, if few interactions exist among defects in a domain, most specific subset diagnosis will often yield an empty diagnosis.

If the evidence function is monotonically decreasing, then most specific subset diagnosis tries to construct the smallest diagnosis possible. Due to the fact that subset diagnosis has much in common with consistency-based diagnosis, most specific subset diagnosis applied to a monotonically decreasing evidence function representing a device, is a special form of kernel diagnosis in the sense of [De Kleer et al., 1992]. Recall that a kernel diagnosis is a partial diagnosis that is the smallest partial diagnosis according to set inclusion (cf. Section 4.4).

Example 5.6. The correspondence between kernel diagnosis and most specific subset diagnosis is illustrated by an example taken from [De Kleer et al., 1992]. Consider Figure 5.4, which depicts an electronic circuit with three multipliers, referred to as M_1, M_2 and M_3, and two adders, denoted by A_1 and A_2. Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification representing the circuit. As before, the fact that some multiplier M_i is defective, is denoted by m_i; if it is nondefective, this is indicated by $\neg m_i$. A similar notational convention is adopted with regard to the two adders. It is convenient to assume that the input to the circuit is fixed (as assumed in [Davis & Hamscher, 1988] and [De Kleer et al., 1992]), as indicated in Figure 5.4. The normal output of the circuit, $O_1 = 12$ and $O_2 = 12$, is denoted by o_1 and o_2; abnormal output is denoted by $\neg o_j$, $j = 1, 2$.

The following values of the evidence function are among those that correspond to the circuit:

$$\begin{array}{ll}
e(\{\neg m_1, \neg m_2, \neg m_3, \neg a_1, \neg a_2\}) & = \{o_1, o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1, \neg a_2\}) & = \{o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1, a_2\}) & = \varnothing \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1\}) & = \{o_2\} \\
e(\{a_1\}) & = \{o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3\}) & = \{o_1, o_2\} \\
 & \vdots \\
e(\varnothing) & = \{o_1, o_2\}
\end{array}$$

The most specific subset diagnosis with respect to the hypothesis $H = \{a_1\}$ is equal to

$$\mathrm{SS}_{\Sigma, e_{|\{a_1\}}}(\{\neg o_1, o_2\}) = \{a_1\}$$

which is indeed a kernel diagnosis for the diagnostic problem $\mathcal{P} = (\Sigma, E)$ using consistency-based diagnosis. Note that

$$\mathrm{SS}_{\Sigma, e_{|H}}(\{\neg o_1, o_2\}) = \{a_1\}$$

if $a_1 \in H$, for example, $H = \{\neg m_1, \neg m_2, \neg m_3, a_1, \neg a_2\}$. ◊

The main reason for the similarity between kernel diagnosis in consistency-based diagnosis and most specific subset diagnosis is that any hypothesis H' for which $e_{|H}(H') \subseteq E$ is also consistent with E.

Proposition 5.9. *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. If $\mathrm{SS}_{\Sigma, e_{|H}}(E) = H$, $H \neq u$, is a diagnosis for some diagnostic problem with monotonically decreasing evidence function e, then $\mathrm{SS}_{\Sigma, e_{|H'}}(E) = \mathrm{SS}_{\Sigma, e_{|H}}(E)$ for each $H' \supseteq H$.*

Proof. If $\mathrm{SS}_{\Sigma, e_{|H}}(E) = H$, then $e_{|H}(H) \subseteq E$. Since e is monotonically decreasing, it follows that $e_{|H'}(H') \subseteq E$ for each $H' \supseteq H$. ◊

From this proposition, it follows that the notion of diagnosis SS may be viewed as a flexible form of kernel diagnosis, defined as a notion of diagnosis.

Finally, note that most specific subset diagnosis is neither Φ-complete, because not every observed finding need be accounted for, nor Δ-complete, because an evidence function e may have function values like $e(D) = \{f, \neg f\}$ that cannot be used for diagnosing observed findings.

5.3.2 Most specific superset diagnosis

As discussed in Section 5.2.2, most general superset diagnosis will often yield a diagnosis that contains too many defect elements, in particular when an evidence function is monotonically increasing. Most specific superset diagnosis is a more restrictive, and possibly

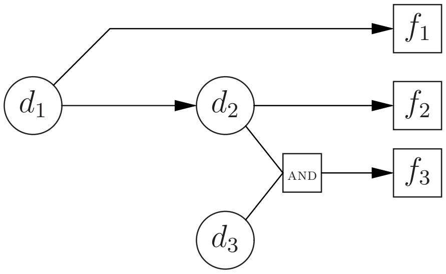


Figure 5.5: Causal net (repeated).

more suitable, notion of diagnosis, than most general superset diagnosis.

Definition 5.5 (*most specific superset diagnosis*). *The notion of* most specific superset diagnosis, *denoted by* SO, *is defined as follows:*

$$\mathrm{SO}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcap_{\substack{H' \subseteq H \\ e_{|H}(H') \supseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \supseteq E \\ u & \text{otherwise} \end{cases}$$

for each $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$. The diagnostic solution $\mathrm{SC}_{\Sigma,e_{|H}}(E)$ *of a diagnostic problem $\mathcal{P} = (\Sigma, E)$ is called the* most specific superset diagnosis of $\mathcal{P}$ with respect to H.

If the evidence function to which most specific superset diagnosis is applied, is monotonically increasing, the result may be intuitively attractive. The basic idea of most specific superset diagnosis is that the observed findings that are common to the accepted subhypotheses are due to common defects of the accepted subhypotheses. Hence, an evidence function is modified along those lines in the process of diagnosis.

Example 5.7. Example 3.1 from Chapter 3 considered a diagnostic specification Σ with the following bottom-up partial specification of an evidence function e:

$$\tilde{e}(D) = \begin{cases} \{f_1, f_2\} & \text{if } D = \{d_1\} \\ \{f_2\} & \text{if } D = \{d_2\} \\ \{f_2, f_3\} & \text{if } D = \{d_2, d_3\} \\ \{f_1, f_2, f_3\} & \text{if } D = \{d_1, d_3\} \\ \bot & \text{if } \{d_1, \neg d_2\} \subseteq D \\ \varnothing & \text{if } D = \{\neg d_i\},\ i = 1,2,3, \text{ or } D = \{d_3\} \end{cases}$$

Recall that this evidence function expresses medical knowledge concerning influenza and related disorders. The figure from the original example is repeated in Figure 5.5. For $E = \{f_2, f_3\}$ (i.e. the patient has a sore throat and dyspnoea), the most specific superset diagnosis is equal to

$$\mathrm{SO}_{\Sigma,e_{|\{d_1,d_2,d_3\}}}(E) = \{d_3\}$$

because, it holds that $e_{|H}(\{d_1, d_3\}) \supseteq E$, $e_{|H}(\{d_2, d_3\}) \supseteq E$ and $e_{|H}(\{d_1, d_2, d_3\}) \supseteq E$, where $H = \{d_1, d_2, d_3\}$. All other subsets of H have associated sets of findings that are no supersets of E. The defect d_3 stands for asthma. It is interesting to note that both d_1 and d_2 participate in subhypotheses that also account for E. However, only the defect d_3 occurs in all accepted subhypotheses, i.e. turns out to be essential. It seems therefore intuitively right to accept d_3 as the most plausible diagnosis. Thus, instead of taking $e(d_3) = \varnothing$ as the proper value, it is assumed that for this particular diagnostic problem the function value $e(d_3) = \{f_2, f_3\}$ holds. In other words, the statement that no findings will be observed if d_2 is present, and the presence or absence of all other defects is unknown, is considered too strong. $\Diamond$

As the example above indicates, the resulting most specific superset diagnosis need not account for all observed findings on the basis of the given evidence function. If an evidence function is interaction free, then most specific superset diagnosis is likely to produce a singleton set diagnosis for a given hypothesis that is very plausible if the associated sets of observed findings $e(d)$ are mutually disjoint.

If the evidence function is monotonically decreasing, it is not easy to come up with an intuitively satisfactory interpretation of most specific superset diagnosis. Similarly, if the evidence function can be represented as a top-down partial specification, the resulting diagnosis will be the empty set. This situation is similar to most general superset diagnosis for monotonically decreasing evidence functions. Finally, the notion of most specific superset diagnosis is neither Φ-complete nor Δ-complete.

5.3.3 Most specific intersection diagnosis

As discussed in Section 5.2.3, most general intersection diagnosis is a very unrestrictive notion of diagnosis. All defects that, either individually or in combination with other defects, have findings in common with the set of observed findings, are included in a diagnosis. The notion of most specific intersection diagnosis is much more restrictive than most general diagnosis.

Definition 5.6 (*most specific intersection diagnosis*). *The notion of* most specific intersection diagnosis*, denoted by* SI*, is defined as follows:*

$$\mathrm{SI}_{\Sigma,e_{|H}}(E) = \begin{cases} \displaystyle\bigcap_{\substack{H' \subseteq H \\ (E = \varnothing \vee e_{|H}(H') = \varnothing \vee \\ e_{|H}(H') \cap E \neq \varnothing)}} H' & \begin{array}{l}\text{if } H \text{ is consistent, and } (E = \varnothing \text{ or} \\ \quad \exists H' \subseteq H : e_{|H}(H') = \varnothing \text{ or} \\ \quad e_{|H}(H') \cap E \neq \varnothing)\end{array} \\ u & \text{otherwise} \end{cases}$$

for each $\Sigma \in \mathcal{S}$, each set of observed findings $E \subseteq \Phi$, and each $H \subseteq \Delta$. The diagnostic solution $\mathrm{SI}_{\Sigma,e_{|H}}(E)$ of a diagnostic problem $\mathcal{P} = (\Sigma, E)$ is called the most specific intersection diagnosis with respect to *H.*

If the evidence function for which most specific intersection diagnosis is defined is monotonically increasing, the resulting diagnosis will often be equal to the empty set if the function values $e(d)$, $d \in \Delta$, have many observable findings in common. It will always be empty if Δ is interaction free. On the other hand, if the function values $e(d)$, $e(d')$, $d, d' \in \Delta$, $d \neq d'$, are disjoint, then the diagnosis will always be a singleton set or empty if the diagnostic specification Σ in the function $\mathrm{SI}_{\Sigma,e_{|H}}$ is exhaustive, and always empty if $e(\varnothing) = \varnothing$ or $e(\varnothing) \cap E \neq \varnothing$.

Because evidence functions that are monotonically decreasing will have many findings in common, in particular the intersection of $e(\varnothing)$ with the set of observed findings E will always be nonempty for exhaustive diagnostic specifications, most specific intersection diagnosis is likely to be equal to the empty set. Most specific intersection diagnosis is Δ-incomplete, because if the function values $e(D)$, $e(D')$, $D, D' \subseteq \Delta$, $D \neq D'$, are disjoint it is only possible to have the empty set or u as a diagnosis; furthermore, SI is Φ-incomplete.

5.3.4 Comparison

Although the notions of most specific diagnosis are very restrictive, they do not stand in a simple restriction relation to the other notions of diagnosis. It is easy to see that

$$\mathrm{SS}_{\Sigma,e_{|H}}(E) \subseteq \mathrm{GS}_{\Sigma,e_{|H}}(E)$$

holds for each consistent $H \subseteq \Delta$. Similar set inclusion relations hold for the other notions of diagnosis. We state without proof that:

$$\mathrm{SS} \trianglelefteq \mathrm{GS}$$
$$\mathrm{SO} \trianglelefteq \mathrm{GO}$$
$$\mathrm{SI} \trianglelefteq \mathrm{GI}$$

5.4 Similarity of components

In this section, some special properties of the three notions of diagnosis – most general subset, superset and intersection diagnosis – introduced above, are investigated. Similar properties can be studied for the other notions of diagnosis.

5.4.1 Basic assumptions

As discussed in Section 3.2.3, the diagnostic components of a notion of diagnosis for a given diagnostic specification can be grouped into equivalence classes on the basis of the sets of findings that can be accounted for. An equivalence class of diagnostic components was denoted by $[R_{\Sigma,e_{|H}}]_{\equiv}$. Diagnostic components in an equivalence class were considered partially ordered by the $\trianglelefteq$ relation. Then, interest in considering unique $\trianglelefteq$-minimal diagnostic component naturally arises, because of the potential computational advantage of limiting the computation to such components when determining minimal diagnoses.

In the study of $\trianglelefteq$ minimality of diagnostic components it will be assumed that the different notions of diagnosis are restricted to diagnostic specifications with evidence functions that are interaction free, or that they are externally described. For interaction-free or externally described evidence functions, changes with respect to the set of observable findings are local with respect to individual defects, dependent on the notion of diagnosis employed.

5.4.2 Interaction-free defects

First, a number of properties of most general subset diagnosis will be studied. Throughout this section, it is assumed that the evidence function e of a given diagnostic specification Σ is interaction free, i.e. the domains of most general subset, superset and intersection diagnosis are assumed to be restricted to such evidence functions. Furthermore, it is sometimes also assumed that the evidence function is synonym free as well, i.e. for each $d, d' \in \Delta$ it holds that if $d \neq d'$ then $e(d) \neq e(d')$.

A basic property of most general subset diagnosis is that the diagnostic solutions of minimal diagnostic components of an equivalence class contain only essentially accounting defects, i.e. defects that cannot be removed from a diagnostic solution without losing the property that each diagnosis accounts for the same observed findings as the other members of the equivalence class.

Proposition 5.10. *Let* GS *be the notion of most general subset diagnosis, and let* $\mathcal{P} = (\Sigma, E)$ *be a diagnostic problem. Let* $\mathrm{GS}_{\Sigma,e_{|H}}$ *be a minimal diagnostic component of an equivalence class* $[\mathrm{GS}_{\Sigma,e_{|H}}]_{\equiv}$ *according to the* $\trianglelefteq$ *relation. If* $d \in \mathrm{GS}_{\Sigma,e_{|H}}(E)$, $H \subseteq \Delta$, *for some* $E \subseteq \Phi$, *then* d *is an essentially accounting defect.*

Proof. Suppose, in contrast, that d is not an essentially accounting defect, i.e. there exists a set $D \subseteq \mathrm{GS}_{\Sigma,e_{|H}}(E)$, with $e(D) \subseteq E$, such that $A(D\backslash\{d\}, E) = A(\{d\}, E)$, or, equivalently, $e(D\backslash\{d\}) = e(d)$. Let $H' = H\backslash\{d\}$. If $d \in \mathrm{GS}_{\Sigma,e_{|H}}(E)$, then

$$A(\mathrm{GS}_{\Sigma,e_{|H}}(E)\backslash\{d\}, E) = A(\mathrm{GS}_{\Sigma,e_{|H'}}(E), E)$$

Furthermore, for each $d' \in \mathrm{GS}_{\Sigma,e_{|H}}(E)$, $d' \neq d$: $d' \in \mathrm{GS}_{\Sigma,e_{|H'}}(E)$, due to the independence assumption, which holds because e is interaction free (cf. Proposition 4.14). In addition $A(D\backslash\{d\}, e(d)) = A(\mathrm{GS}_{\Sigma,e_{|H}}(e(d)), e(d))$, therefore, it holds that for each $E \subseteq \Phi$

$$A(\mathrm{GS}_{\Sigma,e_{|H}}(E), E) = A(\mathrm{GS}_{\Sigma,e_{|H'}}(E), E)$$

It can be concluded that the diagnostic component $\mathrm{GS}_{\Sigma,e_{|H'}}$ is similar to $\mathrm{GS}_{\Sigma,e_{|H}}$, but smaller, yielding a contradiction. This proves the proposition. $\Diamond$

In the following example the proposition is illustrated.

Example 5.8. Consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Phi_P = \{f_1, f_2, f_3\}$ and $\Delta_P = \{d_1, d_2, d_3\}$, with bottom-up partial specification $\tilde{e}$:

$$\tilde{e}(d) = \begin{cases} \{f_1\} & \text{if } d = d_1 \\ \{f_2, f_3\} & \text{if } d = d_2 \\ \{f_1, f_2, f_3\} & \text{if } d = d_3 \\ \bot & \text{if } d \in \Delta_N \end{cases}$$

of the interaction-free evidence function e. Consider the diagnostic components with respect to the sets $H = \{d_1, d_2\}$:

$$\begin{aligned} \mathrm{GS}_{\Sigma,e_{|H}} = \{&(\{f_1\}, \{d_1\}), \\ &(\{f_2, f_3\}, \{d_2\}), \\ &(\{f_1, f_2, f_3\}, \{d_1, d_2\})\} \end{aligned}$$

and $H' = \{d_1, d_2, d_3\}$:

$$\begin{aligned}\mathrm{GS}_{\Sigma,e_{|H'}} = \{&(\{f_1\}, \{d_1\}),\\ &(\{f_2, f_3\}, \{d_2\}),\\ &(\{f_1, f_2, f_3\}, \{d_1, d_2, d_3\})\}\end{aligned}$$

The diagnostic component $\mathrm{GS}_{\Sigma,e_{|H}}$ is $\trianglelefteq$-minimal in this case. Note that the defects that are essential to account for the sets of findings $\{f_1\}$ and $\{f_2, f_3\}$ occur in the diagnoses obtained from both diagnostic components for these sets of observed findings. ◊

A similar proposition does not hold for most general intersection diagnosis, as the following counter-example demonstrates.

Example 5.9. For the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Delta_P = \{d_1, d_2, d_3\}$, $\Phi_P = \{f_1, f_2, f_3, f_4, f_5\}$, and e is given by the following bottom-up partial specification:

$$\tilde{e}(d) = \begin{cases} \{f_1, f_5\} & \text{if } d = d_1 \\ \{f_2, f_4\} & \text{if } d = d_2 \\ \{f_1, f_2, f_3\} & \text{if } d = d_3 \\ \varnothing & \text{if } d \in \Delta_N \end{cases}$$

the diagnostic component $\mathrm{GI}_{\Sigma,e_{|\Delta_P}}$ is minimal with respect to $\trianglelefteq$, because only if all defects are included in the diagnostic component can all possible observed findings be accounted for. However, $A(\{d_1, d_2\}, \{f_1, f_2\}) = A(\{d_3\}, \{f_1, f_2\})$, i.e. d_3 is not essential to account for $\{f_1, f_2\}$. ◊

The following proposition indicates a structure of diagnostic components that is typical for most general subset diagnosis.

Proposition 5.11. *Let* GS *be the notion of most general subset diagnosis. Let* $\mathcal{P} = (\Sigma, E)$ *be a diagnostic problem with interaction-free set of defects* Δ. *Let* $\mathrm{GS}_{\Sigma,e_{|H}}$, $\mathrm{GS}_{\Sigma,e_{|H'}} \in [\mathrm{GS}_{\Sigma,e_{|H}}]_{\equiv}$ *be diagnostic components, then if* $d \in \mathrm{GS}_{\Sigma,e_{|H}}(E)$ *is a defect such that* d *is an essentially accounting defect, then there exists a defect* $d'' \in \mathrm{GS}_{\Sigma,e_{|H'}}(E)$ *with* $e(d) = e(d'')$.

Proof. By the definition of component similarity, it holds that for each $d \in \mathrm{GS}_{\Sigma,e_{|H}}(E)$ there exists a set $D \subseteq \mathrm{GS}_{\Sigma,e_{|H'}}(E)$, such that $e(d) = A(D, e(d)) = e(D)$. Choose the smallest of such sets D that is nonempty. If $|D| > 1$, then $e(d') \subset e(d)$, for each $d' \in D$. Furthermore, for the elements d' of the set D there must exist a set $D_{d'} \subseteq \mathrm{GS}_{\Sigma,e_{|H}}(E)$, such that $e(d') = e(D_{d'})$. The last equality follows from the fact that $e(d)$ is also a set of observed findings that can be accounted for by both $\mathrm{GS}_{\Sigma,e_{|H}}$ and $\mathrm{GS}_{\Sigma,e_{|H'}}$. Since $e(d') \subset e(d)$, it holds that $d \notin D_{d'}$. Component similarity ensures that $e(\bigcup_{d' \in D} D_{d'}) = e(d)$, thus contradicting the fact that d is essential to accounts for $e(d)$. We conclude that $|D| = 1$; hence, $d'' \in \mathrm{GS}_{\Sigma,e_{|H'}}(E)$, $D = \{d''\}$, and $e(d) = e(d'')$. ◊

If, in addition, the set of defects is assumed to be synonym free, then an essentially accounting defect d occurs in $\mathrm{GS}_{\Sigma,e_{|H'}}(E)$ as well, i.e. the proposition above then states that the defects are essential to account for observed findings will occur in every diagnostic component of an equivalence class of most general subset diagnosis. In contrast with most general subset diagnosis, a similar property does not hold in general for most general intersection diagnosis.

Example 5.10. Consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Delta_P = \{d_1, d_2, d_3, d_4, d_5\}$, $\Phi_P = \{f_1, f_2, f_3, f_4\}$, and bottom-up partial specification $\tilde{e}$:

$$\tilde{e}(d) = \begin{cases} \{f_1\} & \text{if } d = d_1 \\ \{f_2, f_3\} & \text{if } d = d_2 \\ \{f_2, f_4\} & \text{if } d = d_3 \\ \{f_4\} & \text{if } d = d_4 \\ \{f_3\} & \text{if } d = d_5 \\ \varnothing & \text{if } d \in \Delta_N \end{cases}$$

If $H = \{d_1, d_2, d_4\}$ and $H' = \{d_1, d_3, d_5\}$, then $\mathrm{GI}_{\Sigma, e_{|H}}$ and $\mathrm{GI}_{\Sigma, e_{|H'}}$ will be both part of the same equivalence class. However, for the set of observed findings $E = \{f_1, f_2\}$, it holds that $\mathrm{GI}_{\Sigma, e_{|H}}(E) = \{d_1, d_2\}$, $A(\{d_1\}, E) \neq A(\{d_2\}, E)$, and also $A(\varnothing, E) \neq A(\{d_2\}, E)$, but $d_2 \notin \mathrm{GI}_{\Sigma, e_{|H}}(E) = \{d_1, d_3\}$. $\Diamond$

The proposition above implies that, in contrast to most general intersection diagnosis, any equivalence class in the quotient set $P_{\mathrm{GS}_\Sigma}/{\equiv}$ contains a unique least element, if it is assumed that the evidence functions are interaction and synonym free.

Corollary. *Let* GS *be the notion of most general subset diagnosis, and let* $\equiv$ *be the component similarity relation. Let* Σ *be a diagnostic specification with interaction and synonym-free evidence function. Then, every equivalence class in the quotient set* $P_{\mathrm{GS}_\Sigma}/{\equiv}$ *contains a unique least element by set inclusion of diagnosis.*

Proof. Let $\mathrm{GS}_{\Sigma, e_{|H}}(E)$ and $\mathrm{GS}_{\Sigma, e_{|H'}}(E)$ be two most general subset diagnoses from diagnostic components belonging to the same equivalence class. Let $\mathrm{GS}_{\Sigma, e_{|H}}$ be a $\trianglelefteq$ minimal diagnostic component. Then, from Proposition 5.10, Proposition 5.11 and synonym freeness of e, it follows that for each $d \in \mathrm{GS}_{\Sigma, e_{|H}}(E)$: $d \in \mathrm{GS}_{\Sigma, e_{|H'}}(E)$. Hence, the minimal diagnostic component $\mathrm{GS}_{\Sigma, e_{|H}}$ is the unique, least element of the equivalence class. $\Diamond$

Although every equivalence class of diagnostic components for most general subset diagnosis contains a least element if the evidence function in the diagnostic specification is synonym free, this does not mean that every diagnosis obtained by application of this element always yield diagnoses that are minimal with respect to set inclusion. As the following example shows, there may be smaller diagnoses accounting for the same observed findings as the diagnosis obtained from the least diagnostic component of an equivalence class.

Example 5.11. Consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Delta_P = \{d_1, d_2\}$, $\Phi_P = \{f_1, f_2\}$, and the bottom-up partial specification

$$\tilde{e}(d) = \begin{cases} \{f_1\} & \text{if } d = d_1 \\ \{f_1, f_2\} & \text{if } d = d_2 \\ \varnothing & \text{if } d \in \Delta_N \end{cases}$$

yields the evidence function e. Let $H = \{d_2\}$ and $H' = \{d_1, d_2\}$, then the diagnostic components $\mathrm{GS}_{\Sigma,e_{|H}}$ and $\mathrm{GS}_{\Sigma,e_{|H'}}$ are not similar; both diagnostic components are least elements of their equivalence class. If $E = \{f_1, f_2\}$, then $\mathrm{GS}_{\Sigma,e_{|H}}(E) = \{d_2\}$, where $\mathrm{GS}_{\Sigma,e_{|H'}}(E) = \{d_1, d_2\}$. Note that $A(\mathrm{GS}_{\Sigma,e_{|H}}(E), E) = A(\mathrm{GS}_{\Sigma,e_{|H'}}(E), E)$. Hence, the smallest diagnosis with respect to set inclusion is not the result of application of the least diagnostic component $\mathrm{GS}_{\Sigma,e_{|H'}}$, but of the other diagnostic component $\mathrm{GS}_{\Sigma,e_{|H}}$. ♢

5.4.3 Externally described defects

The contribution of individual defects to a diagnosis in most general superset diagnosis is only apparent if the evidence function is assumed to be externally described. Thus, from now on, this will be assumed. Then, in contrast with most general subset and most general intersection, most general superset diagnosis has the property that each defect in a diagnostic solution accounts for all observed findings. An immediate consequence of this property is that if a defect in diagnosis is an essentially accounting defect, then the diagnosis must be a singleton set. A proposition, similar to Proposition 5.11 for most general subset diagnosis, holds for most general superset diagnosis.

Proposition 5.12. *Let* GO *be the notion of most general superset diagnosis, and* Σ *be a diagnostic specification with externally described evidence function* e*. Let* $\mathrm{GO}_{\Sigma,e_{|H}}, \mathrm{GO}_{\Sigma,e_{|H'}}$ *be two* $\equiv$ *similar diagnostic components, then if* $d \in \mathrm{GO}_{\Sigma,e_{|H}}(E)$ *is an essentially accounting defect, then there exists a defect* $d' \in \mathrm{GO}_{\Sigma,e_{|H'}}(E)$ *with* $e(d) = e(d')$*.*

Proof. If the defect d is an essentially accounting defect, then there exists a defect $d' \in \mathrm{GO}_{\Sigma,e_{|H'}}(E)$ such that $e(d') \supseteq e(d)$, because $A(\{d\}, e(d)) = A(\{d'\}, e(d))$; otherwise the two diagnostic components would not be similar. Suppose that $e(d') \supset e(d)$, then $A(\{d'\}, e(d')) = A(\{d''\}, e(d'))$, $d'' \in \mathrm{GO}_{\Sigma,e_{|H}}(E)$. But, since $e(d'') \supseteq e(d')$, we have $d'' \neq d$; we may conclude that d is not an essentially accounting defect, because $e(d'') \supseteq E$; contradiction. Therefore, $e(d') = e(d)$; it can be concluded that $d' \in \mathrm{GO}_{\Sigma,e_{|H'}}(E)$. ♢

For all diagnostic components of most general superset diagnosis included in the same equivalent class, it holds that the defect in a singleton diagnostic solutions is included in all other diagnoses in the equivalence class, if the evidence functions are assumed to be synonym free.

5.5 Selection of diagnoses

Rather than considering diagnostic components only, it is also possible to investigate the individual diagnoses that result when applying a diagnostic component $R_{\Sigma,e_{|H}}$ to a set of observed findings E of a diagnostic problem. The basic theory has been developed in Section 3.3. Multiple diagnosis was shown to be particularly important for selecting the diagnoses satisfying particular criteria of parsimony. Criteria of parsimony are not as important here as for the notions of weak, strong and consistency-based diagnosis discussed in Chapter 4. The notions of refinement diagnosis already incorporate a mechanism that selects, or constructs, a diagnosis that is considered most likely. Nevertheless, the theory can be applied, just as before.

5.5.1 Multiple diagnosis

Multiple most general diagnosis is defined as the equivalence class

$$[G_{\Sigma,e_{|H}}(E)]_{\doteq}$$

which is abbreviated to $\mathcal{D}_G(E, E')$, where $E' = A(G_{\Sigma,e_{|H}}(E), E)$. The notion of most general diagnosis G is either most general subset (GS), superset (GO) or intersection (GI) diagnosis. In this section, attention is focussed on the equivalence class $[G_{\Sigma,e_{|\Delta_P}}(E)]_{\doteq}$.

Example 5.12. Consider the diagnostic problem $\mathcal{P} = (\Sigma, E)$, with $\Delta_P = \{d_1, d_2, d_3\}$, $\Phi = \{f_1, f_2, f_3\}$, $E = \{f_1, f_3\}$ the bottom-up partial specification

$$\tilde{e}(d) = \begin{cases} \{f_1\} & \text{if } d = d_1 \\ \{f_1, f_2\} & \text{if } d = d_2 \\ \{f_1, f_3\} & \text{if } d = d_3 \\ \perp & \text{if } d \in \Delta_N \end{cases}$$

for the interaction-free evidence function e. The power set of Δ_P is equal to $\wp(\Delta_P) = \{H_1, \ldots, H_8\}$, where $H_1 = \varnothing$, $H_2 = \{d_1\}$, $H_3 = \{d_2\}$, $H_4 = \{d_3\}$, $H_5 = \{d_1, d_2\}$, $H_6 = \{d_1, d_3\}$, $H_7 = \{d_2, d_3\}$, $H_8 = \{d_1, d_2, d_3\}$. Then, the following most general subset diagnoses can be constructed: $\mathrm{GS}_{\Sigma,e_{|H_1}}(E) = \varnothing$, $\mathrm{GS}_{\Sigma,e_{|H_2}}(E) = \{d_1\}$, $\mathrm{GS}_{\Sigma,e_{|H_3}}(E) = \varnothing$, $\mathrm{GS}_{\Sigma,e_{|H_4}}(E) = \{d_3\}$, $\mathrm{GS}_{\Sigma,e_{|H_5}}(E) = \{d_1\}$, $\mathrm{GS}_{\Sigma,e_{|H_6}}(E) = \{d_1, d_3\}$, $\mathrm{GS}_{\Sigma,e_{|H_7}}(E) = \{d_3\}$ and $\mathrm{GS}_{\Sigma,e_{|H_8}}(E) = \{d_1, d_3\}$. Using the accountability relation $\doteq$, it is possible to partition the set of all diagnoses

$$P_{\mathrm{GS}_\Sigma}(E) = \{\mathrm{GS}_{\Sigma,e_{|H}}(E) \mid H \subseteq \Delta_P\}$$

into equivalence classes. Thus, the sets

$$\{\mathrm{GS}_{\Sigma,e_{|H_1}}(E), \mathrm{GS}_{\Sigma,e_{|H_3}}(E)\}$$

$$\{\mathrm{GS}_{\Sigma,e_{|H_2}}(E), \mathrm{GS}_{\Sigma,e_{|H_5}}(E)\}$$

and

$$\{\mathrm{GS}_{\Sigma,e_{|H_4}}(E), \mathrm{GS}_{\Sigma,e_{|H_6}}(E), \mathrm{GS}_{\Sigma,e_{|H_7}}(E), \mathrm{GS}_{\Sigma,e_{|H_8}}(E)\}$$

are equivalence classes. For the two elements in the first equivalence class, it holds that

$$A(\mathrm{GS}_{\Sigma,e_{|H_1}}(E), E) = A(\mathrm{GS}_{\Sigma,e_{|H_3}}(E), E) = \varnothing$$

and for the second equivalence class it holds:

$$A(\mathrm{GS}_{\Sigma,e_{|H_2}}(E), E) = A(\mathrm{GS}_{\Sigma,e_{|H_5}}(E), E) = \{f_1\}$$

The last equivalence class corresponds to the multiple most general subset diagnosis $\mathcal{D}_{\mathrm{GS}}(E,E)$ for $\mathcal{P}$. It may be observed that the most general subset diagnoses $\mathrm{GS}_{\Sigma,e_{|H_i}}(E)$, $i \in \{4,6,7,8\}$, are accountable, because

$$\begin{aligned} A(\mathrm{GS}_{\Sigma,e_{|H_8}}(E),E) &= A(\{d_1,d_3\},\{f_1,f_3\}) \\ &= \{f_1,f_3\} \\ &= A(\mathrm{GS}_{\Sigma,e_{|H_i}}(E),E) \ \text{ for } i \in \{4,6,7\} \end{aligned}$$

Hence,

$$\mathcal{D}_{\mathrm{GS}}(E,E) = \{\{d_3\},\{d_1,d_3\}\}$$

The most general subset diagnoses $\mathrm{GS}_{\Sigma,e_{|H_4}}(E)$ and $\mathrm{GS}_{\Sigma,e_{|H_7}}(E)$ represent the possible situation where only the single defect d_3 is present, i.e. the defect d_3 accounts for all the observed findings that can be accounted for by any other diagnosis. $\mathrm{GS}_{\Sigma,e_{|H_6}}(E)$ and $\mathrm{GS}_{\Sigma,e_{|H_8}}(E)$ express the situation where both d_1 and d_3 are present at the same time. Note that d_1 and d_3 both account for the same finding f_1. ♢

Multiple diagnosis can also be used for examining the various diagnostic solutions resulting from the notion of most general intersection diagnosis. Multiple diagnosis makes it possible to compare the diagnoses resulting from various notions of diagnosis to each other, in the sense of the restriction and subdiagnostic relations. As the following example shows, $\mathcal{D}_{\mathrm{GS}}(E,E')$ and $\mathcal{D}_{\mathrm{GI}}(E,E')$ will generally differ.

Example 5.13. Consider the following diagnostic problem $\mathcal{P} = (\Sigma, E)$, with $\Delta_P = \{d_1,d_2,d_3\}$, $\Phi = \{f_1,f_2,f_3\}$, $E = \{f_1,f_2\}$ and the bottom-up partial specification

$$e(d) = \begin{cases} \{f_1,f_2\} & \text{if } d = d_1 \\ \{f_2\} & \text{if } d = d_2 \\ \{f_1,f_3\} & \text{if } d = d_3 \\ \bot & \text{if } d \in \Delta_N \end{cases}$$

The multiple subset diagnosis is equal to $\mathcal{D}_{\mathrm{GS}}(E,E) = \{\{d_1\},\{d_1,d_2\}\}$, and the multiple intersection diagnosis is equal to $\mathcal{D}_{\mathrm{GI}}(E,E) = \{\{d_1\}, \{d_1,d_2\}, \{d_1,d_3\}, \{d_2,d_3\}, \{d_1,d_2,d_3\}\}$. ♢

From the previous example, it is tempting to conjecture that

$$\mathcal{D}_{\mathrm{GS}}(E,E') \subseteq \mathcal{D}_{\mathrm{GI}}(E,E')$$

The following counterexample demonstrates that this is not the case. Let $E = \{f_1,f_2\}$, $e(d_1) = \{f_1\}$, $e(d_2) = \{f_2,f_3\}$, then $\mathcal{D}_{\mathrm{GS}}(E,E) = \{\{d_1\}\}$ and $\mathcal{D}_{\mathrm{GI}}(E,E) = \{\{d_1,d_2\}\}$.

5.5.2 Minimal diagnosis

Having constructed the multiple, most general diagnoses, diagnoses that are minimal with respect to set inclusion can be selected. For most general subset diagnosis, the minimal multiple diagnosis is denoted by $\mathcal{D}^{\subseteq}_{\mathrm{GS}}(E,E')$, for most general superset diagnosis by $\mathcal{D}^{\subseteq}_{\mathrm{GO}}(E, E')$, and for most general intersection diagnosis by $\mathcal{D}^{\subseteq}_{\mathrm{GI}}(E, E')$. As stated before, the construction of minimal diagnoses may be useful if the number of different diagnoses is very large. In the case of refinement diagnosis, it may be more naturally to focus on one diagnosis with respect to a general hypothesis, such as Δ_P.

5.6 Discussion

In this chapter, several notions of diagnosis have been proposed that are less rigorous in dealing with observed findings and evidence functions than the notions of diagnosis that have appeared in the literature. As was shown, the particular properties of evidence functions to which a notion of diagnosis is applied, are important with respect to the appropriateness of a notion of diagnosis. There are several ways in which the notions of diagnosis discussed in this chapter can be enhanced. A number of these will be briefly reviewed.

In the formalization of a diagnostic problem in Chapter 3, findings associated with a defect were listed, without making an explicit distinction between those findings that are important and those that are not. However, the set of findings may be subdivided into subsets according to several, not necessarily mutually exclusive, criteria, taken as measures of the 'importance' or relevance of findings. Two examples of such criteria are:

- Frequency of occurrence: some findings may always be present given a set of present or absent defects, while others may only be observed occasionally.
- Discriminatory power: the observation of a finding associated with some set of defects, but not with other sets, makes the occurrence of that particular set of defects more likely than the occurrence of the other sets. In medicine, findings with high discriminatory power are known as *pathognomonic* findings.

Many other criteria are possible. For example, ease of observation of a finding may be an additional criterion. These criteria could be incorporated into our notions of diagnosis by decomposing an evidence function into several different evidence functions with different meanings. In fact, the treatment of modularization of a diagnostic specification in Section 4.2.7 is an example of this.

In many problem-solving situations, only a subset of all findings that may be observed is explicitly entered by the observer in reports concerning these situations. Findings that could have been observed, but were not explicitly entered as being positive, are often assumed to be negative, i.e. absent. This is a special case of the closed world assumption (CWA) that is often assumed to hold for database systems [Reiter, 1977]. A typical example of the useful application of this assumption in medical diagnostic problem solving is in the interpretation of the data a clinician has entered for a patient in a medical record. A clinician usually writes down all positive findings that have been observed in the patient and that may be associated with one or more defects. The remaining findings that could have been observed if present, but have not been observed, are implicitly assumed to be negative. This negative information is not written down, but is implicitly assumed. The notions of diagnosis may be enhanced by interpreting the set of observed findings E in this way. The diagnostic interpretation of negative findings could be made more subtle by distinguishing between negative findings that have been entered, and those that have been derived through the CWA.

Part II

A Diagnostic System in Hepatology

Chapter 6

Medical Diagnosis in Hepatology

In the previous chapters, the theoretical aspects of diagnosis have been discussed in much detail. This second part of the thesis focusses on the application of some of these principles to the medical diagnosis of disorders of the liver and biliary tract.

In this chapter, an overview of the problem of diagnosing disorders of the liver and biliary tract is presented from a medical perspective. This chapter may be viewed as a summary of the knowledge acquired to build the HEPAR system. First, some anatomical and physiological principles concerning the liver and biliary tract are briefly discussed. Next, the clinical approach to the patient with a disorder of the liver or biliary tract is reviewed. The general diagnostic strategy followed for these patients is outlined. Furthermore, some of the more frequently applied laboratory tests and procedures in diagnosis are briefly described. A small number of disorders of the liver and biliary tract are described in some detail.

6.1 The liver and the biliary tract

The liver is the largest solid organ in the human body, weighing about 1.5 kg. This organ has an enormous number of different functions, including the formation of bile and urea, carbohydrate and fat metabolism, reduction and conjugation of steroid hormones, and the production of plasma proteins. Most of these functions are metabolic in nature. The liver consists of three major cell types: the *hepatocyte*, the *biliary epithelial cell* and the *Kupffer cell*. The hepatocytes are responsible for most of the metabolic functions mentioned above. The liver is supplied with oxygen by the common hepatic artery, a side-branch of the aorta. The portal vein carries blood that is saturated with nutrients from the gastrointestinal tract to the liver. The blood leaves the liver by the hepatic veins.

The bile, a complex solution composed of water, bile acids (among others cholic acid), bile pigments (biliverdin glucuronide, bilirubin glucuronide), fatty acids, cholesterol, etcetera, is excreted by the hepatocytes through a complicated mesh of small bile ducts into the common bile duct, which drains into the duodenum. Between meals, bile is stored in the gallbladder, a pouch emerging from the common hepatic duct.

Studies show that the liver's reserve capacity to damage is large; recovery of patients has been reported following 80 to 90 per cent resection due to the large capacity of the liver to regenerate [Karran & McLaren, 1985]. In view of the central role the liver plays in bile metabolism, liver damage is often associated with bile excretion derangements, frequently resulting in jaundice. *Jaundice*, which is characterized by, among others, yellow eyes and skin, is due to increased plasma levels of bilirubin. These and other similarities between the symptoms and signs of disorders of the liver and those of disorders of the biliary tract are the main reason to treat these two systems as a single entity. The duodenum and pancreas are considered additional potential sources of disorders of the liver and biliary tract, due to their close anatomical relationship to the common bile duct.

Although a large variety of laboratory tests are available to assess the liver function, there is no single combination of tests that is informative for all patients. As a consequence, the principal approach to the patient with a disorder of the liver and biliary tract is strongly clinical in nature, i.e. information from history and physical examination is an essential ingredient for the proper diagnostic management of these patients.

6.2 Clinical diagnosis in hepatology

In this section, a review of diagnosis in hepatology is presented.

6.2.1 Approach to the patient

The central problems in the diagnosis of disorders of the liver and biliary tract in the patient are [Karran et al., 1985]:

(1) To determine whether the disorder is primarily affecting the hepatocytes (*hepatocellular disorder*) or primarily affecting the biliary tract (*biliary obstructive disorder*);

(2) To establish whether the disorder is acute or chronic in nature;

(3) To recognize whether the disorder has benign or malignant features.

Based on this information, it is often possible to develop a plan for further diagnostic assessment to reach an acceptable *differential diagnosis*, i.e. a small set of disorders where each disorder more or less fits the findings observed in the patient. From this differential diagnosis, a single disorder with strongest evidence, often called the *'final diagnosis'* or *'definite diagnosis'* may be selected if sufficient evidence is available.

As a matter of terminology, in the medical literature the collection of derangements associated with the diseases in the group of hepatocellular disorders is sometimes referred to as intrahepatic cholestasis, whereas the set of derangements associated with the diseases in the group of biliary obstructive disorders is sometimes referred to as extrahepatic cholestasis. It should, however, be noted that disorders with extrahepatic cholestasis may also involve, or even be limited to, the intrahepatic parts of the biliary tract. Other authors, therefore, use the terms intrahepatic and extrahepatic cholestasis in another sense, namely to signify a distinction at the anatomical level (inside the liver – intrahepatic – versus outside the liver – extrahepatic). Another possible source of confusion is the use of the term 'cholestasis' in the literal sense of stagnation (stasis) of bile; it has traditionally been used to describe the accumulation of bile as seen under a light microscope. In this thesis, the terms hepatocellular disorder and biliary obstructive disorder are used instead, because of their less ambiguous meaning. When the connotation implied by the term 'disorder' seems undesirable, the pathophysiological terms *hepatocellular derangement* or *hepatocellular damage* and *biliary obstruction* will be used instead.

The aetiology of hepatocellular disorders varies widely. Some causes of disease belonging to the class of hepatocellular disorders are:

- Alcohol abuse, which may give rise to alcoholic hepatitis, alcoholic cirrhosis, steatosis hepatis, Zieve's syndrome.
- Viral infection; examples of viral disease which affects the liver are hepatitis A, B and C, and cytomegalic inclusion disease.
- Autoimmune disease, such as causing autoimmune chronic hepatitis.
- Inborn errors of metabolism, such as in Gilbert's syndrome, which is caused by a deficiency of the enzyme glucuronyl transferase in the hepatocyte, or Wilson's disease in which a lack of the copper-binding α-globulin caeruloplasmin is found.

Since in all these disorders, the hepatocyte is affected, they share several symptoms and signs.

In biliary obstructive disorders, some form of obstruction always exists in the small, intrahepatic bile ducts or in the large bile ducts, i.e. the left and right hepatic ducts and the common bile duct. Obstruction of the biliary tract may be caused by:

- The presence of gallstones in the common bile duct.
- A benign or malignant tumour, such as pancreatic carcinoma, which may obstruct the common bile duct, bifurcation carcinoma, which may cause obstruction of the hepatic ducts and common bile duct at the porta hepatis, or a metastatic tumour in the liver, which may cause obstruction of the intrahepatic biliary tract due to compression of the surrounding tissue.
- Destruction of the small bile ducts, as found in primary and secondary biliary cirrhosis.

In advanced liver disease, there is often destruction of both the hepatocytes and the bile ducts, and the two clinical pictures will merge. Several of the disorders mentioned above will be treated in more detail in Section 6.7.

One of the interesting features of the area of hepatology is its strong clinical basis. Although a rapidly increasing number of diagnostic tests have become available in the past two decades, a careful and thorough history and physical examination, supplemented with a small number of laboratory tests, are still of overriding importance in the diagnosis of disorders of the liver and biliary tract. Clinical information often provides sufficient evidence concerning the underlying pathology and aetiology of the disease, and may even indicate a definite diagnosis. Supplementary diagnostic investigations, such as endoscopic retrograde cholangio-pancreaticography (ERCP), need therefore only be performed in a limited number of carefully selected patients. To this end, a clear plan of investigation

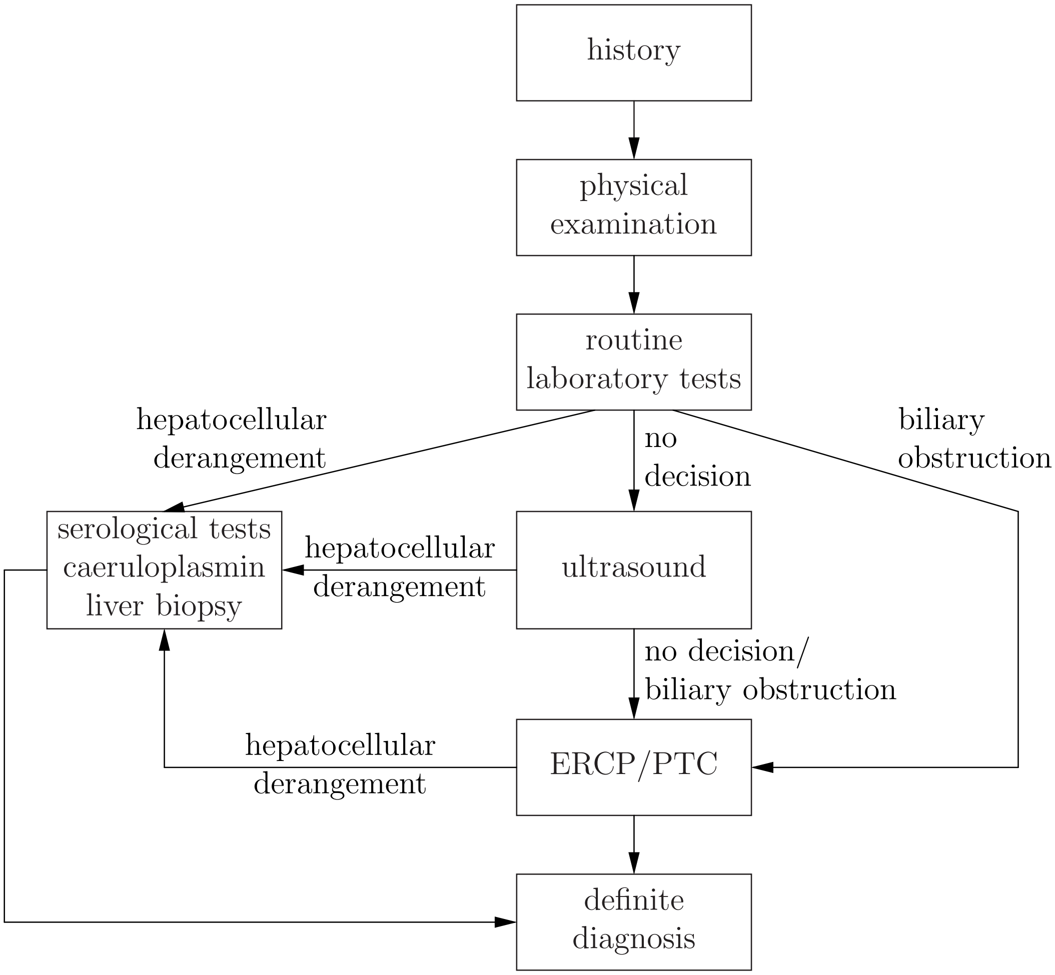


Figure 6.1: Diagnostic plan in patients with liver or biliary tract disease.

is of the utmost importance. The diagnostic plan that has been taken as the basis for the development of the HEPAR system is depicted in Figure 6.1. This diagnostic plan is commonly followed by the clinician in dealing with a disorder of the liver or biliary tract. This same structure has been adopted in the HEPAR system (cf. Chapter 7). We shall first review the process of history taking and physical examination in hepatology, and discuss some of the most frequently applied laboratory tests, before discussing some of the diagnostic procedures in more detail.

6.2.2 Patient history

A patient with a disease of the liver or biliary tract will typically have jaundice, i.e. yellow sclerae, dark urine and pale, clay-coloured stools. However, in some patients none of the features may be present, and the suspicion of the presence of a liver disease may only be based on coincidentally detected abnormal laboratory findings. Information from the patient history may point to a specific disease. For example, if a patient has been in contact with a jaundiced subject, has visited a (sub)tropical area, has been transfused with blood products or is a heroin addict, some form of viral hepatitis may be suspected.

A patient with jaundice due to a biliary obstructive disorder (cholestatic jaundice) will typically have dark urine and pale stools. A history of biliary colics is strong evidence that the jaundice is caused by stones in the common bile duct. A dry mouth and burning eyes may indicate the disorder to have an autoimmune origin. If the patient is suffering from pain, its nature (paroxysmal or continuous), location, possible radiation to other parts of the body, and the relation to food intake is important information to differentiate between various disorders. Information from the *family history* may provide evidence for a genetic origin of a disorder.

6.2.3 Physical examination

Physical examination of the patient may yield important information about the severity and aetiology of the disorder, and may sometimes even yield a definite diagnosis. For example, the presence of Kayser–Fleischer rings on inspection of the patient's eyes leads immediately to the conclusion that the patient probably has Wilson's disease. Certain cutaneous stigmata, such as erythema of the palm of the hands (palmar erythema), indicate the presence of a chronic dysfunction of the hepatocytes. Splenomegaly (enlarged spleen) and caput medusae (tortuous veins around the navel) indicate portal hypertension, i.e. increased pressure in the portal venous system usually associated with cirrhosis of the liver. A palpable gallbladder in a patient without pain (Courvoisier's sign) or an epigastric mass discovered by palpation, may be explained by the presence of a pancreatic carcinoma. Percussion of the abdomen may yield evidence about the presence of ascites (abnormal protein-rich fluid in the abdominal cavity).

6.2.4 Diagnostic tests

The routine laboratory tests to investigate the nature of a disease of the liver and biliary tract are the serum levels of conjugated and unconjugated bilirubin, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, $5'$-nucleotidase, and γ-glutamyl transferase. These laboratory tests are used to differentiate between hepatocellular derangement and biliary obstruction.

Unconjugated bilirubin is a break-down product of haemoglobin, an oxygen-carrier metalloprotein present in erythrocytes. Most of this unconjugated bilirubin is reversibly bound to plasma albumin. It is taken up by the hepatocytes by a carrier-mediated transport mechanism, and inside the cell conjugated to bilirubin glucuronide by the enzyme glucuronyl transferase. Conjugated bilirubin is actively excreted in the bile. The serum level of conjugated bilirubin is elevated in most diseases of the liver and biliary tract. The serum level of unconjugated bilirubin is increased in haemolytical anaemia (not covered in HEPAR), and in several liver diseases such as Gilbert's disease and Zieve's syndrome.

Urine is usually examined on the presence of urobilin and conjugated bilirubin. Conjugated bilirubin is partially oxydated to urobilinogen in the intestines, part of which is reabsorbed to the blood. As a water-soluble substance, some urobilinogen is excreted into the urine where it is oxydated to urobilin. The absence of urobilin in urine indicates that the transport of conjugated bilirubin to the intestines is blocked, due to biliary obstruction. Since biliary obstruction leads to the accumulation of conjugated bilirubin in the blood, again a water-soluble substance, increased amounts of bilirubin can be found in urine. Unconjugated bilirubin is not excreted in the urine due to its binding to plasma albumin; under normal conditions albumin cannot cross the glomerular basement membrane in the kidney.

The tests mentioned above belong to a wide range of biochemical tests of liver function and presence of hepatocellular damage or biliary obstruction, commonly called '*liver function tests*'. More about these tests will be said in Section 6.3. It should be noted that the outcomes of the routine laboratory tests merely guide the selection of further diagnostic tests as indicated in Figure 6.1; they do not offer conclusive evidence.

If the evidence indicates hepatocellular damage, several supplementary laboratory tests are available to differentiate between various hepatocellular disorders. In particular serological tests, such as hepatitis B serology, and the presence of hepatitis A IgM, cytomegalovirus, smooth muscle, nuclear, or mitochondrial antibodies provide important evidence for a particular disease. Section 6.4 reviews the most important serological tests.

If insufficient evidence concerning the hepatocellular or biliary obstructive nature of a disorder has been gathered for a patient, ultrasonography may be carried out, providing more specific evidence regarding disease of the biliary tract and the pancreas. Ultrasonography of the liver and biliary tract, and of anatomical structures in their direct neighbourhood, is discussed in Section 6.5.

6.3 Biochemical hepatobiliary assessment

In this section, the most important biochemical tests used in the assessment of liver function in general, extent of hepatocellular damage, presence of biliary obstruction or hepatocellular malignancy are reviewed. For a more detailed treatment of the subject, the reader is referred to [Price & Alberti, 1985].

6.3.1 Assessment of liver function

Hepatocytes play an important role in the synthesis of plasma proteins. The α- and β-globulins are only produced in the liver. Quantitatively the largest amount of protein synthesized by the liver is albumin, the protein that is responsible for a significant part of the colloid osmotic pressure of plasma. A low level of albumin, hypoalbuminaemia, is often found in chronic liver disease. However, there are several other conditions that may give rise to hypoalbuminaemia, such as malnutrition, which therefore should be ruled out.

The liver also plays an important role in the metabolism of steroid hormones, such as progesterone and the oestrogens, and the androgenic hormone testosterone. Although the precise endocrinological basis is unclear, in male patients chronic liver disease may result in gynaecomastia (female-like breast development) and testicular atrophy; disturbance of the steroid metabolism may be causally related to these signs.

In addition to metabolic functions, the liver also has a function in the storage of several compounds. Only the functions that are important for the diagnosis of liver disease, i.e. the storage of iron and copper, will be mentioned. Iron is bound to transferrin, a protein that transfers iron to the tissues. Excess of iron is stored in the liver as ferritin and haemosiderin. A chronic overload of iron may result in a form of liver damage called haemochromatosis. Copper is another compound stored in the liver. In the recessively inherited liver disease, Wilson's disease, there is an excess of copper deposits in the liver, especially in homozygous patients, eventually causing symptoms and signs of liver disease. Furthermore, the levels of the copper-binding protein caeruloplasmin, that is produced by the liver, is reduced. The serum level of caeruloplasmin may be determined if Wilson's disease is suspected in the patient.

6.3.2 Tests of hepatocellular damage

The levels of the aminotransferases, aspartate aminotransferase (ASAT) and alanine aminotransferase (ALAT), are the tests most frequently employed by the clinician to assess hepatocellular damage. The information obtained by these tests is not very specific, i.e. which liver disease may be involved cannot be determined by these tests. Both enzymes are present in high concentration in the hepatocyte. The serum levels of these enzymes may be increased due to leakage of cytoplasm into the circulation, caused by increased membrane permeability or breakdown of the membrane. In acute liver disease, the concentrations of these enzymes may rise 20 to 50 times the upper limit of the normal range; in chronic liver disease, the elevation will be more moderate, about five times the upper limit of the normal range. In disorders that only lead to increased cell membrane permeability, such as acute hepatitis, the ALAT level is usually more increased than ASAT levels. Similarly, in biliary obstruction ALAT levels are usually more elevated than ASAT levels. On the other hand, for hepatic necrosis, the levels of ASAT may rise above those of ALAT. A possible explanation of these observations is that although both enzymes are present in the cytoplasm of the hepatocyte, in contrast with ALAT, ASAT can also be found in the mitochondria; only cytoplasm will leak through the cell membrane.

The levels of ASAT and ALAT are usually studied relative to the serum concentrations of alkaline phosphatase, γ-GT or $5'$-nucleotidase, three compounds used in the assessment of biliary obstruction (cf. Section 6.3.3). In case of hepatocellular derangement, typically the levels of ASAT and ALAT are increased whereas the levels of alkaline phosphatase, $5'$-nucleotidase and γ-GT are normal, or slightly increased. The elevation of alkaline phosphatase, $5'$-nucleotidase and γ-GT can be explained by the fact that in many patients with a hepatocellular disorder, some features of biliary obstruction are present.

6.3.3 Tests of biliary obstruction

The biochemical tests of biliary obstruction are used to determine whether or not the bile flow is obstructed somewhere between the small bile canaliculi enclosed between the hepatocytes and the large common bile duct. The pathophysiological basis of biliary obstructive symptoms and signs is the (possibly partial) inability to excrete bile. As a consequence, several bile components appear in blood. One of these components is conjugated bilirubin discussed in Section 6.2.4. The most frequently employed laboratory test to investigate the presence of biliary obstruction in the patient is alkaline phosphatase (AP). The concentration of this enzyme is raised up to ten times the upper limit of normal range in patients with biliary obstructive disease. In particular, the concentrations are high in extrahepatic obstruction, e.g. obstruction due to pancreatic carcinoma, but may also be high in intrahepatic obstruction.

A disadvantage of the alkaline phosphatase test is that the concentration may also be increased in the presence of bone disease. Whether increased alkaline phosphatase levels are related to liver or bone disease can be established by the serum levels of $5'$-nucleotidase, an enzyme that is only raised in serum for liver disease. There are patients, however, for whom this does not hold: the concentration of $5'$-nucleotidase does not follow that of alkaline phosphatase, even though liver disease is present. The enzyme γ-glutamyl transferase (γ-GT) is used in a similar way to $5'$-nucleotidase; serum levels of this enzyme are particularly high in biliary obstructive disorders.

In summary, classical biliary obstruction causes increased levels of alkaline phosphatase, $5'$-nucleotidase and γ-GT, and normal levels of ASAT and ALAT. However, in some biliary obstructive disorders features of hepatocellular derangement can be observed, as demonstrated by the fact that the concentrations of ASAT and ALAT are above the normal upper limits.

6.3.4 Tests for hepatobiliary malignancy

In hepatic malignancy, several biochemical tests discussed above may yield abnormal results, depending on the state of the tumour. For example, the tumour mass may, due to compression of surrounding liver parenchyma, give rise to intrahepatic biliary obstruction causing the concentration of alkaline phosphatase to rise. It has been shown that infiltration of liver parenchyma by tumour cells may also produce increased levels of the aminotransferases.

For primary hepatocellular carcinoma, a malignant tumour arising from the liver parenchyma, a frequently applied test is the detection of α-foetoprotein in blood. The compound α-foetoprotein is a protein normally present in the fetus but not in the adult; it is synthesized again by the tumour cells.

Tumours of the biliary tract and pancreas may give rise to disturbances of biochemical parameters as discussed above for biliary obstruction.

6.4 Immunological and serological tests

There are several disorders of the liver and biliary tract that have associated immunological disturbances. Below, the immunological changes that are directly relevant for diagnosis will be briefly reviewed.

6.4.1 Auto-antibodies in hepatobiliary disease

In a number of disorders of the liver and biliary tract antibodies directed against the patient's own cellular components can be detected. Antibodies directed against the components of the cell nucleus can be observed in autoimmune chronic hepatitis and primary biliary cirrhosis (cf. Section 6.7). These antibodies may give rise to the LE-cell phenomenon, i.e. polymorphonuclear leukocytes that incorporate engulfed nuclear material from lymphocytes as a large homogeneous mass.

Antibodies directed against smooth muscle cells, in particular binding to actin molecules in these cells, can be found in high titres in patients with autoimmune chronic hepatitis; these antibodies may also be present in primary biliary cirrhosis and viral hepatitis.

Antibodies directed against the mitochondria may be found in over 85 per cent of all patients with primary biliary cirrhosis. Because in other hepatobiliary disease mitochondrial antibodies are less frequently encountered, the demonstration of mitochondrial antibodies in blood is a valuable test in the diagnosis of primary biliary cirrhosis.

6.4.2 Antigens and antibodies in viral hepatitis

Hepatitis A

Hepatitis A is a viral infection caused by a small RNA virus, called HAV (Hepatitis A Virus). In the presence of this infection, antibodies against HAV, consisting of IgG and IgM fractions, can be detected. A raised anti-HAV IgM titre is diagnostic evidence for acute infection with HAV. The total titre of anti-HAV remains high after hepatitis A infection, and actually is evidence that the patient is now immune to infection with the virus. However, anti-HAV IgM will not be detected after about ten weeks following the initial symptoms.

Hepatitis B

Hepatitis B is a viral infection of the liver, caused by a small DNA virus, called HBV (Hepatitis B Virus). Several components of the virus act as antigens. The complete virus is broken down to a core and an envelope. Core antigens are known as HBcAg and HBeAg; the envelope antigen is called HbsAg. The presence of these antigens in blood gives rise to the production of antibodies directed against these antigens. The antibodies are known as HBcAb, HBeAb and HBsAb, respectively. They are also referred to as anti-HBc, anti-HBe and anti-HBs, respectively. The HBV antigens and antibodies in blood are clinically of significant importance in diagnosing and monitoring the course of hepatitis B infection. The titres of the HBV antigens and antibodies vary during the course of the disease. Figure 6.2 depicts the titres of the HBV antigens and antibodies as a function of time.

In acute hepatitis B, there is an initial response consisting of HBcAb, even preceding the symptoms of the disease, together with the presence of HBsAg. In time following the detection of HBcAb in serum, HBeAb can be found. HBeAg is only detectable during the early phase of acute hepatitis B. HBsAb titres slowly increase after the disappearance of HBsAg from serum.

In chronic hepatitis B, the substances HBsAg, HBcAb and HBeAb can be detected in the patient's serum. (This cannot be read off from Figure 6.2, which assumes the virus to be nonpersistent.)

Hepatitis C

In more than 90 per cent of the patients, viral hepatitis following blood transfusion is caused by HCV (Hepatitis C Virus), which is a small RNA virus [Alter et al., 1989;

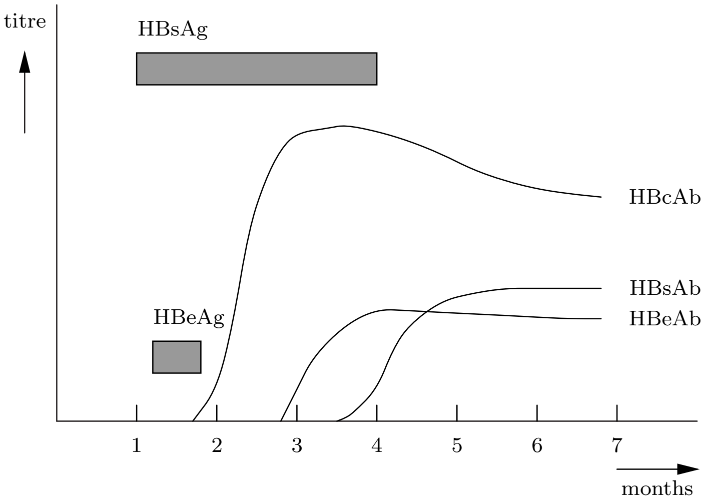


Figure 6.2: Course of hepatitis B infection [Wright et al., 1985b].

Robbins et al., 1994]. It has been shown that hepatitis C also may be transmitted by organ transplantation [Pereira et al., 1991]. Hepatitis C was previously known as hepatitis non-A non-B. Symptoms and signs of hepatitis C are similar to those of hepatitis B, but jaundice is encountered less frequently in patients with hepatitis C than with hepatitis B. In general, symptoms and signs are milder than those found in hepatitis A or B. Antibodies to HCV can be demonstrated in about 85 per cent of patients having hepatitis C [Alter et al., 1989].

Hepatitis D

Hepatitis D is an infection of the liver that can only occur when there is concomitant hepatitis B infection. It is caused by an RNA virus also called 'delta agent' or HDV. IgM anti-HDV antibodies are detectable in blood [Robbins et al., 1994]. This disease has not been included in the HEPAR system because of its strong relationship with hepatitis B.

Infectious mononucleosis

Infectious mononucleosis is an infection caused by the Epstein-Barr virus (EBV). Although infectious mononucleosis is not primarily a liver disease, in the majority of patients with infectious mononucleosis, biochemical evidence of disturbed liver function can be demonstrated. Tests for infectious mononucleosis are therefore used to differentiate infectious mononucleosis from hepatitis A, B and C. By means of the Paul–Bunnell test, serological evidence for the presence of the disease can be collected.

Cytomegalic inclusion disease

Cytomegalic inclusion disease is an infection due to the cytomegalovirus; inclusion of this virus in the hepatocyte and vascular endothelial cells can be demonstrated in about 20 per cent of cases. Elevated IgM antibodies against the virus can usually be demonstrated in serum of patients with this disease. As for infectious mononucleosis, this tests is primarily used to differentiate this infection from hepatitis A, B and C.

6.4.3 Serological tests in bacterial and parasitic liver disease

Serological tests to demonstrate that a particular organism is the cause of a disorder of the liver or biliary tract, are usually specific for that organism. We therefore review the most important tests in relationship with the specific disorders caused by the organisms tested for.

Syphilis

Syphilis is a bacterial infection caused by Treponema pallidum. The liver may be affected in all forms of this disease. In particular, the disease may present itself in a way very similar to acute hepatitis. There are many different serological tests available, with varying sensitivity, to demonstrate the presence of syphilis in a patient. Tests that demonstrate the presence of antibodies to Treponema pallidum, such as the Fluorescent Treponema Antibody Absorption (FTA-Abs) test, have high sensitivity and specificity.

Toxoplasmosis

Toxoplasmosis is a parasitic infection caused by the protozoa Toxoplasma gondii. The organism may affect all tissues in the body. The clinical presentation of toxoplamosis may be very similar to that of viral hepatitis. A frequently applied serological test in the diagnosis of toxoplamosis is the Sabin–Feldman test, a specific test to demonstrate the presence of antibodies directed against Toxoplasma gondii in the serum.

Amoebic liver disease

Amoebic liver disease is usually caused by Entamoeba histolytica, a protozoa that may give rise to intestinal ulcers (intestinal amboebiasis). Entamoeba histolytica may reach the liver through the portal vein, giving rise to the formation of liver abscesses of varying size. There are a number of serological tests available to demonstrate the presence of antibodies directed against Entamoeba histolytica in the serum.

Echinococcosis of the liver

Echinococcosis of the liver is caused by the flatworms Echinococcus granulosus, Echinococcus multilocularis or Echinococcus oligarthrus. In about 70 per cent of patients, these worms will give rise to the formation of cysts in the liver, known as hydatid cysts. Specific serological tests to demonstrate the presence of echinococcus antibodies in the patient are the immunoelectrophoresis test and indirect haemagglutination test.

6.5 Ultrasonographical and radiological investigation

In recent years, ultrasonography has become one of the major diagnostic techniques in the diagnosis of disorders of the liver and biliary tract. The conventional plain radiograph of the abdomen is therefore only of limited use.

6.5.1 Ultrasonography of the liver and biliary tract

Ultrasonography has the advantage that it is non-invasive in nature, and that it is easy to perform. A disadvantage is that its accuracy varies widely between various clinical centres. The lack of a standardized terminology hinders the communication between physicians on the results.

In general, ultrasonography of the liver and biliary tract is employed to investigate:

- The presence and cause of hepatomegaly (enlargement of the liver);
- The presence of hepatic metastases after diagnosis of primary malignancy at some other site;
- The cause of jaundice.

The main use of ultrasonography in hepatology is to differentiate hepatocellular causes of jaundice and disturbed liver function from biliary obstructive causes. As shown in Figure 6.1, ultrasonography of the liver and biliary tract is a diagnostic technique applied early in the diagnostic process.

Normally, the biliary tract cannot be visualized by ultrasonography due to its small diameter. In case of biliary obstruction, dilatation of the bile ducts can be demonstrated in almost all patients. Furthermore, it is often possible to localize the level of obstruction. In case of obstruction at the level of the left and right hepatic ducts or higher, only the intrahepatic bile ducts are dilated, whereas in case of biliary obstruction at the level of the common bile duct, e.g. due to pancreatic carcinoma or carcinoma of the papilla of Vater, both intrahepatic and extrahepatic bile ducts are dilated. Demonstration of the level of biliary obstruction can be hindered by the presence of gas in the duodenum or stomach, making it impossible to visualize the dilated common bile duct.

Some pathological conditions of the gallbladder, e.g. gallstones, or of the pancreas, e.g. pancreatic pseudocyst or carcinoma, can also be detected by ultrasonography.

Ultrasonography, being an important technique in the early assessment of hepatobiliary disease, has been incorporated as one of the tests in the HEPAR system (cf. Chapter 7). As stated above, one of the problems with ultrasonography of the liver and biliary tract is the lack of a standardized terminology to describe the findings observed. For the two validation studies of the performance of HEPAR, a systematic terminology was designed. The basic requirement for this terminology was that ultrasonography reports from Dijkzigt University Hospital and from Leiden University Hospital could be translated automatically to this standard terminology.

In the following sections, we briefly review the terminology used to describe the contents of ultrasonography reports in the HEPAR project.

Ultrasonography of the liver

Findings of ultrasonographical assessment of the liver have been described using the following concepts:

(1) Size of the liver (normal, enlarged, too small);

(2) Density of the liver parenchyma (normal, hyperdense, hypodense);

(3) Specific findings of liver parenchyma (normal, acoustic shadows, haemangioma, hypodense areas, hyperdense areas, multiple cysts, solitary cyst, solid mass(es));

(4) The liver contour (smooth, nodular).

Ultrasonography of the biliary tract

Assessment of the bile ducts:

(1) Findings concerning the intrahepatic bile ducts (normal, dilated, hyperreflective);

(2) Findings concerning the extrahepatic bile ducts (normal, dilated, stone, obstruction at the papilla of Vater, obstruction in the pancreas, common bile duct obstruction, stricture, tumour).

Other ultrasonographical findings

A number of possible findings of ultrasonographical investigation have been classified under the heterogeneous heading of ‘other findings’:

(1) Presence of ascites in the patient;

(2) Description of the gallbladder (normal, acoustic shadows, dilated, polyp, shrivelled, signs of cholecystitis, sludge/debris, stone, thickening of wall, tumour);

(3) Description of the hilar region (normal, shows cyst, solid mass);

(4) Description of the hepatic veins (normal, thrombosis, without flow, not visible);

(5) Description of the pancreas (normal, calcifications, cystic process(es), diffusely enlarged, dilated pancreatic duct, signs of pancreatitis, solid mass, visualized);

(6) Description of the portal vein (normal, dilated, occluded, revised flow);

(7) Description of the vena cava inferior (not visible by ultrasound – which is normal –, thrombosis);

(8) Description of the spleen (normal, enlarged, splenic vein occluded).

6.5.2 Radiological investigations

A plain radiograph of the abdomen may provide valuable information about the presence of calcified gallbladder stones, or demonstrate a sentinel loop sign which is typical for pancreatitis.

6.6 Other diagnostic techniques

Other frequently employed diagnostic techniques in hepatology are percutaneous transhepatic cholangiography (PTC), endoscopic retrograde cholangio-pancreaticography (ERCP) and liver biopsy. These procedures are all invasive in nature, and are only used in the diagnostic process if no conclusion has been reached concerning the hepatocellular or obstructive nature of the disorder, or to confirm a particular diagnosis (See Figure 6.1). PTC and ERCP provide information about alterations in the biliary tract. For example, in carcinoma of the head of the pancreas, obstruction of the common bile duct and the main pancreatic duct can often be demonstrated. In primary sclerosing cholangitis, the typical findings of ERCP are multiple strictures (narrowings) with 'beading' of ducts between the narrow segments, and involvement of both the intrahepatic and extrahepatic duct systems.

PTC and ERCP are not usually carried out early in the diagnostic process; for this reason and because of the invasive nature of these diagnostic techniques, results of PTC and ERCP have not been incorporated in the HEPAR system.

6.7 Disorders of the liver and biliary tract

In this section, we briefly review the clinical features of a few disorders of the liver and biliary tract. This gives the reader an impression of the information with which the clinician starts when faced with the problem of establishing which disorder is responsible for the symptoms and signs observed in the patient. Eight of the about eighty disorders covered in HEPAR are briefly reviewed. We shall refrain from providing pathological and pathophysiological detail in describing these disorders, because such information is not essential in the very early stages of diagnosis. For a more detailed account on the subject, the reader may consult the standard textbooks, for example [Wright et al., 1985a]. The disorders are subdivided into four different categories, the meaningful combinations obtained from the diagnostic categories introduced in Section 6.2.1.

6.7.1 Acute hepatocellular disorders

An acute hepatocellular disorder generally develops within two weeks; they are associated with hepatocellular derangements.

Hepatitis B

Early symptoms in patients with hepatitis B are anorexia, nausea, malaise, weight loss, fever, dark urine and pale stools. Sometimes the patient has also generalized pruritus (itch). After some time, the severity of these symptoms decreases, and jaundice develops in addition to abdominal (hepatic) pain. Jaundice and hepatomegaly may be the only findings detected by physical examination. All these symptoms and signs may also be found in other forms of viral hepatitis, malaria, amoebic liver disease and several other disorders.

There are several ways in which hepatitis B can spread through the population. Information from the disease history may sometimes indicate that the patient has been infected through one of these familiar routes. Several years ago, hepatitis B could be the result of transfusion with blood products, which is rare nowadays, due to improved quality control at the blood banks. Nowadays, hepatitis B is more commonly spread by close personal contact, contaminated syringes (such as used by heroin addicts). Personal contact as a cause is less common for hepatitis B than for hepatitis A.

Alcoholic hepatitis

Alcoholic hepatitis is an acute liver disease in patients with alcohol abuse, in which the typical symptoms are anorexia, nausea, vomiting, abdominal pain. Physical examination may bring to light hepatomegaly and a painful liver on palpation. The biochemical findings are those of hepatocellular damage, with elevated total (conjugated and unconjugated) bilirubin in blood and a γ-GT/ALAT ratio lower than 5. By means of ultrasonography of the liver an enlarged liver that is hyperdense may be demonstrated, but in contrast to alcoholic cirrhosis the liver is not nodular.

6.7.2 Chronic hepatocellular disorders

Chronic hepatocellular disorders develop over a period of several months. In the initial phase, symptoms and signs are primarily due to hepatocellular damage.

Autoimmune chronic hepatitis

Autoimmune chronic hepatitis is a chronic liver disease of unknown origin, characterized by hepatocellular failure. Typically, the disease occurs in young women. On physical examination, the patient may have mild jaundice, spider angiomas (small blood vessel tumours), palmar erythema and butterfly erythema resembling the skin rash encountered in lupus erythematosus, an autoimmune disease affecting multiple organs. In young women, amenorrhoea and acne may be present; in the male patient one may observe cutaneous striae and gynaecomastia. Serum aminotransferase levels are almost invariably elevated. Immunological disturbances that may demonstrated in the serum are smooth muscle antibodies, antinuclear antibodies and in about 15 per cent of the patients LE cells may be found (LE-cell phenomenon). The presence of these and other antibodies is also reflected by the hypergammaglobulinaemia (increased levels of γ-globulin in the serum) that is almost universally encountered in these patients.

Wilson's disease

Wilson's disease (hepatolenticular degeneration) is caused by a genetic defect of copper metabolism, with a recessive inheritance pattern. Only homozygous patients develop the full clinical picture, although heterozygous patients may show some signs of the disease. The disease is slowly progressive; it usually takes until the patient is seven years old before symptoms and signs of hepatic disease become apparent. Somewhat later, neurological signs become manifest. These symptoms and signs are caused by the toxic effects of increased levels of copper on hepatic and neural tissue. Hepatic symptoms of the disease include fatigue, abdominal pain, spider angiomas, jaundice, oedema, ascites, oesophageal varices and splenomegaly. Deposits of copper in Descemet's membrane in the cornea cause brown rings at the peripheral cornea, known as Kayser–Fleischer rings.

Biochemical evidence for the disease can be collected by measuring the caeruloplasmin concentration, a copper-binding protein, in serum which is in 95 per cent of the patients below 200 mg/l. In addition, there is excess free (i.e. not bound to caeruloplasmin) copper in serum, and excretion of copper in urine is increased.

6.7.3 Benign biliary obstructive disorders

The principle feature of benign biliary obstructive disorders is obstruction to bile outflow, either at the level of the large, extrahepatic bile ducts or at the level of the small intrahepatic bile ducts.

Common bile duct stones

Symptoms and signs of common bile duct stones are caused by the mechanical obstruction of the common bile duct to the outflow of bile produced by the liver. It is believed that most of these stones originate in the gallbladder, although some stones have their origin in the bile ducts. Symptoms due to common bile duct stones include abdominal pain, often colicky (i.e. paroxysmal, spasmodic and severe) in nature probably due to the obstruction, jaundice, which is often associated with pruritus, dark urine and pale stools. If partial obstruction of the common bile duct is present, jaundice need not always develop. If obstruction is complete, urobilinogen cannot be detected in urine (cf. Section 6.2.4).

Accumulated bile is a good substrate for bacterial growth, causing infection of the common bile duct (acute cholangitis). The principal symptom of such an infection is fever associated with chills. Results of physical examination are tenderness at the right upper quadrant of the abdomen and signs of jaundice. Biochemical tests show an increase in alkaline phosphatase; if the concentrations of ASAT and ALAT are also elevated, hepatocellular damage due to raised biliary pressure should be suspected. Although jaundice is not always marked, total bilirubin concentration is usually increased.

Primary biliary cirrhosis

Primary biliary cirrhosis is a chronic liver disease in which the small intrahepatic bile ducts are affected. Its aetiology is unknown. Typically, the patient is a middle-aged woman; only about 10 per cent of the patients are male. An important early symptom in this disease is generalized pruritus. In some patients, xanthomas may be present, as well as Kayser–Fleischer rings. Alkaline phosphatase is usually increased as are the levels of γ-GT and $5'$-nucleotidase, indicating biliary obstruction. Immunological investigation shows mitochondrial antibodies in the serum in about 90 per cent of the patients. The cholesterol concentration is often elevated.

6.7.4 Malignant biliary obstructive disorders

In malignant biliary obstructive disorders, obstruction of the small or large bile ducts is caused by the tumour mass of the malignancy. Obstruction of the small intrahepatic bile ducts is caused by compression of the surrounding normal hepatic tissue by tumour mass. When the tumour involves the biliary ducts, as it does in common bile duct carcinoma, or when the biliary ducts run through tumour tissue, as in pancreatic carcinoma, obstruction is more direct.

Pancreatic carcinoma

Carcinoma of the pancreas is usually localized in the pancreas' head. The average age of patients is about sixty years; the disease affects about twice as many males as females.

The patient experiences significant weight loss (often more than $\frac{1}{2}$ kg per week), significant jaundice, pruritus, nausea and anorexia. Routine laboratory tests indicate biliary obstruction.

Ultrasonography may reveal a pancreatic solid tumour, and will almost always demonstrate dilatation of the intrahepatic and extrahepatic bile ducts.

Primary hepatocellular tumour

Primary hepatocellular tumour is a carcinoma that frequently develops in the patient with chronic liver disease, e.g. cirrhosis, haemochromatosis and chronic hepatitis B. In addition to the symptoms and signs caused by these chronic disorders, the patient may experience abdominal pain, jaundice, malaise, anorexia, weight loss, nausea and fever. Characteristic findings of physical examination are enlarged, tender liver and sometimes a palpable hepatic mass, ascites and sometimes distended abdominal veins. In about 25 per cent of the patient a bruit over the liver may be heard on auscultation. The routine biochemical tests show elevated concentrations of total bilirubin, as well as elevated levels of alkaline phosphatase, ASAT en ALAT. A useful and frequently applied test that is specific for primary hepatocellular tumour is the demonstration of α-foetoprotein (cf. Section 6.3.4).

6.8 Discussion

It will be clear from the overview above that diagnosis in the field of hepatology is a complicated matter. The question now arises which of the diagnostic theories treated in the first part of this thesis is most suitable for diagnosing disorders of the liver and biliary tract.

When using a model of the normal structure and behaviour of the liver and biliary tract it is possible, at least in principle, to discover new, previously unknown disorders in patients. Unfortunately, this approach is not feasible in practice, because the information that would be required to drive such models is only available under laboratory conditions, and not in the clinic. This is certainly true during the early assessment of patients, when only information from history, physical examination and routine laboratory tests is available. But the most important limitation of this approach is that known disorders cannot be characterized sufficiently precisely in terms of normal structure and function of the liver and biliary tract. It was our aim to develop a diagnostic system, sufficiently accurate for real-life application. Therefore, knowledge about abnormal structure and function would be indispensable.

Malfunction of the liver and biliary tract is traditionally described in the medical literature in terms of disorders, which in turn are described as specific clinical patterns. Above, several disorders have been described in this fashion. Some of this knowledge is causal in nature; another part is empirical in nature. Thus, from a practical point of view, causal as well as empirical knowledge about disorders, as available from the literature and specialists in the field of hepatology, offer the best foundation for a diagnostic knowledge base in hepatology. In the terminology of our diagnostic framework, the evidence function e that is used to represent knowledge concerning disorders of the liver and biliary tract may be interpreted as denoting causal relations and empirical associations.

A complementary issue concerns the notion of diagnosis that best fits diagnostic problem solving in hepatology. Although co-occurring disorders in a patient are likely to interact with each other, it is unlikely that a patient suffers from more than one disorder of the same organ system at the same time. In particular, is it unlikely that a patient has more than one disorder of the liver and biliary tract at the same time, unless disorders are causally related to each other. Furthermore, in establishing a diagnosis, a physician tries to account for as many of the observed findings as possible. In a field like hepatology, it is often not possible to establish a diagnosis that accounts for all observed findings. Collecting disorders, each accounting for part of the observed findings, disregarding as few of the relevant disorders as possible, seems an acceptable approximation. Hence, the notion of diagnosis that underlies diagnosis in hepatology is similar to the notion of refinement diagnosis as discussed in Chapter 5; it is unlike notions of diagnosis such as strong causality diagnosis.

Chapter 7

Development and Implementation of HEPAR

The major objective of the development of the HEPAR system was to investigate the hypothesis that techniques from the field of expert systems can be used for the development of an accurate diagnostic expert system in a field of the size of hepatology. In particular, the suitability of using empirical associations with associational diagnosis (or hypothetico-deductive diagnosis in a logical framework) as the principal notion of diagnosis was investigated. In this chapter, the principles underlying the HEPAR system, including its development, are discussed. In the next chapter, the evaluation of the system is addressed.

As discussed in Chapter 1, various systems have been developed to support the clinician in the diagnosis of disorders of the liver and biliary tract. Most of these systems were built to aid the clinician in the selection of diagnostic tests early in the diagnostic process. A rough classification of a patient's disorder into one of 2–4 different categories is often sufficient for that purpose. However, in order to assess the treatment and prognosis of a patient's disease, a more detailed classification in terms of disease entities is required. Hence, in contrast to other diagnostic (not only computer-based) systems in hepatology, the HEPAR system was developed to produce a detailed classification of a patient's disease. The HEPAR system may therefore be viewed as the point of departure for a decision-support system covering the entire process of patient management, not only diagnosis. The system discussed in this thesis, however, is limited to the support of the diagnostic process.

7.1 Diagnosis in hepatology

The diagnostic framework developed in the first part of this thesis provides a good basis for clarifying the notion of diagnosis that was considered suitable for diagnosis in the field of hepatology. The framework has been developed after the construction of HEPAR was completed. It has, therefore, not been applied in the design of the HEPAR system. In this section, the diagnostic framework shall be utilized to motivate particular choices made in the development of HEPAR.

As the term ‘defect’ is not customarily used in medicine to denote causes of malfunction, from now on, the term ‘defect’ will often be replaced by the term ‘disorder’. The distinction that is made in medicine between *categories* of disorders and *specific* disorders plays a key role in the organization of knowledge in the medical domain. The concept of ‘disorder category’ is also explored in guiding the process of diagnosis. In the previous chapter, the group of hepatocellular disorders was presented as an example of a disorder category, whereas acute hepatitis-B was an example of a specific disorder.

The concept of ‘disorder category’ was introduced in Chapter 3 (Definition 3.13). Formally described in terms of an evidence function e, a ‘disorder category’ was defined as the union of the observable findings associated with the disorders more specific than the category. If associational diagnosis is taken as a basis for diagnostic problem solving, a notion corresponding to the *common* findings, instead of the union of the findings, of a set of defects is more appropriate. A defect corresponding to the common findings of sets of defects will be called a classifier.

Definition 7.1 (*classifier*). *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. A defect $d \in \Delta$ is called a* more specific classifier *than a defect $d' \in \Delta$, denoted by $d \ll d'$, if $e(d') \subset e(d)$. If $D \subseteq \Delta$ is a set of defects, such that there exists a defect $d \in \Delta$ for which each $d' \in D$ is a more specific classifier and*

$$e(d) = \bigcap_{d' \in D} e(d')$$

then d is called a general classifier *of D.*

Focussing on findings that are common to a number of disorders is a frequently applied diagnostic strategy in medicine.

In the following example, we review several possible notions of diagnosis in the field of hepatology in terms of the framework.

Example 7.1. Consider the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where the following set of defects $\Delta = \Delta_P \cup \Delta_N$, standing for present and absent hepatological disorders, is given:

$$\begin{aligned}\Delta_P = \{&\textit{hepatocellular disorder},\\ &\textit{autoimmune chronic hepatitis},\\ &\textit{acute hepatitis-A}\}\end{aligned}$$

$$\begin{aligned}\Delta_N = \{&\neg\textit{hepatocellular disorder},\\ &\neg\textit{autoimmune chronic hepatitis},\\ &\neg\textit{acute hepatitis-A}\}\end{aligned}$$

Furthermore, the following set of findings $\Phi = \Phi_P \cup \Phi_N$ is defined:

$$\begin{aligned}\Phi_P = \{&\textit{subtropical journey},\\ &\textit{contact with jaundiced subjects},\\ &\textit{hypergammaglobulinaemia},\\ &\textit{butterfly erythaema},\end{aligned}$$

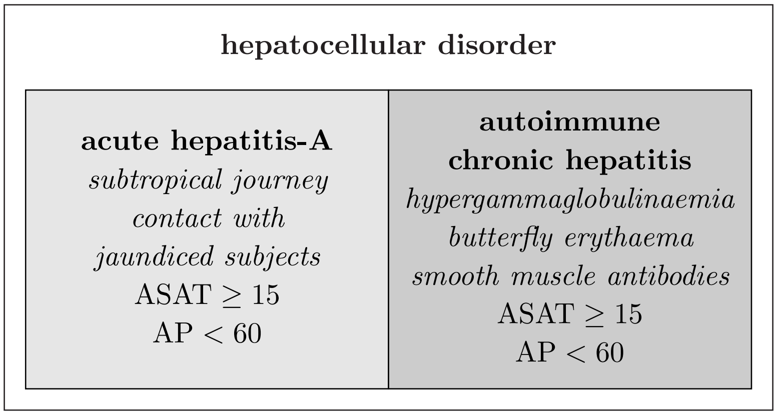


Figure 7.1: Partial specification of evidence function e.

$$\begin{aligned} &\textit{smooth muscle antibodies},\\ &\text{ASAT} < 15,\\ &\text{AP} < 60\} \end{aligned}$$

$$\begin{aligned} \Phi_N = \{&\neg\textit{subtropical journey},\\ &\neg\textit{contact with jaundiced subjects},\\ &\neg\textit{hypergammaglobulinaemia},\\ &\neg\textit{butterfly erythaema},\\ &\neg\textit{smooth muscle antibodies},\\ &\text{ASAT} \geq 15,\\ &\text{AP} \geq 60\} \end{aligned}$$

where '$\text{ASAT} \geq 15$' is used as an abbreviation for '$\neg(\text{ASAT} < 15)$'; a similar abbreviation is used for '$\neg(\text{AP} < 60)$'. Furthermore, it is assumed that findings with equality, such as '$\text{AP} = 75$', are allowed as observed findings.

Consider two different evidence functions, both expressing similar hepatological knowledge. Based on knowledge extracted from HEPAR, these evidence functions are defined in terms of a bottom-up partial specification. The first evidence function, e, is defined as follows:

$$\begin{aligned} \tilde{e}(\textit{acute hepatitis-A}) &= \{\textit{subtropical journey},\\ &\quad\ \textit{contact with jaundiced subjects},\\ &\quad\ \text{ASAT} \geq 15,\\ &\quad\ \text{AP} < 60\}\\ \tilde{e}(\textit{autoimmune chronic hepatitis}) &= \{\textit{hypergammaglobulinaemia},\\ &\quad\ \textit{butterfly erythaema},\\ &\quad\ \textit{smooth muscle antibodies},\\ &\quad\ \text{ASAT} \geq 15,\\ &\quad\ \text{AP} < 60\}\\ \tilde{e}(\textit{hepatocellular disorder}) &= \quad\ \tilde{e}(\textit{acute hepatitis-A}) \cup\\ &\quad\ \tilde{e}(\textit{autoimmune chronic hepatitis}) \end{aligned}$$

where $\tilde{e}(D) = \bot$ if there exists $d \in D$ with $d \in \Delta_N$, thus restricting to positive defects. The function e is depicted schematically in the Venn diagram in Figure 7.1. It represents

acute hepatitis-A *subtropical journey* *contact with* *jaundiced subjects*	**hepatocellular** **disorder** ASAT ≥ 15 AP < 60	**autoimmune** **chronic hepatitis** *hypergammaglobulinaemia* *butterfly erythaema* *smooth muscle antibodies*

Figure 7.2: Partial specification of evidence function e'.

empirical associations between disorders and findings, where the disorder '*hepatocellular disorder*' is a (disorder) category.

Next, consider the evidence function e', which is defined as follows (see Figure 7.2):

$$\begin{aligned}
\tilde{e}'(\textit{hepatocellular disorder}) &= \{\text{ASAT} \geq 15, \text{AP} < 60\} \\
\tilde{e}'(\textit{acute hepatitis-A}) &= \{\textit{subtropical journey}, \\
&\quad \textit{contact with jaundiced subjects}\} \cup \\
&\quad \tilde{e}'(\textit{hepatocellular disorder}) \\
\tilde{e}'(\textit{autoimmune chronic hepatitis}) &= \{\textit{hypergammaglobulinaemia}, \\
&\quad \textit{butterfly erythaema}, \\
&\quad \textit{smooth muscle antibodies}\} \cup \\
&\quad \tilde{e}'(\textit{hepatocellular disorder})
\end{aligned}$$

where $\tilde{e}'(D) = \perp$ if there exists a disorder $d \in D$ with $d \in \Delta_N$. This evidence function is part of the diagnostic specification $\Sigma' = (\Delta, \Phi, e')$. Note that '*hepatocellular disorder*' is a general classifier in this case.

Now, consider the diagnostic problems $\mathcal{P} = (\Sigma, E)$ and $\mathcal{P}' = (\Sigma', E)$, where

$$\begin{aligned}
E = \{&\neg \textit{hypergammaglobulinaemia}, \\
&\textit{contact with jaundiced subjects}, \\
&\textit{subtropical journey}, \\
&\text{ASAT} = 120\}
\end{aligned}$$

is the set of observed findings for a given patient. Based on the discussion in Section 6.8, a reasonable choice for a notion of diagnosis might be general intersection diagnosis, GI (cf. Definition 5.3). It states that the set of findings associated with each positive defect included in a diagnosis must have at least one finding in common with the set of observed findings. For this specific problem the following diagnostic solution is obtained:

$$\begin{aligned}
\text{GI}_{\Sigma, e_{|\Delta_P}}(E) = \{&\textit{hepatocellular disorder}, \\
&\textit{acute hepatitis-A}, \\
&\textit{autoimmune chronic hepatitis}\}
\end{aligned}$$

This diagnosis states that for this patient the hepatocellular disorders acute hepatitis-A and autoimmune chronic hepatitis are present. The same result is obtained by $\text{GI}_{\Sigma', e'_{|\Delta_P}}(E)$ for $\mathcal{P}'$. Note that if 'ASAT $= 120$' is deleted from E yielding E', the diagnosis

$$\text{GI}_{\Sigma', e'_{|\Delta_P}}(E') = \{\textit{acute hepatitis-A}\}$$

would result, where

$$\mathrm{GI}_{\Sigma,e_{|\Delta_P}}(E') = \{\textit{acute hepatitis-A}, \textit{hepatocellular disorder}\}$$

Hence, it seems that the evidence function e provides a more natural representation of the domain with regard to general intersection diagnosis than the evidence function e'.

There are other notions of diagnosis that do not seem less reasonable than GI. For example, given the diagnostic problem $\mathcal{P}'' = (\Sigma, E'')$, where

$$\begin{aligned} E'' = \{&\neg \textit{hypergammaglobulinaemia},\\ &\textit{contact with jaundiced subjects},\\ &\textit{subtropical journey},\\ &\mathrm{ASAT} = 120,\\ &\mathrm{AP} = 55\} \end{aligned}$$

a possible diagnostic solution is

$$\mathrm{GN}_{\Sigma,e_{|\Delta_P}}(E'') = \{\textit{acute hepatitis-A}\}$$

using the notion of diagnosis GN, which expresses that only those defects (disorders) $d \in \Delta_P$ are admitted to $\mathrm{GN}_{\Sigma,e_{|\Delta_P}}(E'')$ that have at least one finding from $e(d)$ in common with E'', and, in addition, the difference set $E''\backslash e(d)$ must not contain any finding f with $\neg f \in e(d)$. This yields a new notion of diagnosis: negative information acts as a kind of cancelling information. (This is similar to the consistency condition in abductive diagnosis; cf. Definition 2.3.) For the (second) evidence function e' from the diagnostic problem $\mathcal{P}''' = (\Sigma', E'')$, we have

$$\begin{aligned} \mathrm{GN}_{\Sigma',e'_{|\Delta_P}}(E'') = \{&\textit{acute hepatitis-A},\\ &\textit{hepatocellular disorder}\} \end{aligned}$$

Hence, the two evidence functions e and e' yield different results when interpreted by the same notion of diagnosis, although these evidence functions essentially express similar, although formally different, knowledge. This time, the evidence function e' seems more natural than e, but only with respect to the diagnostic interpretation of observed findings contradicting those predicted.

If we choose for the notion of general subset diagnosis, GS (cf. Definition 5.1), or associational diagnosis AD, because e and e' are monotonically increasing, the diagnostic solution $\mathrm{GS}_{\Sigma,e_{|\Delta_P}}(E'')$ expresses that every disorder admitted to $\mathrm{GS}_{\Sigma,e_{|\Delta_P}}(E'')$ must have its associated set of findings contained in E''; it is equal to

$$\mathrm{GS}_{\Sigma,e_{|\Delta_P}}(E'') = \{\textit{acute hepatitis-A}\}$$

which differs from $\mathrm{GS}_{\Sigma',e'_{|\Delta_P}}(E'')$, which is equal to

$$\begin{aligned} \mathrm{GS}_{\Sigma',e'_{|\Delta_P}}(E'') = \{&\textit{acute hepatitis-A},\\ &\textit{hepatocellular disorder}\} \end{aligned}$$

The latter diagnosis rightfully expresses that when acute hepatitis-A is diagnosed, the group of disorders to which it belongs is also included in the diagnosis. Again, the evidence function e' seems more natural than e. ◊

This extensive example illustrates that it is not an easy matter to decide on the notion of diagnosis that is most suitable to a particular problem. Clearly, as was already evident from the theory developed in the chapters 3–5, characteristics of the evidence function are important in this respect. In the development of the HEPAR system it appeared to be more convenient to represent groups of disorders as general classifiers rather than as disorder categories. The notion of diagnosis that has been chosen for the HEPAR system lies somewhere between the notion of general intersection diagnosis and the notion of general subset diagnosis. It may be characterized more precisely as the notion of structural associational diagnosis, SAD (cf. Definition 4.10). The main advantage of this notion is that a disorder is included in a diagnosis if at least one of the alternative sets of observable findings is included in the set of observed findings E. It does not have the disadvantage of general intersection diagnosis that it suffices for a disorder to have one finding in common with the set of observed findings to be included in a diagnosis. The notion is less rigorous than the notion of general subset diagnosis, but how much more flexible it is, depends on the specification of the structural evidence function. As argued in the previous chapter, in the field of hepatology it is not feasible to express information regarding simultaneously present (multiple) disorders. It was assumed that specific disorders are interaction free, but not strongly interaction free. For example, acute cholangitis is usually caused by cholelithiasis. This is represented in the knowledge base by expressing these disorder to have observable findings in common.

Summarized, the notion of diagnosis adopted in the HEPAR system can be best viewed as structural associational diagnosis, SAD, applied with respect to the set of positive disorders (defects) Δ_P and the set of negative disorders Δ_N as hypotheses. More formally,

$$\mathrm{SAD}_{\Sigma,e^s_{|\Delta_P}}(E) \cup \mathrm{SAD}_{\Sigma,e^s_{|\Delta_N}}(E)$$

yields a diagnosis in the sense of HEPAR for a diagnostic problem $\mathcal{P} = (\Sigma, E)$. Only a few disorders in HEPAR are represented by the evidence function in both positive and negative form, and none of the disorders is represented in negative form only. The knowledge concerning negative disorders expresses the fact that if one disorder is present, the other disorder is absent, i.e. $e(\neg d) \subseteq e(d')$, $d \neq d'$, holds (i.e. d' precludes d; cf. Section 3.1, preclusion). Disorders that are represented in both positive and negative form cannot be included both positively and negatively in a diagnosis, because the knowledge has been modelled in such a way that $e(d) \subseteq E$ and $e(\neg d) \subseteq E$ do not hold at the same time.

7.2 Structure of the HEPAR system

In the previous section, diagnosis in the field of hepatology was related to the diagnostic framework developed in chapters 3–5. In the present section, the knowledge incorporated in the HEPAR system, i.e. the set of disorders Δ and the set of observable findings Φ, is reviewed in detail; the associated evidence function e is merely sketched. The aim of the development of the HEPAR system has been described in Section 1.3.

7.2.1 Data used in the HEPAR system

Globally, the following data are included in the knowledge base of the HEPAR system, corresponding to the set of observable findings Φ of a diagnostic specification $\Sigma = (\Delta, \Phi, e)$:

- General data, such as age and sex of a patient;
- Data from a medical interview, personal and family history of a patient;
- Data from physical examination;
- Blood-chemical and serological data;
- Data from non-invasive diagnostic procedures, such as ultrasonography and radiography.

No use is made of data obtained by endoscopic retrograde cholangiography, percutaneous transhepatic cholangiography or liver biopsy. The results of these tests have been left out of the HEPAR system, as the system has primarily been designed for the initial assessment of a patient. (See chapters 1 and 6 for further motivation of this choice.) The findings from history and physical examination are always requested initially from the user. These findings are used for the early pruning of the search space of alternative hypotheses (see below).

There are three kinds of disorder in the system, that may be viewed as elements of various hypotheses $H \subseteq \Delta$ of a diagnostic specification Σ:

(1) The *type of the disorder*, which is *hepatocellular* and/or *biliary-obstructive*;

(2) The *nature of the disorder*, which is either *benign* or *malignant*;

(3) *Specific disorders*, which may be one or more of the 77 disorders listed in Table 7.1; a *final diagnosis* is a subset of the set of specific disorders.

A diagnosis by HEPAR consists of elements from these three sets. The 'type of the disorder' and 'nature of the disorder' are general classifiers of the specific disorders. As there are several specific disorders which may have the mixed features of hepatocellular derangements and biliary obstruction, it is possible to have both at the same time as conclusions for a particular patient. Hence, hepatocellular and biliary obstructive disorders are actually two different (general classifier) disorders that have several specific disorders in common. An example of such a disorder is primary biliary cirrhosis, a disorder that does affect both the small bile ducts and the hepatocytes. Other disorders, such as hepatitis-B, may have both features of a hepatocellular and a biliary obstructive disorder during at least part of their course. Although a patient may have symptoms and signs indicating both benign and malignant disease, only one of these can hold for a given patient case. Below, some examples of relationships between observable findings and disorders that have been represented, are discussed.

acute cholangitis	12	haematoma	2
acute cholecystitis	6	haemochromatosis	5 (6)
acute hepatitis-B	13	hepatitis-A	6
acute hepatitis-C	5 (6)	hereditary conj. hyperbilirubinaemia	1
acute hepatitis-E	0 (1)	hydatid cyst	3
alcoholic cirrhosis	15 (16)	infected polycystic disease	3
alcoholic hepatitis	17 (18)	infected solitary liver cyst	2
α_1-antitrypsin deficiency	3	infectious mononucleosis	3 (8)
amoebic liver abscess	4	invaded gallbladder carcinoma	4
amyloidosis	6	jaundice of pregnancy	1
autoimmune chronic hepatitis	5	liver cell adenoma	5
bacterial liver abscess	6	malaria	1
bifurcation carcinoma	4	malignant lymphoma	4
biliary dyskinesia	1	metastatic tumour	11
Budd Chiari syndrome	7	Mirizzi syndrome	3
carcinoma of papilla of Vater	4	nodular regenerative hyperplasia	4
Caroli's disease	5	other malignant liver tumours	2
cholelithiasis	1	pancreatic carcinoma	7 (8)
choledochal cyst	3	pancreatitis	6 (8)
choledochocele	1	polycystic disease	8
chronic hepatitis-B	16	portal lymphnode enlargement	4
chronic hepatitis-C	10	portal vein obstruction	2
circulatory liver damage	1	posttraumatic cyst	2
common bile duct carcinoma	9	primary biliary cirrhosis	14
common bile duct stone	15	primary hepatocellular tumour	15 (21)
congenital hepatic fibrosis	3	primary sclerosing cholangitis	10 (9)
congenital solitary liver cyst	5	regenerating nodule	1
Crigler-Najjar syndrome	1	Rendu Osler Weber disease	1
cryptogenic chronic hepatitis	5 (8)	retention cyst	2
cryptogenic cirrhosis	11 (7)	sarcoidosis	2
cystic liver metastases	2	secondary biliary cirrhosis	1
cytomegalic inclusion disease	1	sepsis	1
dermoid cyst	2	steatosis hepatis	6 (10)
duodenal carcinoma	2	toxaemia of pregnancy	1
early acquired syphilis	2	toxoplasmosis	3
fatty liver of pregnancy	1	veno-occlusive disease	4
focal nodular hyperplasia	6	Wilson's disease	4
Gilbert's syndrome	4	Zieve's syndrome	5
haemangioma	5		

Table 7.1: Disorders covered by the version of the HEPAR system that has been validated; each disorder is followed by the number of rules in which it appears; numbers between parentheses are numbers of rules in the most recent version of the system when different from the number of rules in the validated version.

7.2.2 Diagnostic strategy in HEPAR

Above, hepatological data relevant for diagnosis have been described, without referring to the way in which the data are used. In this section, some of the details of the diagnostic strategy incorporated into HEPAR are described. It has been modelled according to the diagnostic plan depicted in Figure 6.1. As discussed in Chapter 1, diagnostic problem solving in hepatology consists of a sequence of steps; during each step evidence is gathered in order to accept, reject or adjust a hypothesis. The data in the HEPAR system have been grouped in accordance with the traditional sequence of steps followed by the clinician in diagnostic problem solving. Part of the structure of the diagnostic process in hepatology has been modelled by imposing a static ordering on the data. The other part is a dynamic process, brought about by the inclusion of disorders as part of a hypothesis in the process of reasoning. More will be said about the dynamic nature of the diagnostic strategy in Section 7.3.

When the system is consulted for advice concerning a particular patient, only part of all patient data is required to reach a diagnostic conclusion. A diagnosis that accounts for the observed findings for a patient is determined dynamically. As discussed above, if the (dynamic aspects of) information gathering are disregarded, diagnosis can be viewed as structural associational diagnosis. The system first assesses a hypothesis, consisting of the following elements:

(1) Whether the patient is suffering from an acute, subacute or chronic disease;

(2) Whether the entered patient data have benign or malignant features;

(3) Whether the disorder is caused by a hepatocellular derangement, or a disorder of the small or large bile ducts (biliary obstructive disorder).

This constitutes the first step in the diagnostic process. Subsequently, the system produces a subset of all distinguished specific disorders as a differential diagnosis, based on the data entered, the three intermediate conclusions mentioned above, and additional information required in the remainder of the diagnostic process. In forming the differential diagnosis, the result for the intermediate hypothesis is applied as a means of cancelling most of the specific disorders from further consideration in the diagnostic process.

Prompting the user for additional information, i.e information gathering, is also controlled by the conclusions reached for the elements of the intermediate hypothesis. As a consequence, the system normally asks the user only about the serological test results for hepatitis-A, hepatitis-B, cytomegalic inclusion disease, toxoplasmosis, etcetera, when there is sufficient evidence for an acute, hepatocellular disorder. Data from ultrasonography of the liver and biliary tract are usually not required in that case, as all biliary obstructive disorders have already been rejected on the basis of available clinical evidence. On the other hand, if the system has gathered sufficient evidence for a malignant biliary obstructive disorder, the user is asked to enter ultrasonographical data from the liver and biliary tract. Serological data are then not required. The final output of the HEPAR system is a differential diagnosis, i.e. a list of specific disorders, of the patient's disease, ordered by the amount of evidence, expressed by a measure of plausibility (called certainty factor, see below). The derivation of a differential diagnosis is the last step in

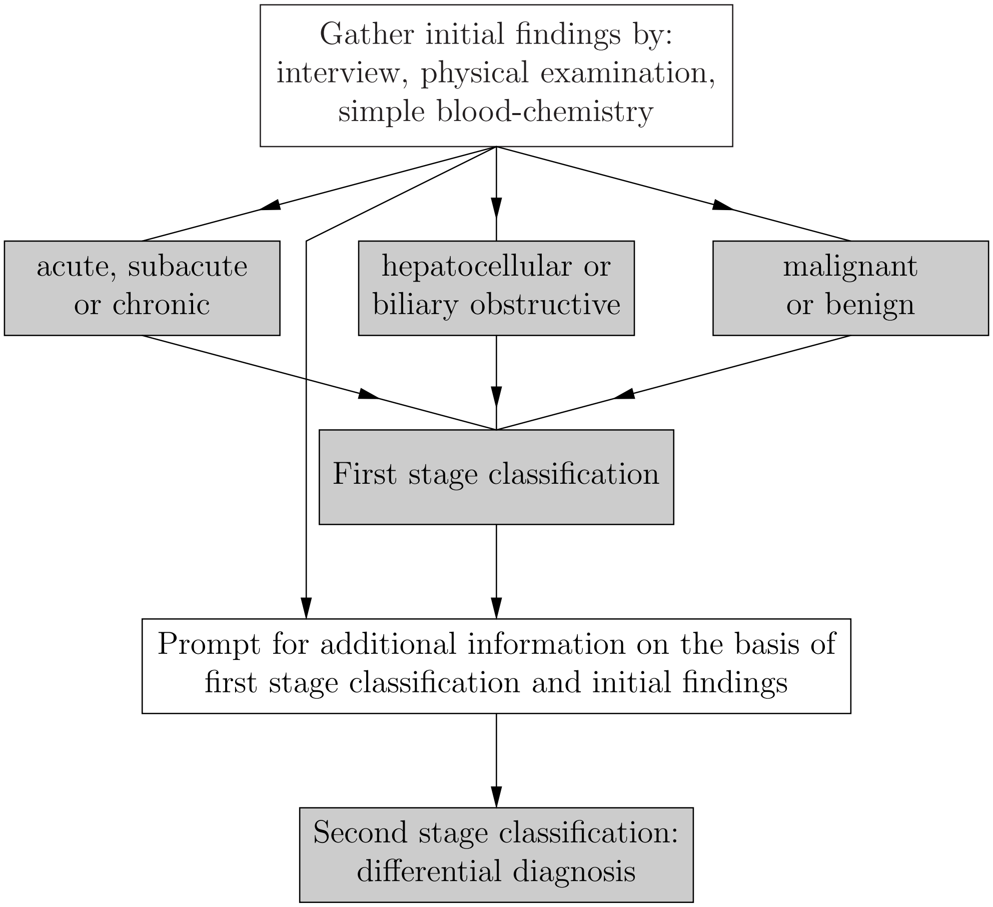

Gather initial findings by:
interview, physical examination,
simple blood-chemistry
acute, subacute
or chronic
hepatocellular or
biliary obstructive
malignant
or benign
First stage classification
Prompt for additional information on the basis of
first stage classification and initial findings
Second stage classification:
differential diagnosis

Figure 7.3: Diagnostic strategy in HEPAR.

the diagnostic process. Note that according to Table 7.1, diagnosis in HEPAR is not restricted to disorders of the liver and biliary tract; several diagnostic descriptions of disorders of the duodenum and pancreas have also been included in the knowledge base. Thus, a smooth transition of the field of hepatology to related areas is obtained. The structure of the diagnostic process is summarized in Figure 7.3.

It should be remarked that the system is not capable of establishing whether or not the patient suffers from single alternative or multiple disorders. As holds for refinement diagnosis, the purpose of the system is to establish a diagnosis consisting of disorders that together account for as many of the observed findings as possible. In principle, multiple disorder diagnoses can be generated, but then, more than the two hypotheses Δ_P and Δ_N must be investigated. The empirical, uncertain nature of most of the diagnostic knowledge embodied in the HEPAR system does make distinguishing between disorders as alternative and combined explanations of little value.

7.3 Implementation aspects of HEPAR

In this section, some details of the implementation of HEPAR are discussed.

7.3.1 Knowledge-representation and inference techniques

The HEPAR system has been implemented using the DELFI-2 system, [Lucas, 1986; Lucas & De Swaan Arons, 1987], a rule-based expert system shell that has much in common with the EMYCIN system (cf. [Lucas & Van der Gaag, 1991]). The DELFI-2 system provides a form of hypothetical reasoning, called top-down inference or backward chaining, that is suitable as a basis for implementing hypothetico-deductive (associational) diagnosis. The knowledge base of the HEPAR system consists of two separate components:

- A set of *object definitions*, where an object describes an entity from the domain by listing its relevant features, called *attributes*;

- A *rule base*, consisting of a collection of production rules, containing knowledge concerning diagnosis in hepatology.

For details concerning object and production-rule representation in rule-based (or production) systems, the reader is referred to [Lucas & Van der Gaag, 1991]. Some alternative representation formalisms, notably many-sorted logic, [Bezem, 1988; Lucas, 1993], and probabilistic belief networks, [Korver & Lucas, 1993], have been subjects of experiments with the HEPAR knowledge base. The rule-based version of the system, however, has remained the most important implementation. The organization of its knowledge base is briefly discussed.

The hepatological data are organized as objects with attributes and associated (sets of) possible values. For example, *patient* is an object with attributes *age*, *complaints*, *diagnosis*, etcetera. Most of the data have been placed in meaningful collections, as possible values of attributes. For example, all data concerning the symptoms of the patient, such as jaundice and abdominal pain, are defined as values of the (multi-valued) attribute *complaint*. The data described in Section 7.2.1 have been represented by means of 118 attributes, grouped within 11 objects. The organization of these objects in the form of a tree structure is shown in Figure 7.4, where only a small part of the attributes incorporated in HEPAR are displayed. Attributes that are defined as *goals* are explored by the system as hypotheses. Three such goal attributes have been defined for the *patient* object:

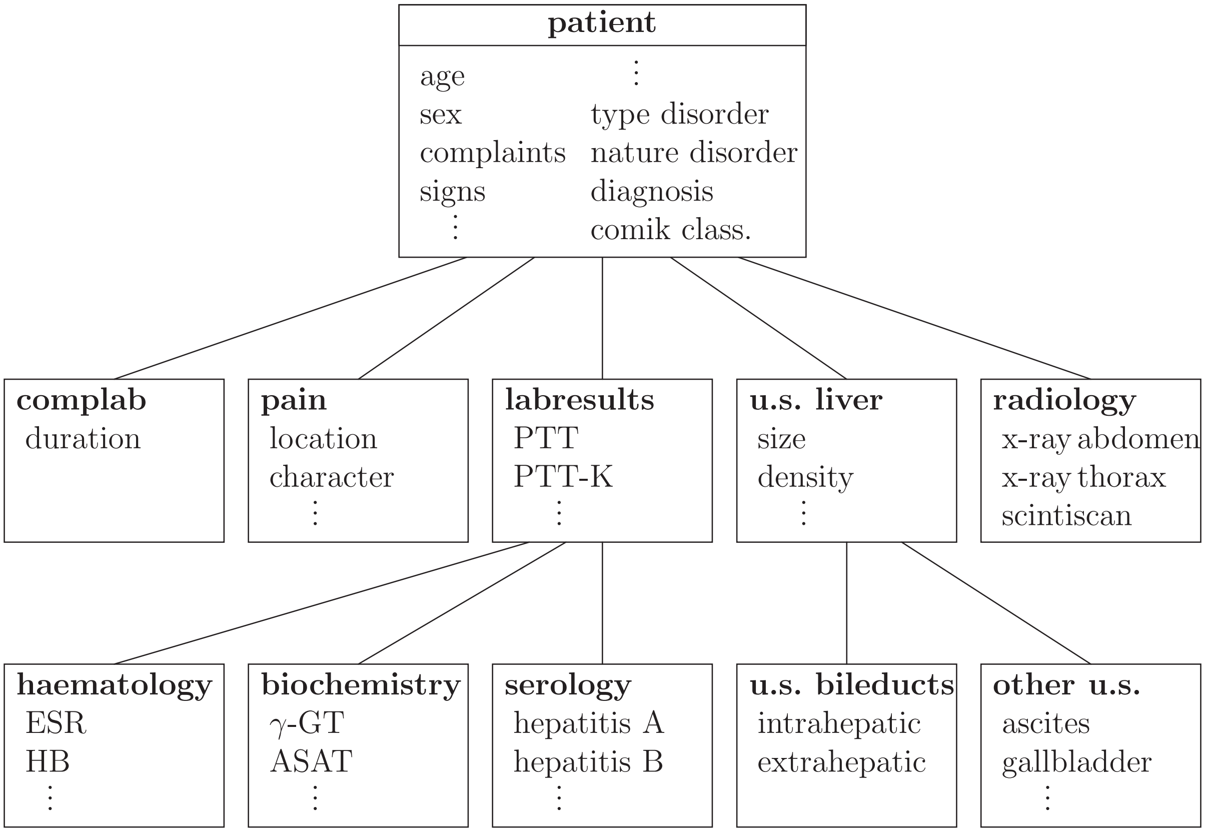
(1) The multi-valued attribute *type-disorder* with *hepatocellular* and *biliary-obstructive* as possible values;

(2) The single-valued attribute *nature-disorder* with *benign* or *malignant* as possible values;

(3) The multi-valued attribute *diagnosis* with the possible values listed in Table 7.1.

Note that the possible values of these attributes correspond to the set of disorders (defects) Δ in terms of our diagnostic framework.

The data from history and physical examination are represented as so-called initial attributes of the object *patient*; for these attributes values are initially requested from the user in the order of their specification in the HEPAR knowledge base. The specification order corresponds to medical practice. For example, the system first asks the user to



complab = complaints, clinical signs, lab abnormalities
u.s. = ultrasound

Figure 7.4: Objects in the HEPAR system.

enter a name of the patient (i.e. to enter a value for the attribute *name* of the object *patient*), next to enter the age, sex and complaints of the patient etcetera. After all the initial attributes of the object *patient* have been processed, the system proceeds to trace the three goal attributes defined in the object *patient*.

The main part of the knowledge base consists of production rules, mostly representing empirical associations. In the next section, the methods employed in designing these rules, is discussed. A *production rule* is an expression of the form

if A **then** C **fi**

where A represents a Boolean expressions consisting of tests on values of attributes of the form $P(o, a, v)$, called *conditions*, with o the name of an object, and a the name of an attribute; P is called a predicate that tests whether some relation P holds between entered or derived values of attribute a and the given value v. A disjunction of conditions is called a *clause*. The expression C denotes the sequential composition of *conclusions* concerning object–attribute pairs. In HEPAR, all conclusions include the action ‘conclude’; it expresses that a (new) value is to be added to the set of values already derived for a given attribute. Each value derived for an attribute is attached with a measure of uncertainty, called a *certainty factor*, in the sequel sometimes abbreviated to CF. Certainty factors take values between -1 and $+1$ inclusive, where a value of -1 means absolutely false and a value of $+1$ means absolutely true; a certainty factor equal to 0 expresses that the value to which it is associated is unknown. The certainty expressed by means of certainty factors is propagated from entered findings (or facts) to derived conclusion using a certainty calculus, referred to as the certainty-factor model [Buchanan & Shortliffe, 1984]. The reader is referred to [Lucas & Van der Gaag, 1991; Van der Gaag, 1989] for a detailed description of the underlying principles of the certainty-factor model and production systems; [Van der Gaag, 1994] discusses the pragmatics of these techniques. Certain factors only play a role in ordering the elements of diagnostic conclusions produced by the system.

7.3.2 Description of the rule base of HEPAR

In this section, a detailed account of the contents of the rule base of the HEPAR system is provided. Firstly, a characterization of the rule base is provided in terms of simple parameters, such as the number of production rules, the number of conditions and conclusions, etcetera. In the literature on expert systems, detailed descriptions of the contents of a knowledge base are seldom encountered. An exception to this is a rule-based system for the diagnosis of abdominal pain of gynaecological origin, developed by B.S. Todd and R. Stamper [Todd & Stamper, 1993]. However, their production rules represent causal knowledge instead of empirical associations. A straightforward comparison of the two systems is therefore impossible. Secondly, the distribution of the domain knowledge among the various production rules in the rule base is analysed. Several types of production rules will be distinguished to help the reader in understanding the system.

Table 7.2 gives an overview of several features of the HEPAR rule base. Column A enumerates the features for the entire knowledge base. As can be seen, the number of clauses in production rules varies between 1 and 20, with an average of 4.76 clauses per rule. On the average there is less than 1 (0.67) disjunctive condition per production rule; hence, most clauses are actually single conditions. The average number of conclusions per production rule in the rule base is about 2. Most production rules concern the object *patient*; this object occurs in 993 conclusions of rules. The average number of conclusions per object is, however, considerably lower: 168.5. Most of the production rules (493) contain the attribute *type-disorder*. These conclusions are approximately equally distributed among the values *hepatocellular* (249 conclusions, the maximum number of conclusions for any attribute value in the rule base), and *biliary-obstructive*. Most of the production rules that contain the attribute *type-disorder* have conditions testing whether the results of the routine biochemical tests, such as ASAT, ALAT, γ-GT, alkaline phosphatase (AP), etcetera, lie within normal range. An example of such a rule is given below.

Example 7.2. The following production rule, taken from the HEPAR knowledge base,

```
if
   greaterthan(biochemistry,AP,120) and
   between(biochemistry,gamma_GT,[50,100]) and
   lessthan(biochemistry,ASAT,15) and
   lessthan(biochemistry,ALAT,15) and
   lessequal(biochemistry,total_bili,17)
```

	A	**B**	**C**
Feature	(complete rule base)	(biochemical rules excluded)	(only diagnostic rules)
total number of rules	513 (533)	288 (307)	201 (208)
number of conjunctive clauses			
average	4.76	4.92	5.30
minimum	1	1	1
maximum	20	20	20
number of disjunctive conditions			
average	0.67	0.79	0.65
minimum	2	2	2
maximum	13	13	4
number of conclusions			
average	1.97	1.95	1.88
minimum	1	1	1
maximum	14	14	6
number of conclusions			
per object			
average	168.50	94.17	378
minimum	1	1	378
maximum	993	547	378
per attribute			
average	36.11	19.48	378
minimum	1	1	378
maximum	493	378	378
per value			
average	9.28	5.00	4.97
minimum	1	1	1
maximum	249	54	17
certanty factor			
average	0.43	0.50	0.50
minimum	-1.0	-1.0	-1.0
maximum	1.0	1.0	1.0

Table 7.2: Some features of the HEPAR rule base; numbers between parentheses are for the most recent version of the system.

```
then
  conclude(patient,type-disorder,biliary_obstructive)
  with CF = 0.5 also
  conclude(patient,type-disorder,hepatocellular)
  with CF = 0.25
fi
```

states that in a patient with a high alkaline phosphatase (AP) level in the serum, with slightly increased γ-glutamyl transferase (γ-GT), and with normal levels of aspartate and alanine aminotransferase (ASAT and ALAT), there is more evidence for a biliary obstructive than for a hepatocellular disorder. It is a typical example of a biochemical rule. ♢

The results of the routine biochemical tests used in HEPAR were originally formulated by the clinician in the form of a multidimensional table. This table was reformulated into production rules for the pragmatic reason that most of DELFI-2's facilities are production-rule related. Production rules, such as the rule in the example above, essentially translate knowledge concerning the relationship between abnormal ranges of laboratory tests and disorders into weighted evidence for the presence of a certain disorder for a patient. If a large number of different intervals are distinguished, production rules may not be the most natural formalism for representing such knowledge. Production rules are sufficiently expressive, but scattering knowledge over a large number of different production rules makes a rule base difficult to expand and maintain.

The attribute *diagnosis* occurs in 378 conclusions of production rules in HEPAR (see column C in Table 7.2). The average number of conclusions per attribute was 36.11, but the average number of conclusions for the attribute *diagnosis* was lower, about 2 varying between 1 and 6. The average number of conclusions for a specific disorder was about 5. In Table 7.1 the number of rules available for each specific disorder is shown. Hence, there is usually more than one production rule available to conclude about a specific hepatological disorder. For most of these disorders, at least one production rule has been included in HEPAR applying mainly clinical evidence to reach a diagnostic conclusion. These production rules may be viewed as the implementation counterparts of empirical, clinical associations. For most disorders, additional production rules are present using the results of specific laboratory tests (such as serological tests). Finally, to deal with cases in which particular data, such as data from serology or ultrasonography, are not available, special production rules have been included which explicitly handle incomplete information. The certainty factor associated with the conclusions of these rules are always lower than the certainty factors associated with the rules dealing with a complete set of patient data. Furthermore, the certainty factors attached to conclusions of production rules dealing with the results of specific tests are higher than those dealing with less specific clinical findings. The average certainty factor in the conclusions of the rules was about 0.5.

A further characterization of the rule base of HEPAR is obtained by distinguishing several types of production rules. For example, a distinction can be made between definitional, associational and strategic production rules. The different types of production rules will be illustrated by example rules, taken from the HEPAR knowledge base.

Definitional production rules define a certain concept as an object–attribute–value triple in the conclusion of a production rules. The features of a concept are stated as the conditions of the rule.

Example 7.3. The following HEPAR rule defines the concept of ‘Courvoisier’s sign’:

```
if
   notsame(patient,pain,colicky) and
   same(patient,signs,palpable_gallbladder) and
   same(patient,jaundice,yes)
then
   conclude(patient,Courvoisier,positive)
   with CF = 0.9
fi
```

It expresses that ‘Courvoisier’s sign’ is the observation of a palpable gallbladder in a jaundiced patient without colicky pain. ◊

It should be remarked that the definitional rule presented in Example 7.3, as all other definitional rules in HEPAR, deviates from a formal definition in two ways. Firstly, the logical analogue of this rule is not a bi-implication, as in a formal definition, but an implication. Secondly, the certainty factor occurring in the conclusion of the rule is not equal to 1. This indicates that, even if the findings have been observed in a patient, conclusions in a diagnostic situation may still be uncertain.

Associational production rules are used to associate disorders – either categories (general classifiers in HEPAR) or specific disorders – with certain empirical evidence. Most of the production rules in HEPAR are of the associational type.

Example 7.4. A typical example of an associational production rule is the following rule:

```
if
   greaterthan(patient,age,50) and
   same(patient,weightloss,significant) and
   same(patient,complaint,fever) and
   same(patient,complaint,generalized_pruritus)
then
   conclude(patient,type-disorder,biliary_obstructive)
   with CF = 0.40 also
   conclude(patient,type-disorder,hepatocellular)
   with CF = 0.25
fi
```

It states that a biliary obstructive disorder is more likely than a hepatocellular disorder in a patient older than 50 years, having a fever, pruritus and significant weight loss. ◊

The last type of production rule to be considered is the *strategic rule*; it is used to express a domain-specific reasoning strategy on top of the standard hypothetical reasoning obtained by backward chaining. Only a few of such strategic rules have been incorporated in the HEPAR system. The strategic reasoning behaviour discussed above, was realized by including conditions concerning disorder categories (general classifiers) in production rules concerning specific disorders.

Example 7.5. The first two conditions of the following production rule correspond to two general classifiers, viz. 'chronic duration' of the disorder and 'hepatocellular type' of the disorder.

```
if
  same(complab,duration,chronic) and
  same(patient,type-disorder,hepatocellular) and
  lessthan(biochemistry,alpha-globulin,2) and
  notsame(labresult,alpha-1-antitrypsin_phenotype,MM)
then
  conclude(patient,diagnosis,alpha-1-antitrypsin_deficiency)
  with CF = 0.7
fi
```

This expresses that α_1-antitrypsin deficiency is a chronic hepatocellular disorder. ◊

The order imposed on the specification of production rules affects the order in which individual disorders are explored.

Many researchers feel that when only a single layer of production rules is distinguished in an expert system, one should be careful in including strategic rules, since these rules disturb the declarative reading of the knowledge base, i.e. the meaning of the knowledge base as a logical theory [Jackson et al., 1990]. When the rule base contains a large number of strategic rules, it is usually advantageous to separate the rule base into two layers, a layer of declarative production rules, then called *object-rules*, and a layer of strategic production rules, then called *meta-rules* [Davis, 1980]. Because the number of strategic production rules in HEPAR is limited, such a separation has not been made. However, improving the structure of the system, e.g. in a way resembling CENTAUR [Aikins, 1980; Aikins, 1983], may be necessary when the system is expanded further.

7.4 Development of the HEPAR system

Above the structure, content and implementation of the HEPAR system have been reviewed. With this information as a basis, it is now possible to focus on the methods adopted in the development of the system (cf. [Lucas, 1994]).

7.4.1 Software engineering perspective

In the field of software engineering it is generally recognized that the implementation of large software systems must be supported by methods and tools for their verification and validation [Sommerville, 1992]. Often a distinction is made between static and dynamic approaches to verification and validation. Typical examples of *static* methods are program code inspection and methods for proving program correctness. The application of static methods does not require the program to be executed. In contrast, *dynamic* verification and validation methods involve execution of the program and examination of its output in relationship to specific input. As has been pointed out repeatedly by many researchers, dynamic methods can be used only to demonstrate the presence of errors in a program, and generally not to demonstrate their absence [Backhouse, 1986]. Despite this fundamental limitation, software engineers consider dynamic methods indispensable aids in the software-development process, and supporting software tools are invariably included in programming environments. It is therefore ironical that in the development of expert systems, where it is much more difficult to ensure that the system meets its specifications and expectations than in software engineering, tools that aid in the dynamic verification and validation of the system are not generally available.

In this section, several static and dynamic methods and associated software tools, which were applied during the development of the HEPAR system, are discussed. Taking the HEPAR system as a real-life example, the situation in which the need for static and dynamic verification and validation methods and tools in the development process was identified, will be described. This experience provides further evidence that more support by methods and tools for verification and validation is needed than is presently provided.

7.4.2 Knowledge acquisition and design

It is now well-recognized that the acquisition of domain knowledge in the process of building an expert system is a difficult task [Guida & Tasso, 1989]. In recent years, many methodologies have therefore been proposed, providing systematic methods to be followed in building an expert system. Examples of such methodologies are KADS [Breuker & Wielinga, 1989], CommonKADS [Schreiber et al., 1994] and KEATS [Motta et al., 1989; Motta et al., 1990]. Some of these methodologies include a set of software tools which help the knowledge engineer in building a specific application, mainly by assisting in the analysis of the problem domain. Some assistance by software tools may be provided in the design of the expert system as well. Examples of such tools are Shelley [Anjewierden et al., 1990], Acquist [Motta et al., 1989; Motta et al., 1990], KADStool [Albert & Jacques, 1993], KARL [Angele et al., 1994; Fensel, 1995] and DESIRE [Van Langevelde et al., 1993]. Most methodologies place considerable emphasis on the process of gathering domain knowledge to be incorporated into the expert system, and on the development of conceptual models of the domain, being the result of the analysis of the knowledge collected.

Although the HEPAR system was actually designed before such methodologies came into play, the development of the system was initially carried out in a structured way, mainly following the stepwise-refinement design methodology from software engineering [Dijkstra, 1972; Wirth, 1971]. The knowledge concerning diagnosis in liver and biliary disease incorporated into the HEPAR system was derived from the experience of a specialist in internal medicine and hepatology and from the medical literature. The analysis of the problem of diagnosis of disorders of the liver and biliary tract indicated that the following aspects were important in this domain:

- Expert hepatologists follow a clear and unambiguous strategy in diagnosis. The early classification of a patient's disorder into general categories, such as whether or not the disorder is biliary-obstructive in nature, is used for the selection of supplementary tests to reduce the number of alternative diagnoses to be considered.

- Early in the diagnostic process only a limited amount of patient data is available, mainly obtained from medical history and physical examination. Still, a hepatologist is often capable of stating a working diagnosis of the patient's disorder.

The process of knowledge acquisition was carried out in five main stages:

(1) The main concepts and their interrelationship in the process of medical diagnosis in hepatology were identified.

(2) The global strategy followed by clinicians in reaching a diagnosis in liver and biliary disease was determined.

(3) An inventory was made of all data required in the diagnostic process, grouped into several categories, such as biochemical data, serological data, ultrasound data, etcetera.

(4) Informal rules dealing with the concepts and the diagnostic strategy mentioned in steps (1)–(2), using the data mentioned in step (3), were drafted by the medical specialist.

(5) The knowledge communicated to the author by the medical specialist at stage (4), was subsequently formalized yielding object definitions and production rules. Stages (4) and (5) were repeated several times.

The development of the system was decomposed into several subtasks, according to the domain structure determined after the first three stages in the development. In the design of the HEPAR system, the problem-solving strategy followed by the hepatologist has been taken as the point of departure for problem decomposition of the diagnostic process, by distinguishing several subtasks [Lucas & Janssens, 1991a]. The reader is referred to Section 7.2.2, where the decomposition of the diagnostic problem solving task in HEPAR into subtasks is discussed.

The requirement that the ultimate system ought to be able to assist the clinician in the initial assessment of the patient, for whom only a limited amount of data is available, as well as in the assessment of a patient for whom more specific test results are known, proved to be extremely difficult. In general, to explicitly deal with data not available in the patient would yield an exponential number of combinations of conditions on known and unknown patient findings to be taken into account. Although the hepatologist involved in the project was able to reduce the number of useful combinations considerably, we felt quite uncertain with respect to the suitability of this knowledge for classifying actual patient cases.

Note that current knowledge-acquisition methodologies do not offer much help in solving this problem. In most popular methodologies, the design process is essentially viewed as the process of abstraction from reality. Our problem was that we required some form of experimental feedback in refining the expert system to accommodate to reality. Likewise, only limited attention has been given to tools that provide information about the diagnostic quality of the advice produced by the expert system.

```
if
  (same(complab,duration,chronic) or
   same(patient,type-disorder,biliary_obstructive)) and
  same(patient,sex,female) $\Rightarrow$ same(patient,sex,male) and
  same(serol,mitochondrial_Ab,yes)
then
  conclude(patient,diagnosis,primary_biliary_cirrhosis)
  with CF = 0.80 $\Rightarrow$ with CF = 0.60
fi
```

Figure 7.5: Rigorously formulated production rule.

7.4.3 Experimental feedback

After completing a considerable portion of the knowledge base, some experiments with the system using data from real patients were done to investigate whether the system was able to meet our expectations. The system was unable to provide acceptable advice in many cases. An analysis of the results of this initial experiment yielded the following reasons for the disappointing performance:

- Many rules were formulated too rigorously, such that these rules almost never applied to a patient with the given disease.

- Many rules were defined without explicitly mentioning the medical context in which they should hold. These rules frequently succeeded for patients for whom they had not been designed.

These problems may actually be taken as an indication of the knowledge clinicians draw upon in medical practice. Firstly, the knowledge of the clinician is partly based on the descriptions given in medical textbooks, in which there is little place for the description of atypical disease patterns, and partly on experience in the management of specific disorders. Rules which have been formulated too rigorously tend to describe the typical picture of the disease, and may assume the availability of an unrealistic amount of data for the patient. The production rule shown in Figure 7.5 is an example of a too rigorously formulated rule, since the condition left from the right, double arrow is only applicable to female patients. Typically, a patient having primary biliary cirrhosis is female, but the disorder is not limited to the female sex. A new production rule was therefore added to the knowledge base in which the expressions specified right from the arrows replaced the condition and CF specified left from the arrow. Secondly, the clinician has considerable experience with disorders frequently observed in clinical practice. However, these disorders carry a clinical context which the clinician may not be able to make explicit. Formalizing such knowledge may yield rules with a wider application than intended. As an example, consider the production rule depicted in Figure 7.6. This rule was originally formulated without the second condition (indicated by the right arrow); for the modified rule to be applicable, fever must be absent in the patient. So, in its original form it was too weakly

```
if
   same(patient,complaint,abdominal_pain) and
⇒notsame(patient,complaint,fever) and
   same(patient,signs,hepatomegaly) and
   same(pain,character,continuous) and
   same(ultrasound_liver,parenchyma,multiple_cysts) and
   same(patient,nature-disorder,benign)
then
   conclude(patient,diagnosis,polycystic_disease)
   with CF = 1.00
fi
```

Figure 7.6: Weakly formulated production rule.

formulated. This missing condition caused the original rule to interact with Caroli's disease, infected liver cysts and cystic liver metastases.

As said above, clinicians often have to base their early decisions on incomplete clinical evidence; likewise, expert systems must be able to deal with incomplete evidence as well. In designing and implementing the HEPAR system we have tried to explicitly handle incomplete patient data in the following two ways:

(1) By distinguishing several different conceptual levels in the diagnostic problem-solving process;

(2) By explicitly incorporating knowledge about unknown diagnostic test results for a patient into the knowledge base.

To deal with the first source of incompleteness of information, rules were drafted covering only the symptoms and signs of a disorder obtained early in the diagnostic process, whereas other rules were drafted covering only the results of supplementary tests obtained later in the diagnostic process. In this fashion, the knowledge base obtained a layered structure. The second source of incompleteness was dealt with by inspecting rules for conditions on data not always available in the patient. Some of these rules were used as a basis for new rules containing conditions concerning unknown data.

The reason for explicitly dealing with known and unknown laboratory test results was that in the hepatological patient, the clinician may, for example, only require data concerning alkaline phosphatase and not concerning γ-GT serum levels. (Note that the serum levels of these enzymes provide information about the severity of biliary obstruction; cf. Section 6.3.) Data about the alkaline phosphatase levels in the patient often provide sufficient information to the clinician. Furthermore, it is sometimes simply not possible to obtain all necessary laboratory data, or data from supplementary investigations such as ultrasonography. For example, in ultrasonography of the abdomen it is sometimes not possible to visualize the pancreas due to intestinal reflections. This is important information that must be taken into account in the diagnostic process.

As an example, consider the production rule from HEPAR shown in Figure 7.7. In this rule it is assumed that the usual routine laboratory tests have been carried out for a

```
if
   lessthan(biochemistry,AP,60) and
   greaterthan(biochemistry,gamma_GT,100) and
   greaterthan(biochemistry,ASAT,15) and
   lessthan(biochemistry,ALAT,15) and
   greaterthan(biochemistry,total_bilirubin,17) and
then
   conclude(patient,type-disorder,biliary_obstructive)
   with CF = 0.25 also
   conclude(patient,type-disorder,hepatocellular)
   with CF = 0.40
fi
```

Figure 7.7: Rule which uses routine laboratory test results.

patient suspected of a disorder of the liver or biliary tract. The routine laboratory tests are serum levels of alkaline phosphatase (AP), γ-GT, ASAT, ALAT and total bilirubin. As alkaline phosphatase, γ-GT gives an indication of the presence and severity of biliary obstruction; increased levels of ASAT and ALAT provide information about the presence and severity of hepatocellular damage (cf. Section 6.3). An increased serum level of total bilirubin is a general indication that there may be a disorder of the liver or biliary tract in the patient. In the rule shown in Figure 7.8, which is very similar to the rule in Figure 7.7, it is assumed that the clinician only requested information concerning serum levels of alkaline phosphatase, ASAT and total bilirubin. In this case only the serum level of alkaline phosphatase is used to assess biliary obstruction and the serum level of ASAT is used to assess the extent of hepatocellular damage. However, the knowledge that no data concerning γ-GT and ALAT is available is explicitly taken into account. Note that in the situation in which the rule shown in Figure 7.8 will succeed, the rule shown in Figure 7.7 will fail.

Mainly static verification and validation methods were employed at this stage; the early experiments were only carried out to validate our design rationale.

7.4.4 Parameters for knowledge base refinement

The process of refining an expert system may be viewed as the iterative process of validating, extending and adapting its knowledge base. In this section, dynamic validation is addressed. Since extension and adaptation are based on the results of validation, it is clearly important to decide on the parameters used for validating an expert system. The requirements imposed on the refinement of an expert system differ slightly from those imposed on final validation. Subtleties that escape final validation, because they cannot be expressed in simple figures, can be captured more easily in the process of refinement. Examples are the structure of the knowledge base and clarity of the reasoning process of the system. These are features that are clearly important in refining an expert system. Both the structure and reasoning strategy of the HEPAR system have been modified

```
if
   lessthan(biochemistry,AP,60) and
   (notknown(biochemistry,gamma_GT) or
    notknown(biochemistry,ALAT)) and
   between(biochemistry,ASAT,[15,60]) and
   between(biochemistry,total_bilirubin,[17,50])
then
   conclude(patient,type-disorder,biliary_obstructive)
   with CF = 0.25 also
   conclude(patient,type-disorder,hepatocellular)
   with CF = 0.40
fi
```

Figure 7.8: Rule taking unknown laboratory data into account.

several times. The global strategy mentioned in Section 7.2.2 remained the same during the development of the system. The hepatologist involved in the project has studied the reasoning process of HEPAR for a considerable number of patients. Refinement of the reasoning process of HEPAR turned out to be very time-consuming. As a consequence other hepatologists have been involved in the project on a limited scale. However, the results produced by HEPAR have been discussed with other hepatologists. The diagnostic performance of the system in terms of correctly or incorrectly diagnosed, or unclassified test cases provides the best basis for refinement of the accuracy of a system. Performance evaluation provides little direct information about the causes of good or poor performance. Inspection of the knowledge that had been used to draw conclusions for a particular test case can be used for this purpose. This view of performance evaluation is also able to cope with situations in which the expert system's conclusions are incorrect, but nevertheless acceptable in the light of the data available. This approach to performance evaluation is particularly valuable when applied in refining an expert system; it is less appropriate in the final validation of an expert system [Hilden & Habbema, 1990].

7.4.5 Standards for dynamic refinement

A diagnostic expert system like HEPAR may be validated by comparing its results to:

- Known clinical diagnoses of (patient) cases;
- The conclusions of some other, but similar decision-support system;
- The judgement of human experts in the field.

In all cases, there is a need for a test, or a combination of tests, that may be taken as a 'gold standard' for the comparison, although this is not as crucial as in the final validation of an expert system, because inspection of the knowledge base may provide additional information. By a gold standard, we mean a test with high accuracy that is taken as a reference point in an evaluation study. Examples of tests that are suitable as a gold standard in the diagnosis of disorders of the liver and biliary tract are the histological examination of liver biopsies, ERCP and surgical exploration. There is not a single test available in hepatology that may be employed as a gold standard in the entire domain, because diagnosis of hepatocellular disorders differs considerably from diagnosis of biliary obstruction.

Initially in the refinement, part of the data from more than 1000 patients obtained from the Danish COMIK group as a source for comparison were used. Originally, these data have been used in the development of the Copenhagen Pocket Chart, a paper chart based on the statistical technique of logistic regression, which may assist the clinician in the early assessment of a patient with jaundice [Matzen et al., 1984]. Unfortunately, this database was of limited value because only 23 disease categories were distinguished in this database, whereas in HEPAR, about 80 disorders of the liver and biliary tract are distinguished, and not all data required for HEPAR to derive final conclusions were included in the database. Therefore, the database was mainly useful for getting insight into the extent to which HEPAR was capable of dealing with incomplete patient data. During the further development of the system, a database with patient data from the Leiden University Hospital was gathered. It was applied as the main device for the refinement of the knowledge base.

The role of human experts in refinement with respect to structure and reasoning strategy has been discussed above. A human expert may also be asked to judge the conclusions drawn by an expert system, but this requires considerable knowledge of the content of the system.

Comparison of an expert system with some similar decision-support system is seldom straightforward, because of differences in required input and produced output among the systems. For the refinement of the HEPAR system, production rules were included that map diagnostic conclusions of HEPAR to the possible diagnostic conclusions of the Copenhagen Pocket Chart. The results produced by the Copenhagen Pocker Chart could thus be used as a simple means for rapidly finding patient cases which deserved further study with regard to the conclusions produced by HEPAR.

Taking the conclusions concerning the final diagnosis produced by the HEPAR system as a point of departure for refining the system was difficult, because the system's advice consists of more than one conclusion. This problem is dealt with in Section 8.1.2. However, due to the layered approach to diagnostic problem solving modelled in HEPAR, it is not only possible to compare specific disorders as a diagnosis with the clinical diagnosis confirmed for the patient, but also to check whether the patient's disorder has been classified in the right diagnostic category (e.g. hepatocellular disorder). This layered approach makes it less likely that the differential diagnosis produced by the system is, as a whole, unacceptable to the clinician.

Most of the information obtained by the dynamic refinement of the HEPAR knowledge base was automatically compiled by a collection of simple software tools which are discussed in the next section. Without the availability of these tools, dynamic refinement would have been too time-consuming to be practically feasible.

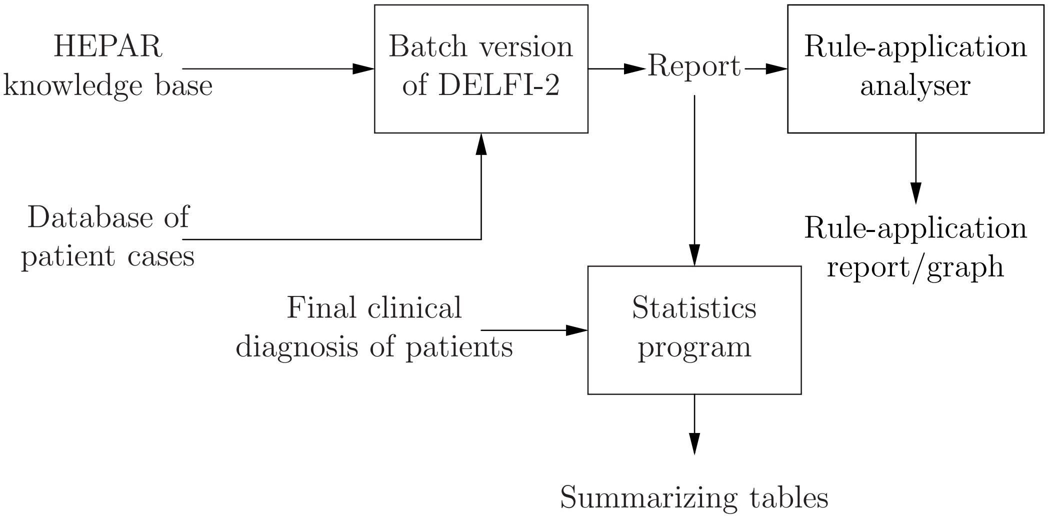


Figure 7.9: Environment for dynamic validation.

7.4.6 Tools for knowledge base refinement

Above, the approach to the development of HEPAR has been reviewed. In this section, several software tools that aided in the development of the system will be addressed.

Testing tools in software engineering

Some of the software tools that have been developed for the dynamic refinement of the HEPAR system, and also applied in the final performance studies of the system, have been inspired by software tools commonly available in programming environments. Two such tools are briefly described.

Test-data generators, programs that systematically produce test data to be used as input to the program to be tested, may be useful for the dynamic refinement and validation of expert systems. However, for realistic testing, data from real-life cases are often indispensable. Therefore, test-data generators have not used in the project.

Another tool is the *dynamic analyser*, also known as the execution flow summarizer. A dynamic analyser adds instrumentation statements to a computer program in order to collect information on how many times a statement is executed. A display part of the dynamic analyser prints a summarizing execution report [Sommerville, 1992]. A tool with similar usage as the dynamic analyser is the *call-graph profiler* [Graham et al., 1982]. These two tools inspired the construction of tools that will be discussed in the next section.

A tool for performance measurement

The environment of software tools developed for the dynamic validation of HEPAR consists of a non-interactive batch version of the expert-system shell DELFI-2 which is able to use a database of patient cases as its input. This system produces a report containing the results for each individual patient. The report together with a file containing information about the final clinical diagnoses and the two intermediate conclusions concerning the

Clinical Diagnosis	n
acute cholangitis	5
acute cholecystitis	1
acute hepatitis-B	2
alcoholic cirrhosis	13
alcoholic hepatitis	4
autoimmune chronic hepatitis	1
bifurcation carcinoma	3
chronic hepatitis-B	9
common bile duct carcinoma	2
common bile duct stone	11
cryptogenic chronic hepatitis	2
cryptogenic cirrhosis	2
early acquired syphilis	1
haemochromatosis	1
hepatitis-A	3
hydrops of gallbladder	1
metastatic tumour	1
pancreatic carcinoma	10
polycystic disease	4
portal lymphnode enlargement	1
primary biliary cirrhosis	7
primary sclerosing cholangitis	2
steatosis hepatis	6
Wilson's disease	1

Table 7.3: Clinical diagnoses of 82 patients of the Leiden University Hospital used as a learning population for HEPAR.

patient is then processed by a program which produces a table summarizing the results. Figure 7.9 shows the overall structure of the validation environment. The tools collect information on the number of correct, incorrect and unclassified patient cases concerning:

(1) The type of hepatobiliary disease (hepatocellular or biliary-obstructive);

(2) The nature of the disorder (benign or malignant);

(3) The final diagnosis.

For the final diagnosis, the system computes the average number of conclusions achieved, in which the clinical diagnosis corresponds to the conclusion highest ranked, or to one of the alternative disorders generated. An example of such a table, produced from data from 82 patients from Leiden University Hospital with clinical diagnosis listed in Table 7.3, is reproduced in Table 7.4. The table shows the results after refinement of the HEPAR system using the software tools. The specific disorders with the highest certainty factor concluded by the system were compared with the clinical diagnoses. The rules which had contributed to the diagnosis established by HEPAR for each patient were assessed. If needed, both the conditions and the certainty factors of the existing rules were adapted,

Conclusion	Correct n (%)	Incorrect n (%)	Unclassified n (%)	Total n (%)
Type of hepato-biliary disorde	74 (90)	4 (5)	4 (5)	82 (100)
Benign/malignant nature of disorder	78 (95)	4 (5)	0 (0)	82 (100)
Final diagnosis	71 (86)	8 (10)	3 (4)	82 (100)
Clinical diagnosis among conclusions	76 (93)	3 (4)	3 (4)	82 (100)

Table 7.4: Diagnostic results for the population of 82 patients with hepatobiliary disease from Leiden University Hospital.

Conclusion	Correct (%)			Incorrect (%)			Unclassified (%)		
	A	B	C	A	B	C	A	B	C
Type of hepato-biliary derangement	90	90	50	5	5	0	5	5	50
Benign or malignant nature of disorder	95	95	94	5	5	6	0	0	0
Final diagnosis	86	49	24	10	9	10	4	43	66

A: All available data presented to system.
B: Only data concerning symptoms, signs, haematology and bloodchemistry (no data from ultrasound or serology presented).
C: Only data from medical interview and physical examination.

Table 7.5: Assessment of the effects of incompleteness of information on the diagnostic conclusions of HEPAR, for the database of 82 patients with hepatobiliary disease from Leiden University Hospital.

and new rules were drafted. This approach proved to be very helpful in detecting and correcting gaps and inconsistencies in the knowledge base. The knowledge base was only modified when considered necessary from a medical point of view, in view of our aim of using only data from simple non-invasive diagnostic procedures. The reader should note that the system does not reach 100% correctness. The incorrect and unclassified cases concerned patients for whom the diagnosis was almost entirely based on the results of diagnostic procedures not applied in the HEPAR system, such as ERCP. In these cases, the specialist considered the conclusions, although incorrect or missing, as acceptable in the light of the data available.

To obtain insight into the capabilities of the system in handling incomplete data, the batch version of the expert-system shell selects part of the patient data from the database; for example, only data obtained from history and physical examination. This part of the analysis of the HEPAR system consisted of the following steps. First, the HEPAR system was supplied with the complete set of patient data available for each patient in the database. Next, data from ultrasonography, radiology and serology were left out. Finally, only the data obtained by the medical interview and physical examination were entered

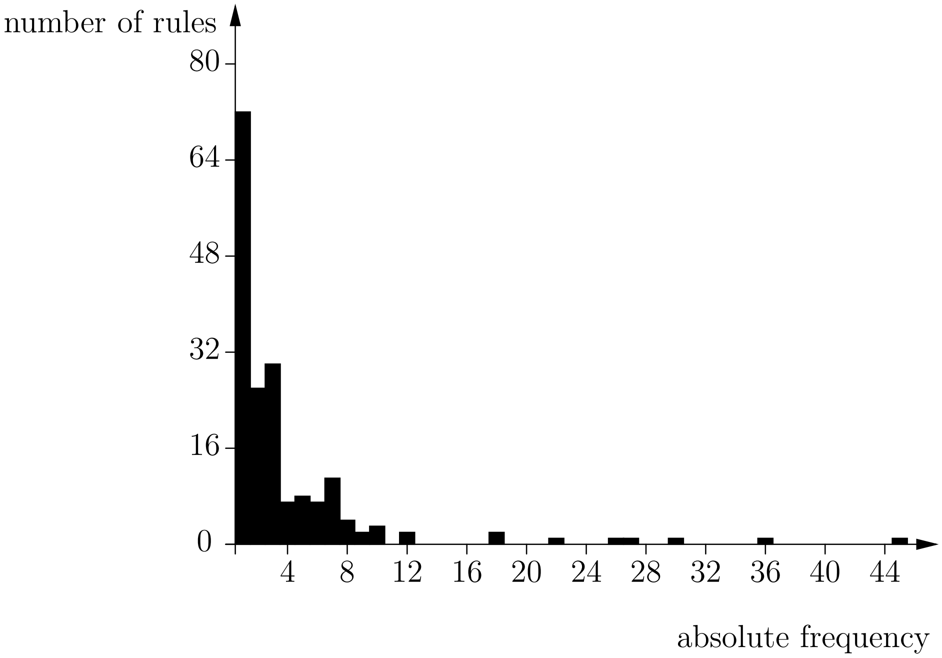


Figure 7.10: Rule-application bar graph for 82 patients from Leiden.

for each case. The final results, after the refinement of the system, are given in Table 7.5. As can be seen in Table 7.5, the number of correct final diagnoses decreased considerably when the patient data entered into the system were more incomplete. However, the percentage of incorrectly classified cases did not increase significantly, only the percentage of unclassified cases did. The remainder of the results in Table 7.5 can be understood in the light of the diagnostic strategy discussed in Section 7.2.2.

Tables such as those presented, were used by the hepatologist as an indication of the effects of changes to the knowledge base. An accompanying textual report provided information for each individual patient, and also provided information about the rules used for deriving the final conclusions. This information served as a point of departure for a more in-depth study of the reasoning behaviour of the system.

Dynamic analysis of the HEPAR knowledge base

In the previous section, we have discussed how the study of the results of the HEPAR system for individual patient cases, has been employed to refine the diagnostic quality of HEPAR. A second source of information that has been used for the refinement was the overall behaviour of the system when provided with a database with patient data. In order to obtain information concerning the frequency of rule application over a given database of patient cases, the testing environment discussed above was extended by a collection of tools, which use the report produced by the batch version of the expert system as input. These tools bear some resemblance to a dynamic analyser as described above. The following results are produced by these programs:

- An enumeration of all production rules used, with, for each rule, information about how often it has been used for a given database;

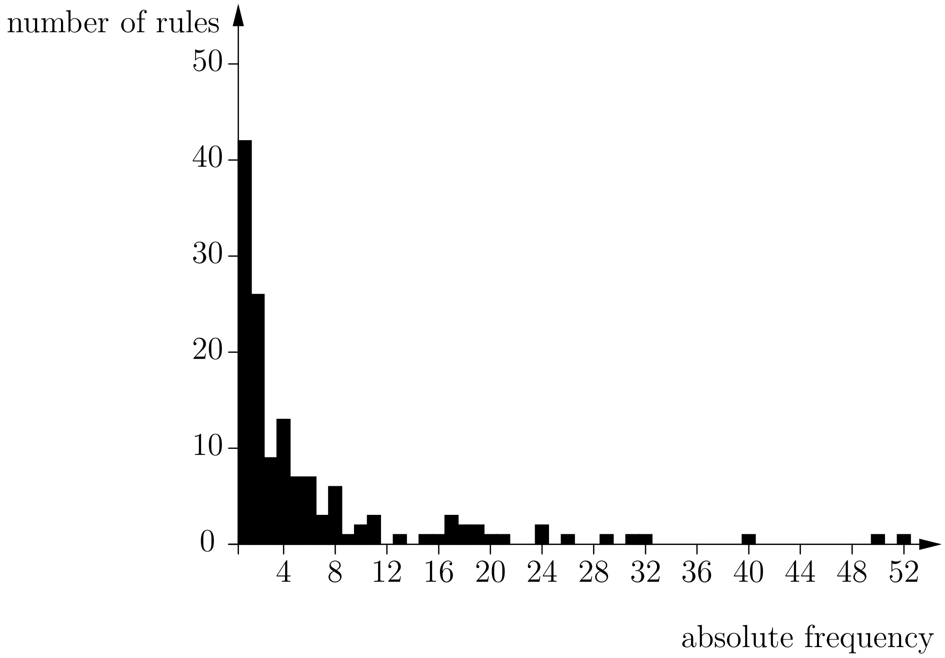


Figure 7.11: Rule-application bar graph for 94 Dijkzigt patient data.

- An overview of the frequency distribution of the rule application, both in textual and in graphical form.

Figure 7.10, automatically produced by the environment, contains the results after refinement of the HEPAR knowledge base for the 82 patient cases from Leiden University Hospital. Most production rules (72 of about 500 rules contained in HEPAR) were applied only once. A similar bar graph is shown in Figure 7.11 for a database with 94 patient from Dijkzigt University Hospital at Rotterdam. The accompanying textual form, which is reproduced in Figure 7.12, shows that from the rules that were applied several times, those with highest frequency were applied to conclude about the intermediate general classifiers. Only a few, one to three, production rules were applied several times to reach a conclusion concerning a specific disorder. The report does not include results for failed rules. The reports were studied by the hepatologist involved in the project as another source for the refinement of HEPAR.

7.5 Discussion

Above we have discussed the methods and tools applied in the development of the HEPAR system. The notion of diagnosis applied in HEPAR assumes that most of the knowledge concerns empirical associations. Hepatology turned out to be a field in which sufficient empirical knowledge was available to be used as a basis for an expert system covering a large part of the field of hepatology. Our initial choice to focus on empirical knowledge dictated the application of the notion of associational diagnosis. The implementation of the acquired knowledge in the form of object declarations and production rules using the DELFI-2 system was rather straightforward.

```
FREQUENCY   #RULES          RULE NUMBER
    1         72    161, 500, 510, 660, 700, 720, 790, 800, 860,
                    900, 930, 950, 980, 1100, 1140, 1150, 1330,
                    1340, 1370, 1410, 1480, 1540, 1610, 1630, 1710,
                    1730, 1820, 1980, 2000, 2010, 2030, 2050, 2140,
                    2240, 2250, 2380, 2530, 2680, 2690, 2750, 3011,
                    3020, 3022, 3090, 3100, 3101, 3120, 3160, 3171,
                    3192, 3220, 3340, 3370, 3425, 3430, 3450, 3460,
                    3470, 3520, 3530, 3610, 3780, 3820, 3870, 3902,
                    3930, 3960, 4000, 4111, 4122, 4162, 4260
    2         26    110, 120, 130, 440, 460, 470, 740, 830, 1090,
                    1130, 1490, 1940, 1950, 2130, 2180, 2370, 3010,
                    3041, 3110, 3390, 3410, 3427, 3428, 3440, 3920,
                    3940
    3         30    111, 112, 131, 132, 140, 190, 230, 240, 560, 580,
                    600, 620, 1220, 1900, 1970, 2060, 2080, 2110,
                    2170, 2220, 2280, 2310, 2720, 3190, 3350, 3970,
                    4230, 4240, 4250, 4252
    4          7    260, 341, 890, 1580, 2160, 3080, 4320
    5          8    1550, 1560, 2120, 2350, 3001, 3060, 3102, 3420
    6          7    160, 300, 2330, 3050, 3070, 3900, 3903
    7         11    90, 100, 332, 1210, 1570, 2200, 2320, 2340, 2970,
                    3000, 3910
    8          4    270, 290, 340, 350
    9          2    150, 180
   10          3    250, 4080, 4310
   12          2    540, 3103
   18          2    220, 4020
   22          1    370
   26          1    5010
   27          1    5000
   30          1    5030
   36          1    711
   45          1    5020
```

Figure 7.12: Textual rule application form.

As we have discussed, iterative refinement methods were essential in the development of HEPAR. Recent knowledge-engineering methodologies place considerable emphasis on the development of conceptual models [Motta et al., 1990; Wielinga et al., 1992]. A suitable conceptual model may be of real help in designing and implementing an expert system, as was also observed in the development of the HEPAR system, where a diagnostic problem-solving strategy was used as the basis for problem decomposition and structuring of the knowledge base. However, in many fields of medicine, the development of an expert system is only possible with sufficient experimental feedback, for which software tools are required. Other software tools that support the building of rule-based expert systems have been developed in the past. TEIRESIAS was an experimental tool that assisted in the refinement of rule-based expert systems by interacting with the user in the analysis of the conclusions concerning single cases, applying meta-knowledge about the rule base [Davis & Lenat, 1982]. Although such an analysis is certainly useful, an approach as embodied by TEIRESIAS does not provide information about how well the system performs for a database of cases. SEEK is a system that automatically suggests generalizations and specialization of production rules, based on the analysis of the success and failure of rules on processing case data [Politakis, 1985]. This system is more in line with our approach. The broadness of the domain of hepatology, and the amount of patient data incorporated in the HEPAR system, suggest that automatic techniques, as provided by SEEK, are as yet not powerful enough to be applied for refining a system like HEPAR. The parameters used for the refinement of the HEPAR knowledge base are only a few of the many that are possible. Another elegant example of knowledge-base refinement has been proposed by Adlassnig and Scheithauer in the context of the CADIAG-2/PANCREAS system [Adlassnig & Scheithauer, 1989]. They have used ROC curves for the optimal adjustment of the internal classification threshold in this expert system. However, the technique is not applicable to the HEPAR system, because here failure of classification is the result of logical falsification, and not of the failure of reaching an internal threshold.

Most current expert-system shells and expert-system builder tools do not provide facilities that support the refinement of an expert system by dynamic validation. In the development of the HEPAR system, we have therefore developed a collection of simple software tools that provide useful information for the refinement of the system. These tools have also been used in two successive validation studies of HEPAR [Lucas et al., 1989; Lucas & Janssens, 1991a; Lucas & Janssens, 1991b], which will be discussed in the next chapter. Although there are many ways in which these tools can be improved, these validation studies would not have been possible without the assistance of these tools. In our opinion, future software tools for building expert systems should provide a wider range of facilities for the detailed analysis, verification and validation of an expert system than is currently provided.

Chapter 8

Validation of the HEPAR System

In the previous chapter, the internal structure, inference strategy and contents of the HEPAR expert system have been described. The approach and techniques followed in the development of the system have also been discussed. The present chapter focusses on the evaluation of HEPAR.

Evaluation of an expert system actually concerns two separate activities. Firstly, it is concerned with its verification to ensure that the expert system meets its specifications. Secondly, it deals with validation in order to investigate the amount of agreement between the behaviour of the system and the problem-solving activity that it intends to model. In verifying an expert system, the subject of investigation is its structure, logical consistency and the soundness and completeness of the inference methods employed. Validation of an expert system encompasses a range of rather diverse activities, such as measuring the level of performance of its advice relative to some gold standard, and determining user acceptance, efficacy and cost effectiveness [Wyatt & Spiegelhalter, 1990]. The results of a validation study may be compared to a set of requirements aimed at, if available. As has been stated succinctly, verification of an expert system is concerned with "building the system right", whereas validation is concerned with "building the right system" [Boehm, 1979]. Research concerning the logical consistency of HEPAR is described in [Bezem, 1988; Lucas, 1993]. We shall not further deal with the subject of verification in this thesis.

Our description of expert systems in Chapter 1 explicitly mentioned their capability to produce advice at a level comparable to experts in the field. Yet, it is often unknown how good the diagnostic performance of a medical expert system is, because only few expert systems have been submitted to a sufficiently rigorous validation process. Although validation of an expert system is known to be difficult, it is of great importance, in particular in the medical field. As the ethical basis of medicine is to strive to improve the patient's health while attempting to do 'no harm' and to use limited health-care resources wisely, it is clearly important to determine whether an expert system is effective and accurate, and to find out how it may change usage of resources before it is widely distributed [Wyatt & Spiegelhalter, 1990].

This chapter begins with a brief overview of several issues that must be considered before starting with the validation of an expert system. Next, two successive validation studies of the diagnostic performance of the HEPAR system are discussed in depth. The chapter is concluded by a critical comparison of what has been achieved in our study to other, similar research.

8.1 Validation of expert systems

In this section, the various issue that are relevant for validation are reviewed. The description focusses on validation of diagnostic systems.

8.1.1 Features to validate

As mentioned above, many features of an expert system may be validated. In this chapter, we concentrate on the level of performance of an expert system, i.e. the assessment of the capability of the system to produce correct or acceptable advice. Advice is defined to be a collection of conclusions, comprising final conclusions and also intermediate results. It should be noted that this assumption restricts the scope of validation considerably. [Kulikowski & Weiss, 1982] have observed that diagnostic expert systems reason by constructing models of patients and compare these models with prestored descriptions of disease patterns in the knowledge base. Under ideal circumstances, validation of a diagnostic expert system should therefore not only pertain to the advice produced by the expert system, but also to the assumptions made by the reasoning process on which the conclusions are based [O'Keefe et al., 1987]. It is, for instance, important to know whether it is possible for an expert system to derive correct conclusions, based on incorrect assumptions. Hence, the conclusions of a medical diagnostic expert system should not be interpreted as unique statements, but as judgements of the patient's status. As a consequence, not unique statements but judgements should be validated. This view on validation is also able to handle situations in which the expert system's conclusions are incorrect for a number of test cases, but nevertheless acceptable in the light of the data available. The assessment of an expert system's capabilities is therefore grossly simplified if the results of validation are described in simple terms of 'correct', 'incorrect' and 'unclassified'. Unfortunately, in-depth validation of an expert system is often unfeasible, because of the considerable amount of time and money involved. In addition, in-depth analysis of the performance of an expert system requires considerable knowledge of the problem domain, which a developer is not likely to have.

8.1.2 Approach to validation

As discussed in Section 7.4.5, insight into the diagnostic performance of a diagnostic expert system can be gained by comparing its conclusions to known conclusions associated with cases in a database, a similar decision-support system or the conclusions expressed by human experts in the field. The validation of an expert system according to one or more of these methods may be viewed as *laboratory validation*, i.e. validation outside the real-life context in which the expert system is finally intended to operate.

A frequently applied method for validation of an expert system is to employ a database with data from cases, such as patients in medicine, for which a final conclusion is known.

To control bias, validation should always be done in a (double) blinded fashion, where the team involved in the development of the expert system is unfamiliar with the specific characteristics of the test population of cases, and the team responsible for data collection is unfamiliar with the precise contents of the knowledge base of the expert system.

A significant disadvantage of only comparing the results of an expert system with known results is that it does not yield information concerning the performance of experienced and less experienced field experts, when presented with the same information as the expert system. It is, of course, unfair to expect an expert system to perform well in situations in which even experienced experts perform poorly. Information concerning the performance of experts is therefore required. This information may be available in the literature, as was true, for example, for the problem of diagnosis of disorders of the liver and biliary tract (cf. Chapter 1). It may also be determined experimentally. An example of a limited study in which measurements of the diagnostic performance of human experts were incorporated in the study design is described in [Van Daalen, 1993].

Often only data from patients for whom a clinical diagnosis is known, are used in validation of a medical diagnostic expert system. It is often better to extend the validation study to include data concerning healthy subjects. The advantage of the latter approach is that useful information is obtained about the diagnostic sensitivity and specificity of the expert system, i.e. how well the system is capable of establishing the presence of disorders in the situation in which the disorders are known to be present (the sensitivity), and how well the system is capable of establishing the absence of disorders in the situation in which the disorders are known to be absent (the specificity) [Wulff, 1981]. Sensitivity and specificity are measures that are typically determined in probabilistic systems.

Validation may also be achieved by comparing an expert system to another decision-support system; two options exist:

- The decision-support system may be used as the gold standard, or
- Both the expert system and the other decision-support system can be validated by comparing with known clinical diagnoses.

Usually the last approach is adopted.

The last approach to validation, i.e. comparing the conclusions of the expert system to those of human expert clinicians, implies that each of the members of a group of clinicians is asked to formulate a differential diagnosis when provided with the same patient data as the system. Again a test or a combination of tests is required to be taken as the gold standard. It is also possible to use the group of experts as the gold standard. In this type of study, known as a *cross-validation study*, each of the experts providing diagnostic conclusions for a patient is asked to assess the conclusions of the expert system and those of the other human experts as well. This approach has the additional advantage that the inter-observer variability can be determined, and therefore taken into account in judging the performance of the system. In the past, cross-validation studies have been carried out on a small scale, and always with a small number of patient cases. The validation of MYCIN and ONCOCIN are examples of such cross-validation studies [Yu et al., 1979a; Yu et al., 1979b; Hickam et al., 1985]. The amount of effort involved in a cross-validation study is so large that it is unlikely that it will ever be done on a large scale, unless this kind of research is incorporated in daily, clinical practice. This view is confirmed by the intensionally limited scope of a cross-validation study regarding the performance of the PLEXUS expert system, [Van Daalen, 1993], a recent expert system for diagnostic and treatment advice regarding lesions of the brachial plexus [Jaspers, 1990]. In this study, the data collected were incomplete, a problem that occurred even for the small number of patient cases submitted for clinical judgement to four expert clinicians.

Instead of determining the opinion of the experts using the cross-validation technique, the *Delphi procedure*, developed in the early 1960s at the RAND corporation to obtain consensus from a panel of experts, can be used [Shannon, 1975]. Each member of a panel of experts gives an opinion on some subject, and is then confronted with a summary of the opinions of the other experts in terms of averages and extremes. The experts are subsequently asked to re-assess the problem now the opinions of the other experts are known. In the entire procedure, anonymity between panel members is maintained. Ideally, the iteration terminates when the opinions of the experts have converged to a common opinion. In practice, complete convergence to a single opinion seldom takes place. In the context of expert systems, validation can be accomplished by taking the expert's opinion about the advice of the expert system as the subject of the procedure.

An expert system can also be submitted to a field test. In a *field test*, each advice produced by the system is reviewed by a team of experts, while the system is being used in the environment for which it was developed. However, many medical applications are critical in nature; the validation of an insufficiently accurate expert system may therefore interfere with normal clinical procedures, making a field test difficult to implement. It is sometimes possible to use an expert system, or some other decision-support system, in parallel with the normal clinical procedures. This has been done with the probabilistic system of De Dombal which is used as an assistance in the diagnosis of acute abdominal pain [De Dombal et al., 1972; De Dombal, 1984; De Dombal et al., 1991].

Validation of the conclusions of a medical diagnostic system can be particularly difficult if the system's advice consists of more than one conclusion. This problem is known as the *multiple response problem* [Shannon, 1975]. Consider for example the situation in which an expert system produces the correct answer with highest certainty, as well as a conclusion with slightly lower certainty which is totally unacceptable to the physician. Restricting validation to only the single conclusion with the highest certainty gives a distorted account of the actual performance of the system. Consider the situation in which the conclusion with highest certainty is incorrect, but the differential diagnosis as a whole is acceptable to the physician. Taking only the conclusion with highest certainty into account gives an inadequate impression of the system's capabilities. If only a single clinical diagnosis is known for a patient, this diagnosis can be compared with the conclusion with highest certainty, or to all conclusions produced by the expert system. However, if the clinical assessment of the patient includes a differential diagnosis, the multiple response problem arises. The multiple response problem can be solved in several ways. The most obvious solution is to reduce the multivariate response to a univariate response by assessing the conclusions of the expert system, using a composite measure, such as acceptability (using, for example, terms such as 'acceptable' and 'unacceptable'). There are also multivariate techniques available for the paired comparison of multiple responses, which will be discussed below.

8.1.3 Performance measurement of diagnosis

Most measures to describe the performance of a diagnostic system derive from early work on the evaluation of probabilistic diagnostic systems [Habbema et al., 1978]. These measures cannot be used directly for the measurement of the performance of diagnostic expert systems, because of differences between the output produced by probabilistic diagnostic systems and expert systems based on (symbolic) reasoning methods. Probabilistic systems are always capable of producing a diagnosis – except when a probability threshold is used, because then cases may also be unclassified – even when no (patient) findings are available. If no findings are available, the system typically generates the prior probability of a disorder in the population. Using a threshold value, this probability may be interpreted as expressing a disorder to be absent. However, the situation for expert systems, in particular rule-based expert systems, is different. Typically, a rule-based expert system can produce no diagnosis for situations for which no findings are available. This is, of course, not a consequence of the rule-based representation, but of the notion of associational diagnosis underlying many, but not all, diagnostic rule-based systems. Examples of exceptions are the PROSPECTOR system (cf. [Duda et al., 1979]) and a diagnostic system for the diagnosis of abdominal pain of gynaecological origin developed by Todd and Stamper [Todd & Stamper, 1993]. Both systems are rule-based, but use probability theory as the underlying mechanism for reasoning with uncertainty. If associational diagnosis is employed, a diagnosis produced by an expert system expresses that a disorder is either present, absent or unknown, whereas a diagnosis produced by a probabilistic system expresses that a disorder is either present or absent. As a consequence, the measures developed for probabilistic systems need some slight adaptation for use in the performance validation of a system based on the notion of associational diagnosis, and similar notions of diagnosis developed in the previous chapters. The measures discussed below are similar to those proposed in [Indurkhya & Weiss, 1989].

Case-specific performance measures

Let D denote a database comprising N cases, where, for each case, one or more diagnostic results are available. Assume that a diagnostic expert system produces zero, or one or more conclusions for each case in the database D. For each case k in the database, $1 \leq k \leq N$, we define:

- c_k to be the number of results correctly diagnosed by the expert system;
- u_k to be the number of results not diagnosed (or unclassified) by the expert system;
- i_k to be the number of incorrect diagnostic conclusions by the expert system.

The number of correct diagnostic conclusions for case k produced by the expert system is equal to the number of correctly diagnosed results included in the database. Hence, it is not necessary to distinguish between these two numbers. The number of results available in the database D for each case k is equal to $r_k = c_k + u_k$, and the number of conclusions produced by the expert system for case k is equal to $e_k = c_k + i_k$.

Based on these measures, several other measures can be defined, providing information concerning the diagnostic performance of the expert system. The *accuracy* of an expert system, denoted by α_m, is defined as the weighed fraction of results that have been interpreted correctly:

$$\alpha_m = \frac{1}{N} \sum_{k=1}^{N} \frac{c_k}{c_k + u_k} \tag{8.1}$$

This measure expresses how well the system is capable of correctly diagnosing results associated with cases in the database. The subscript m indicates that this measure is suitable to interpret *m*ultiple conclusions associated with cases. The complementary measure,

$$\epsilon_m = 1 - \alpha_m \tag{8.2}$$

is called the *error rate*. It gives the false impression that all disorders that have not been diagnosed correctly, have been inferred to be absent. However, an expert system may establish no conclusions at all for certain results mentioned in the database; the error rate produces an overestimation of incorrectly diagnosed results. A limitation of the accuracy measure is that it does not provide precise information regarding how well a diagnosis produced by the expert system predicts a disorder to be present or absent. This information depends on the number of incorrect conclusions produced by the expert system for a particular case. The *predictive value* of an expert system, denoted by π_m, represents such information. It is defined as follows:

$$\pi_m = \frac{1}{N} \sum_{\substack{k=1, \\ c_k + i_k > 0}}^{N} \frac{c_k}{c_k + i_k} \tag{8.3}$$

If $c_k + i_k = 0$, for case k, $1 \leq k \leq N$, no diagnosis has been produced by the expert system. If the cases for which no diagnosis is produced are removed from the database D, a new measure is obtained, called the *adjusted predictive value* of the expert system, denoted by π_m^a:

$$\pi_m^a = \frac{1}{N^a} \sum_{\substack{k=1, \\ c_k + i_k > 0}}^{N} \frac{c_k}{c_k + i_k} \tag{8.4}$$

where N^a is the number of cases in the database after removal of the not diagnosed (unclassified) cases.

As was mentioned above, the performance measures introduced above are applicable to situations in which multiple results are available for every case in the database and possibly also multiple conclusions produced by the expert system. Formula (8.1) can be simplified if only a single diagnostic result is available for each case, as follows

$$\alpha_s = \frac{1}{N} \sum_{k=1}^{N} c_k = \frac{C}{N} \tag{8.5}$$

where $C = \sum_{k=1}^{N} c_k$ is the total number of cases for which a correct diagnostic conclusion has been established by the expert system. The subscript s indicates that this measure

Case	Database results	Conclusions by the expert system
1	$\{a, b\}$	$\{b, d, e, f\}$
2	$\{f, g, u\}$	$\{h, i\}$
3	$\{p, q\}$	$\varnothing$
4	$\{u\}$	$\{u\}$

Table 8.1: Example database.

concerns a *s*ingle result. This simplification is allowed, because $c_k + u_k = 1$, $k = 1, \ldots, N$. This is not an unrealistic restriction, because in medical databases, the diagnostic result is often expressed in terms of a single disorder, being the outcome of a final, clinical diagnosis, where the symptoms and signs included in the case description of a patient in the database are completely determined by this single disorder.

In contrast with the information available in databases, a diagnostic expert system often produces more than one conclusion, e.g. a set of alternative disorders that account for the observed findings. If we assume that only a single diagnostic result is available in the database for a case, it holds that $c_k \in \{0, 1\}$, $k = 1, \ldots, N$. If it is possible to select one conclusion for comparison from the diagnosis produced by the expert system, then this conclusion is either correct or incorrect. Hence, it holds that $(c_k + i_k) \in \{0, 1\}$, $k = 1, \ldots, N$. For example, if conclusions are associated with some measure of uncertainty, then the conclusion with the highest certainty might be selected. It might also be possible to just select the conclusion, if present, that matches a result in the database. Then, the adjusted predicted value can be simplified to the following expression:

$$\pi_s^a = \frac{1}{N^a} \sum_{k=1}^{N} c_k = \frac{C}{N^a} \tag{8.6}$$

Note that we now have that $\pi_s = \alpha_s$, and if all cases have been classified, in addition it holds that $\pi_s = \pi_s^a$.

Example 8.1. Consider the database D with associated results and diagnostic conclusions by the expert system presented in Table 8.1. Here, the performance of the system is determined by the match between the result included in the database D for a case and a conclusion produced by the expert system. From the case information included in this database it follows that the accuracy of the expert system is equal to

$$\alpha_m = \frac{1}{4}\left(\frac{1}{2} + \frac{0}{3} + \frac{0}{2} + \frac{1}{1}\right) = \frac{3}{8}$$

and its predictive value is equal to

$$\pi_m = \frac{1}{4}\left(\frac{1}{4} + \frac{0}{2} + \frac{1}{1}\right) = \frac{5}{16}$$

Furthermore, the adjusted predictive value is equal to

$$\pi_m^a = \frac{1}{3}\left(\frac{1}{4} + \frac{0}{2} + \frac{1}{1}\right) = \frac{5}{12}$$

Case	Database results	Conclusions by the expert system
1	$\{a, b, c\}$	$\{b, e\}$
2	$\{g, h, a\}$	$\{a, i\}$

Table 8.2: Example database for disorder-based measures.

In the adjusted predictive value, case (3) is simply ignored, resulting in a higher value. ◇

The accuracy and predictive-value measures only provide information concerning correct classification. Similar measures can be defined for incorrect or unclassified cases. Together, these measures offer a full, global description of the diagnostic performance of an expert system.

Disorder-specific measures

If there are many alternative diagnoses in a diagnostic systems, such as in the HEPAR system, in which about 80 disorders are dealt with, measurement of the performance with respect to individual disorders d yields useful information. Let Δ_{db} denote the set of disorders in the database, and Δ_{es} the number of disorders included in conclusions produced by the expert system for database D. Furthermore, let $N_{\text{db},d}$, $d \in \Delta_{\text{db}}$, be the number of cases concerning disorder d included in database D, and let $N_{\text{es},d}$, $d \in \Delta_{\text{es}}$, denote the number of diagnoses regarding disorder d produced by the expert system.

The *accuracy with respect to disorder d* of the expert system is now defined as follows:

$$\alpha_d = \frac{C_{\text{db},d}}{N_{\text{db},d}}$$

$d \in \Delta_{\text{db}}$, where $C_{\text{db},d}$ denotes the number of cases with disorder d that have been diagnosed correctly by the expert system. Similarly, the *predictive value with respect to disorder d* of the expert system is defined as follows:

$$\pi_d = \frac{C_{\text{es},d}}{N_{\text{es},d}}$$

$d \in \Delta_{\text{es}}$, where $C_{\text{es},d}$ denotes the number of diagnoses with respect to disorder d that were correct. Note that where cases regarding a certain disorder d can be either diagnosed correctly, incorrectly or may be unclassified, a diagnosis produced with respect to a disorder d by the expert system is either correct or incorrect.

It is usually instructive to compute measures for incorrectly diagnosed and unclassified cases, and incorrectly produced diagnoses by the system. These measures, however, are quite similar to those defined above regarding correctness.

Example 8.2. Consider the database with associated diagnostic conclusions produced by the expert system, given in Table 8.2. We have, for example, $\alpha_a = 1/2$, but $\pi_a = 1$ – because the disorder a has diagnosed only for case (1), but all diagnoses concerning disorder a produced by the expert system are correct – and $\alpha_b = 1$, $\pi_b = 1$. ◇

The measures defined above can be taken as point estimators of statistical parameters. The uncertainty with respect to their value can be determined by computing their associated confidence intervals, assuming a binomial or, for a sufficiently large database, an approximating normal distribution [Kreyszig, 1970]. However, when only a few cases are available for validation, confidence intervals will vary widely, and, therefore, convey little useful information. In particular, computation of confidence intervals for the disorder-specific measures will not be informative.

Statistical measures

Standard statistical techniques may also be used to compare the performance of an expert system with known diagnoses of patients or with the capabilities of clinicians. Some of these will be briefly reviewed in this section.

A method for paired comparison can be applied to validate the level of performance of an expert system in comparison with the performance of human experts or a decision-support system. Let x_k be the result of the expert system and y_k be the result of a human expert, or a decision-support system, for a given test case k, where $x_k, y_k \in \{0, \ldots, p\}$, $1 \leq k \leq N$, $p \geq 1$, if the results can be mapped to an ordered (ordinal) scale. If only two single results are compared, the result available for the test case is assigned the number '1', and the conclusion of the expert system is assigned either '1' or '0', depending on whether or not the conclusion is correct or incorrect (or there is no diagnosis at all). For each case k, the difference $\delta_k = x_k - y_k$ is computed, and finally the sample mean $\bar{\delta}$. Using the Student's t statistic, the null hypothesis that there is no difference between the results of the two systems, i.e. that the population has mean $\mu = 0$, is tested against the alternative that one system's performance is superior to the other.

If a system produces results that can be interpreted probabilistically, a possible statistical test is 'goodness of fit', in which the predicted and actual probability of a disorder given certain evidence is compared to each other in terms of a χ^2 probability distribution. This approach is not useful, if the results cannot be interpreted probabilistically, as is true for systems based on subjective probability assessment, and even more for rule-based expert systems using ad-hoc models of uncertainty handling, such as the certainty-factor model.

8.1.4 Data used in the validation process

Data to be used in validation of an expert system may be collected in several ways. In an informal validation of an expert system, data are simply obtained from expert physicians who enter real or imaginary cases into the system; the results of the system are assessed by the same physicians. An example of a medical expert system for which an informal validation was performed, is CASNET [Kulikowski & Weiss, 1982]. The field test is the other extreme.

Probabilistic systems are commonly constructed using information from large databases with cases. It is often possible to decompose the database into two separate parts: one part is used for learning purposes; the other part is used for testing purposes. However, typically there are no, or no really suitable, databases available for the automatic construction of an expert system. As discussed in Section 7.4, a small database with patient cases has been used as a learning population during the construction of the HEPAR system. This database, however, would have been too small for the construction of a probabilistic system. In medicine, though, it is always possible to collect data from patients with a disorder, diagnosed in the past, for use in a validation study, yielding what is called a *retrospective validation study*. A disadvantage of retrospectively collected data is that they are often incomplete and imprecise. In a *prospective validation study*, data are intensionally collected to be used for validation purposes. It is not surprising that data gathered during a prospective validation study are more complete and precise. But even a validation of an expert system using prospective data is seldom accomplished in a fully satisfactory manner. Firstly, the number of cases required for reaching statistically significant results is large. Secondly, a validation must cover most of the disease categories distinguished in the expert system. However, many medical expert systems include descriptions of highly rare disorders, making collecting data covering the entire range of disorders difficult or even impossible. Nevertheless, validation of any type yields valuable information about the expert system; two such studies of HEPAR have been performed.

8.2 Validation studies of HEPAR

As diagnostic performance was considered the first and most important feature to be assessed, only the diagnostic conclusions of HEPAR, including the intermediate conclusions, have been subject to investigation. In two successive validation studies of HEPAR, retrospectively collected data from patients with known clinical diagnoses were used. These clinical diagnoses were compared to the conclusions reached by the system. Furthermore, the Pocket Diagnostic Chart was used as a decision-support system to which the performance of HEPAR was compared. The testing environment discussed in the previous chapter was applied for collecting the results of validation. As discussed in Section 1.3.2, the Pocket Diagnostic Chart is a diagnostic system, usually applied in the form of a paper chart, that classifies a patient into one of four categories, based on values of 21 variables (see Table 8.3). On the basis of the three total scores for non-obstructive (medical, med) versus obstructive (surgical, surg) jaundice, acute (acu) versus chronic (chro) jaundice, and benign (ben) versus malignant (mal) jaundice, the probabilities of acute non-obstructive, chronic non-obstructive, benign obstructive and malignant obstructive jaundice can be calculated by multiplying the scores by each other, after inverse transformation to probabilities (cf. the discussion on logistic regression in Section 1.2.1).

During the process of data collection, various problems arose due to differences in the terminology employed at the various sites. For example, chronic active hepatitis with certain autoimmune features was called ‘chronic active hepatitis’ at the data collection site, whereas at the development site of HEPAR the disease was known as ‘autoimmune chronic hepatitis’. There were also differences in the terminology used for assessing the ultrasonography of the liver and biliary tract. A small number of problems were due to disagreement on the aetiology of some diseases between the hepatologist involved in the development and the one involved in validation of the system. Finally, the normal range of some of the data used in validation of the HEPAR system differed from the normal

	Med vs. Surg	Acu vs. Chro	Ben vs. Mal		Med vs. Surg	Acu vs. Chro	Ben vs. Mal
Age: 31 – 64 years	+7	+5		Physical			
≥ 65 years	+12	+5		examination:			
Previous history:				Spiders	−6	+11	
Jaundice due to	−7	+8		Acites	−3	+6	
cirrhosis				Liver surface nodular		+5	
Cancer in GI-tract,				Gall bladder:			
pancreas, bile	+10		+7	Courvoisier	+16		+11
system, or breast				firm or tender	+5		
				Clinical chemistry:			
Leukaemia or	−13			bilirubin ≥ 200μmol/l	+5	−5	+5
malignant							
lymphoma							
Previous biliary				Alkaline phosphatase:			
colics or proven				400 – 1000 U/l	+6		
gallstones	+3	+7	−7	> 1000 U/l	+11		+6
In treatment for							
congestive heart							
failure		−5					
Present history:				ASAT:			
				40 – 319 U/l		+5	
≥ 2 weeks			+7	≥ 320 U/l	−10	+1	+6
Upper abdominal pain:				Clotting factors:			
sever	+9		−6	≤ 0.55		+8	+5
slight or moderate	+4			0.56 – 0.70		+5	+5
Fever:				LDH ≥ 1300 U/l		−5	+7
without chills		−3	−5				
with chills		−6	−10				
Intermittent jaundice	+5		−5				
Weight loss (≥ 2 kg)			+4				
Alcohol:							
1 – 4 drinks per day	−4			SUM left			
≥ 5 drinks per day	−4	+4		CONSTANTS	−19	−21	−8
SUM left				TOTAL SCORE			

Table 8.3: Pocket Diagnostic Chart [Matzen et al., 1984].

	Normal Range		Normal Range		Conversion
Laboratory Data	**Original Database**		**HEPAR**		**Factor**
ALAT	5 – 30	U/l	2 – 15	U/l	0.5
Alkaline Phosphatase	25 – 75	U/l	15 – 60	U/l	0.8
α_1-antitrypsine	1.9 – 3.5	g/l	200 – 400	mg%	110
Amylase serum	30 – 130	U/l	15 – 60	U/l	2.3
Amylase urine	40 – 700	U/l	15 – 60	U/l	2.8
ASAT	5 – 30	U/l	2 – 15	U/l	0.5
Bilirubin	2 – 12	μmol/l	3 – 17.1	μmol/l	1.5
LDH	160 – 320	U/l	$<$ 160	U/l	0.5
LIBC	50 – 110	μmol	27 – 54	μmol	0.5

Table 8.4: Conversion factors used in the translation of the laboratory data from the databases of patients from Dijkzigt Hospital to values used in HEPAR.

range assumed in the HEPAR system. The problems mentioned above could only be solved after ample discussions with the medical specialists involved.

A special-purpose computer program was developed to carry out a conversion from the original data of the Rotterdam University Hospital (Dijkzigt Hospital), mapping the terminology employed at Dijkzigt Hospital to that of HEPAR. Table 8.4 gives the conversion factors for the laboratory data (data not mentioned had similar normal ranges as data in HEPAR). These factors were computed on the basis of the mid-points of the normal range intervals. Only the data relevant for HEPAR, obtained early after admission of the patient to the clinic, were selected for inclusion into the validation databases.

8.3 First retrospective validation study

In this section, the result of the first performance validation study of the HEPAR system are described.

8.3.1 Patients and methods

The first validation study of the HEPAR system was carried out using data of 101 consecutive jaundiced patients admitted to the Department of Internal Medicine II of Dijkzigt Hospital between 1984 and 1985. This department was not involved in the development of the HEPAR system. The database was originally compiled for the evaluation of the Pocket Diagnostic Chart, [Segaar et al., 1988], but the process of data collection was organized in such a way that the data required by the HEPAR system were collected as well. The data entered into the database were those obtained early after admission of the patient.

The first performance validation of the HEPAR system was carried out by the author in 1987–1988. In order to be accepted as a case in the test population, a patient was required to have an age of more than or equal to 15 years, clinical jaundice and a serum total bilirubin concentration of more than 17 μmol/l. For each patient a final clinical diagnosis was required, possibly confirmed by specific tests such as liver biopsy, endoscopic

Characteristic	First Population	Second Population
Number of patients	101	214
Number of admitted patients	94	181
Sex		
Male	53	91
Female	41	90
Duration of complaints		
≤ 2 wk	14	19
> 2 wk	80	152
unknown	0	10
Age (yr)		
15 - 30	7	18
31 - 50	12	32
51 - 65	49	77
> 65	26	54
Diagnostic categories (COMIK)		
acute non-obstructive disorder	7	20
chronic non-obstructive disorder	49	74
benign obstructive disorder	13	23
malignant obstructive disorder	25	64
Diagnostic categories (HEPAR)		
hepatocellular disorder	47	78
biliary obstructive disorder	47	103
benign disorder	64	117
malignant disorder	30	64

Table 8.5: Characteristics of the test populations from Dijkzigt Hospital.

retrograde cholangiography, surgery or autopsy. Of the admitted 101 patients, seven were withdrawn, four because the clinical diagnosis was unclear and three because the system did not yet contain information concerning these disorders. The remaining population of patients consisted of 53 male and 41 female patients. The average age of the patients was 58 (with minimum 18 and maximum 90). Some of the characteristics of these patients are summarized in the second column of Table 8.5. In the table, the patients are classified according to the four categories distinguished for the Pocket Diagnostic Chart and the four intermediate conclusions distinguished in HEPAR.

8.3.2 Results

Table 8.6 shows the clinical diagnoses of the test population of 94 patients and gives the results provided by the HEPAR system (accuracy with respect to a disorder, α_d), for each of the diagnoses. In Table 8.7 the diagnostic conclusions, as distinct from the clinical diagnoses, concerning the patients classified by the HEPAR system are presented (predictive value with respect to a disorder, π_d). Table 8.8 summarizes the results obtained. A classification concerning the type of hepatobiliary derangement was obtained in 95%

		Results provided by HEPAR (n)		
Final Diagnosis	**Total (n)**	**Correct**	**Incorrect**	**Unclassified**
acute hepatitis B	3	3	0	0
alcoholic cirrhosis	18	16	2	0
alcoholic hepatitis	1	1	0	0
amyloidosis	1	0	0	1
autoimmune chronic hepatitis	8	8	0	0
bifurcation carcinoma	6	4	0	2
carcinoma of papilla of Vater	2	0	1	1
Caroli's disease	1	0	0	1
chronic hepatitis B	8	8	0	0
common bile duct stone	4	1	1	2
common bile duct stricture	2	0	0	2
cryptogenic cirrhosis	2	0	1	1
cryptogenic chronic hepatitis	2	1	0	1
gallbladder carcinoma	3	0	2	1
malignant lymphoma	2	1	0	1
metastatic tumour	6	2	2	2
pancreatic carcinoma	6	4	1	1
pancreatitis	1	0	0	1
primary biliary cirrhosis	10	9	0	1
primary hepatocellular tumour	3	1	2	0
primary sclerosing cholangitis	4	1	2	1
sepsis	1	0	1	0

Table 8.6: Clinical diagnoses of 94 patients of Dijkzigt Hospital, used as a test population for HEPAR, and results.

of the patients; 85% (adjusted predictive value, π_s^a) was classified correctly. The system achieved a conclusion with regard to the benign or malignant nature of the disorder in 65% of the patients; of these conclusions, 92% was correct. Finally, the system established a differential diagnosis in 80% of the cases; 80% of these contained the correct final diagnosis as the one with the highest certainty factor. The accuracy of the system (with 95% confidence intervals assuming a normal distribution) was: 81% (±8.0) for the type of hepatobiliary derangement, 60% (±9.8) for the nature of the disorder, and 64% (±9.7) for the final diagnosis. Note that the number of unclassified patient cases has a large effect on these numbers. The average number of conclusions drawn as part of the differential diagnosis was less than 3. In 87% of the classified cases, the right conclusion occurred as one of the conclusions in the differential diagnosis.

The effects of incompleteness of entered information on the diagnostic performance of the system were also investigated. For this purpose, three tests were carried out. In test A for each patient all available data were presented to the system. The results for this test are identical to those mentioned in Table 8.8. In test B, only data concerning the symptoms, signs, haematology and blood chemistry, but neither data from serological tests nor those from ultrasonographical investigations were presented to the system. In test C, only data from the medical interview and physical examination were presented.

Diagnostic Conclusion	**Total** *n*	**Correct** *n*	**Incorrect** *n*
acute cholangitis	1	0	1
acute hepatitis B	4	4	0
alcoholic cirrhosis	17	16	1
autoimmune chronic hepatitis	10	8	2
bifurcation carcinoma	8	4	4
chronic hepatitis B	10	8	2
cholelithiasis	1	0	1
circulatory liver damage	1	0	1
common bile duct stone	2	1	1
cryptogenic chronic hepatitis	1	1	0
malignant lymphoma	1	1	0
metastatic tumour	4	2	2
pancreatic carcinoma	4	4	0
primary biliary cirrhosis	9	9	0
primary hepatocellular tumour	1	1	0
primary sclerosing cholangitis	1	1	0

Table 8.7: Diagnostic conclusions for the 75 classified patients out of 94 cases from Dijkzigt Hospital, used as a test population for HEPAR.

Conclusion	**Correct** *n* (%)	**Incorrect** *n* (%)	**Unclassified** *n* (%)	**Total** *n* (%)
Type of hepatobiliary derangement	76 (81)	13 (14)	5 (5)	94 (100)
Benign/malignant nature of disorder	56 (60)	5 (5)	33 (35)	94 (100)
Final diagnosis	60 (64)	15 (16)	19 (20)	94 (100)

Table 8.8: Diagnostic results of the HEPAR system for the test population of 94 patients with hepatobiliary disorders from Dijkzigt Hospital.

	Correct (%)			Incorrect (%)			Unclassified (%)		
Conclusion	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Type of hepato-biliary derangement	81	81	60	14	14	17	5	5	23
Benign or malignant nature of disorder	60	60	56	5	5	3	35	35	40
Final diagnosis	64	36	35	16	12	10	20	52	55

A: All available data presented to system.
B: Only data concerning symptoms, signs, haematology and bloodchemistry (no data from ultrasound or serology presented).
C: Only data from medical interview and physical examination.

Table 8.9: Assessment of the effects of incompleteness of information on the diagnostic conclusions of the system, for a database of 94 patients with hepatobiliary disease from Dijkzigt Hospital.

Table 8.9 summarizes the results of this experiment. The results of test A were used as a reference point. We may conclude that the relative number of incorrect conclusions did not increase when a decreasing number of data was presented to the system, although the relative number of unclassified cases did increase.

8.4 Second retrospective validation study

In this section, the results of the second performance validation study of HEPAR are discussed.

8.4.1 Patients and methods

For the second performance validation study of the HEPAR system, data from 214 consecutively admitted jaundiced patients from the same department as for the first validation study were collected. These patients were admitted to the clinic between 1986 and 1988. As for the patients in the first test population, it was required that a clinical diagnosis was established according to internationally accepted criteria. Because the database was originally compiled for the evaluation of the much simpler classification scheme used by the Pocket Diagnostic Chart, for several cases the final diagnosis was unclear, uncertain, or even completely unknown. For these reasons, 16 cases were removed from the database. Furthermore, only those patient cases having a disorder belonging to one of the about 80 disorders covered by the HEPAR system were admitted to the second test population of HEPAR. Therefore, another 17 cases were removed from the databases, leaving 181 cases for use in the validation study. The remaining population of patients consisted of 91 male and 90 female patients. The average age of the patients was 57 (with minimum 15 and maximum 91). Some of the characteristics of these patients are presented in the third column of Table 8.5.

During the time period between the first and the second performance validation study,

		Results provided by HEPAR (n)		
Clinical Diagnosis	**Total (n)**	**Correct**	**Incorrect**	**Unclassified**
acute cholangitis	9	8	0	1
acute hepatitis B	11	9	0	2
alcoholic cirrhosis	31	24	5	2
alcoholic hepatitis	2	1	1	0
amyloidosis	1	0	0	1
autoimmune chronic hepatitis	8	5	2	1
bifurcation carcinoma	9	8	0	1
carcinoma of papilla of Vater	5	4	1	0
chronic hepatitis B	10	8	0	2
chronic hepatitis C	1	0	0	1
circulatory liver damage	3	3	0	0
common bile duct carcinoma	8	3	5	0
common bile duct stone	14	9	2	3
cryptogenic cirrhosis	3	0	0	3
cytomegalic inclusion disease	1	0	1	0
gallbladder carcinoma	3	2	1	0
hepatitis A	3	1	2	0
hydatid cyst	1	0	0	1
metastatic tumour	11	6	3	2
pancreatic carcinoma	21	18	3	0
pancreatitis	1	0	1	0
primary biliary cirrhosis	15	11	2	2
primary hepatocellular tumour	7	2	4	1
primary sclerosing cholangitis	1	1	0	0
secondary biliary cirrhosis	1	0	1	0
sepsis	1	0	0	1

Table 8.10: Clinical diagnoses of 181 patients from Dijkzigt Hospital, used as a test population for HEPAR, and results.

the knowledge base of the HEPAR system was improved and extended to include several disorders previously not covered by the system. In particular, the description of ultrasonography in HEPAR was improved and extended for representing data from Dijkzigt Hospital. The nature of most other data was unchanged with respect to the first performance validation study. Most of the formalized knowledge from the previous version of the system, was included in the new version of HEPAR unmodified. The performance results for the new version of HEPAR using the database of the first validation study are approximately the same as those described in the previous section. The testing environment was also improved in the period between the two validation studies. As a consequence, the results of the second validation study are slightly more detailed than the results of the first validation study.

Diagnostic Conclusion	Total n	Correct n	Incorrect n
acute cholangitis	8	8	0
acute hepatitis B	9	9	0
alcoholic cirrhosis	25	24	1
alcoholic hepatitis	1	1	0
autoimmune chronic hepatitis	8	5	3
bifurcation carcinoma	12	8	4
carcinoma of papilla of Vater	11	4	7
chronic hepatitis B	11	8	3
circulatory liver damage	5	3	2
common bile duct carcinoma	3	3	0
common bile duct stone	12	9	3
Crigler-Najjar syndrome	1	0	1
gallbladder carcinoma	2	2	0
hepatitis A	1	1	0
metastatic tumour	6	6	0
pancreatic carcinoma	21	18	3
primary biliary cirrhosis	12	11	1
primary hepatocellular tumour	2	2	0
primary sclerosing cholangitis	1	1	0
steatosis hepatis	5	0	5
Zieve's syndrome	1	0	1

Table 8.11: Diagnostic conclusions for the 157 classified patients out of 181 cases from Dijkzigt Hospital, used as a tests population for HEPAR.

8.4.2 Results

For each entered patient case, the conclusion produced by HEPAR with the highest certainty was compared with the known clinical diagnosis. For multiple disorders, the disorder which at the time of admission best explained the patient's signs and complaints was selected for comparison. Table 8.10 shows the clinical diagnoses for the patients in the second test population of 181 cases, and gives for each of the diagnoses the results of the HEPAR system. In Table 8.11 the conclusions of HEPAR for the classified cases with respect to the final diagnosis are presented. The results are summarized in Table 8.12. The accuracy of the system, α_s, (with 95% confidence intervals assuming a normal distribution) was: 82% (±5.6) for the type of hepatobiliary derangement, 83% (±5.5) for the nature of the disorder, and 68% (±6.8) for the final diagnosis. When considering only the classified patients (adjusted predictive value), the conclusions were correct in 86%, 83% and 78%, respectively. The average number of conclusions drawn by the system, making up the differential diagnosis, was about four.

Furthermore, it was determined whether the clinical diagnosis occurred as one of the diagnoses in the differential diagnosis. The correct final diagnosis was in the differential diagnosis produced by the system in 79% of the patients, and in 91% of the classified patients, respectively. When the final diagnoses produced by HEPAR were classified into the four categories as distinguished for the Pocket Diagnostic Chart, the final conclusion

Conclusion	Correct *n* (%)	Incorrect *n* (%)	Unclassified *n* (%)	Total *n* (%)
Type of hepato-biliary derangement	149 (82)	25 (14)	7 (4)	181 (100)
Benign/malignant nature of disorder	150 (83)	31 (17)	0 (0)	181 (100)
Final diagnosis	123 (68)	34 (19)	24 (13)	181 (100)

Table 8.12: Diagnostic results of the HEPAR system for the test population of 181 patients with hepatobiliary disorders from Dijkzigt Hospital.

	Correct (%)			Incorrect (%)			Unclassified (%)		
Conclusion	A	B	C	A	B	C	A	B	C
Type of hepato-biliary derangement	82	82	69	14	14	14	4	4	18
Benign or malignant nature of disorder	83	83	82	17	17	18	0	0	0
Final diagnosis	68	39	34	19	25	22	13	36	44

A: All available data presented to system.
B: Only data concerning symptoms, signs, haematology and bloodchemistry (no data from ultrasound or serology presented).
C: Only data from medical interview and physical examination.

Table 8.13: Assessment of the effects of incompleteness of information on the diagnostic conclusions of the system, for a database of 181 patients with hepatobiliary disease from Dijkzigt Hospital.

of HEPAR was correct in 76% of the patients and in 88% of the classified patients.

Again an experiment was carried out concerned the effects on the conclusions of the system when certain data were omitted from the input. In this manner, we were able to study how well the system behaved in the absence of certain patient data, for example ultrasound and serological data. The results of this experiment are shown in Table 8.13. It should be noted that although the percentage of unclassified cases increased, only a limited change in the number of incorrect conclusions occurred.

8.5 Comparison to related work

In the first validation study, the HEPAR system was shown to be capable of establishing the right conclusion concerning the type of the hepatobiliary derangement, the benign or malignant nature of the disorder, and the final diagnosis, in 85%, 92%, and 80% of the classified cases, respectively. In a relatively large number of patients from the test population the system did not reach a conclusion regarding the benign or malignant nature of the disorder. This was due to the fact that rules concerning the benign or malignant nature of hepatocellular disorders were not yet present in the system. In the second validation study, the system was capable of deriving a correct conclusion in 86%, 83% and 78% of the classified cases. The experiments in which the system was evaluated with respect to its sensitivity to incompleteness of entered information showed an increase in the relative number of unclassified cases when the amount of information presented to the system was decreased. However, lack of information did not significantly increase the relative number of incorrect conclusions established by the system.

In three performance validation studies of the Pocket Diagnostic Chart, 75% [Matzen et al., 1984], 78% [Lindberg et al., 1987] and 77% [Segaar et al., 1988], respectively, of retrospectively compiled test populations were classified correctly. The database used in the second validation study has been applied as a learning population to adapt the Pocket Diagnostic Chart, which was originally based on data from 1002 Danish patients with jaundice, to the local characteristics of the Dijkzigt patient population. This work yielded a Rotterdam version of the chart [Segaar, 1991]. Using the database from our first validation study as a test population, in 77% of the patients a correct diagnosis was established. Contrary to expectations, local adaptation yielded no improved performance, since the performance of the unadapted system was already 77% [Segaar, 1991].

Although the Pocket Diagnostic Chart is not a perfect standard for comparison, because it classifies a patient case only into one of four diagnostic categories, comparison to HEPAR provides some useful information. For the purpose of a comparison, in the second validation study, final diagnostic conclusions were classified into the diagnostic categories distinguished for the Pocket Diagnostic Chart, yielding a correct 'chart' diagnosis for 76% of the patients and 88% of the classified patients. Of course, for a really fair comparison of these outcomes an identical database should have been used. But since the data used in the second validation study of HEPAR were employed for the construction of the Rotterdam version of the Pocket Diagnostic Chart, this was impossible. A comparison of limited scope has been carried out for the intermediate conclusions produced by the HEPAR system in the first validation study, using the Rotterdam version of the Pocket Diagnostic Chart. In this study, the Pocket Diagnostic Chart produced a correct conclusion concerning the type of the disorder in 83% of the patients, where the result for HEPAR was 81%, and the nature of the disorder was established correctly by the Pocket Diagnostic Chart in 86% of the patients, whereas in HEPAR only 60% of these conclusions were correct [Segaar, 1991]. However, with regard to HEPAR it should be noted that the large number of unclassified cases had a major effect on these scores. Furthermore, the scores were based on the early version of the HEPAR system. In the improved and extended HEPAR system, the 60% result has been increased to 90%, which seems to provide a better indication of the present performance of the system with respect to the nature of the disorder, as is confirmed by the results of the second validation study.

Another system for the diagnosis of liver disease, which uses Bayes' theorem and which is capable of recognizing 13 disease entities in the field of liver disease, was shown to reach an accuracy of 75% [Malchow-Møller et al., 1986]. In the EURICTERUS project (cf. Chapter 1), a variety of techniques for the development of diagnostic systems for the diagnosis of 17 different hepatobiliary disease entities have been investigated. The accuracy of the systems varied between 57% for a probabilistic system based on the 'independent form' of Bayes' theorem and 62% for a neural network [EURICTERUS, 1993]. Thus, it appears that HEPAR as described in Chapter 7 has a diagnostic performance comparable to, or better than, other diagnostic systems. The main advantage of the HEPAR system over the other systems is that the classification produced by HEPAR is very detailed and corresponds to the differential diagnosis as used in the clinic, a feature not found in the other systems. Recall that the correct final diagnosis often occurred as one of the items in the differential diagnosis, an important condition for clinical acceptance of the system. Moreover, the number of items making up the differential diagnosis was, on average, small relative to the about 80 disorders included in the expert system.

It is also interesting to compare the validation results for HEPAR to the results of four commercially available diagnostic expert systems in the broad field of internal medicine, which encompasses the domain of hepatology: DxPlain [Barnett et al., 1987], Meditel [Waxman & Worley, 1990], Iliad [Warner, 1989; Heckerling et al., 1991] and QMR [Bankowitz et al., 1989]. In a comparative laboratory study of their diagnostic performance on 105 patient cases, their (multiple diagnosis) accuracy α_m with regard to the final diagnosis varied between 52% and 69% for all cases, and between 71% and 89% for 63 patient cases with disorders covered by all four systems [Berner et al., 1994]. However, the average number of disorders generated by the systems varied between 6.6 and 13.3, and only in 10–22% did the clinical diagnosis correspond to the first disorder listed. Hence, as compared to these broader systems, HEPAR is more restrictive, but its predictive capabilities are considerably better. The diagnostic task for the broader systems is much more difficult than for a more narrowly focussed system like HEPAR, but it is doubtful whether the low predictive value of the advice generated by these systems in their current stage of development is acceptable.

8.6 Discussion

It is good practice to submit a medical expert system, whether primarily designed to support diagnostic problem solving or some other medical task, to a laboratory validation study. Before being used in a real clinical context, even when only under experimental circumstances, appropriate insight into its accuracy and usefulness are required. Thus, sufficient confidence in the performance of the system should be gained before considering submitting the system to other validation studies, such as a field test. Yet, only a limited number of medical diagnostic expert systems have reached the stage of a laboratory validation. Examples of such systems are the expert systems mentioned above and MYCIN [Buchanan & Shortliffe, 1984; Yu et al., 1979a; Yu et al., 1979b], PUFF [Aikins, 1980], CADIAG/PANCREAS [Adlassnig & Scheithauer, 1989], MENINGE [François et al., 1992], ANEMIA [Quaglini et al., 1988], PLEXUS [Van Daalen, 1993] and the gynaecological expert systems developed by Todd and Stamper [Todd & Stamper, 1993]. Although all these systems have had their performance validated, the validation was restricted, using an insufficient number of test cases to reach sufficient confidence with respect to their accuracy. Rare exceptions are the systems developed by Todd and Stamper [Todd & Stamper, 1993]. There are various reasons for this, most of them related to problems inherent to every validation of a diagnostic or therapeutic procedure. To perform a reliable validation of an expert system, it is required that the data which are used for this purpose are collected as part of a prospective study, preferably using a double-blind procedure. The main disadvantage of this approach lies in its very expensive and time-consuming nature. Moreover, it is often impossible to collect all necessary data for a patient, because this would imply submitting a patient to a large array of diagnostic tests, only partly necessary from a clinical point of view. In particular, this problem will arise when validating broadly scoped expert systems; it is less likely to occur in more narrowly scoped expert systems that focus on the early assessment of patients. Assessment of the usefulness of system by comparing the user's performance with and without assistance by the expert system may then be the only feasible alternative (cf. [Bankowitz et al., 1989; Heckerling et al., 1991]), but typically such studies meet many difficulties (cf. [Van Daalen, 1993]). The amount of patient data needed for a full-scale validation of an expert system is very large, thus preventing a formal validation in most situations. Also, disorders which are met infrequently in clinical practice are virtually excluded from such validation, otherwise it may take several years before sufficient data are collected. An alternative approach, which can be realized more easily and which also offers valuable information about the diagnostic quality of the system is to perform a retrospective study, which is the kind of study that has been carried out for most of the systems mentioned above. But even then, the amount of time and money involved may be substantial. The two successive validation studies of HEPAR, discussed in Sections 8.3 and 8.4, are also instances of such attempts. For example, it took about three years before the set of patient data used in the second validation study of HEPAR was of sufficient quality for actual use.

From the discussions above, it seems that the level of performance of the HEPAR system is sufficiently high to warrant further study in a clinical context. As discussed in Section 1.3.1 and in Chapter 6, there are sound, clinical reasons for studying ways to improve diagnosis in disorders of the liver and biliary tract. Whether or not an expert system like HEPAR can influence clinical practice in a positive way, seems a relevant question. Several field tests for other expert systems that have been carried out in the past have been fraught with many difficulties (cf. [Bankowitz et al., 1989; Heckerling et al., 1991; Hickam et al., 1985; Van Daalen, 1993]). Several researchers have attempted to analyse the sources of these difficulties [Kassirer, 1994; Miller & Masarie, 1990; Shortliffe, 1989; Szolovits et al., 1988]. One of the the most important problems may be the lack of integration in the clinical environment that any medical decision-support system, not only expert systems, has to face. The inadequate computational infrastructure in hospitals is one of the main causes for this [Shortliffe, 1991]. In contrast to practitioners in many other disciplines, medical doctors are still not accustomed to the idea of a workstation as a general problem solving tool. This situation is likely to change as soon as information technology in general has gained wider acceptance in the clinic.

Chapter 9

Conclusions and Further Research

In this thesis, diagnostic problem solving has been investigated from several points of view, ranging from a formal basis to the development and validation of an actual medical diagnostic system.

In the first part of this thesis, it was argued that the diagnostic interpretation of knowledge in a diagnostic expert system need not be fixed, but rather may depend on the diagnostic purpose for which the knowledge is employed (e.g. DNSB diagnosis or MAB diagnosis), on information about the accuracy and completeness of modelled knowledge, as well as on the underlying semantics of the represented knowledge. This led to the proposal of a set-theoretical framework within which diagnostic interpretation relations, called 'notions of diagnosis', are central. Next, various notions of diagnosis were described and their basic properties analysed.

In the second part of this thesis, one of the notions of diagnosis, called associational diagnosis, was used in the development of a medical expert system for the diagnosis of disorders of the liver and biliary tract. In this concluding chapter, the scientific contributions of this work are addressed, and several directions for further research are identified.

9.1 Principles of diagnosis

First, our achievements with respect to the theory of diagnosis are summarized.

9.1.1 Notions of diagnosis

The major body of work in this thesis elaborates the idea that an important part of diagnostic problem solving can be viewed as a variable relation between a formal representation of domain knowledge, hypotheses, and diagnoses. A diagnosis is viewed as an accepted or adjusted hypothesis. A given notion of diagnosis, which may be defined in general domain-independent terms, determines on how a knowledge base is to be interpreted. Our set-theoretical framework of diagnosis provides a set of simple mathematical tools for the analysis of diagnostic methods. The purpose of its application is increased insight into the nature of diagnosis. This approach to diagnosis is in accordance with recent research directions in the field of knowledge acquisition, where the development of models of a specific problem domain is guided by generic interpretation models, such as for diagnosis [Karbach et al., 1990]. These interpretation models are informal in nature, and, as a consequence, several more subtle aspects of diagnosis, such as the handling of interactions, have not been dealt with adequately in the literature. As the need for a formal underpinning of development methodologies for knowledge-based systems is now clearly recognized (cf. [Fensel & Van Harmelen, 1994; Fensel, 1995; Treur & Wetter, 1993; Van Harmelen & Balder, 1992]), the framework of diagnosis proposed in this thesis may be viewed as being complementary to that work.

9.1.2 Expressiveness

In comparison to other work on diagnosis, the framework of diagnosis developed in this thesis is more expressive. In contrast to the set-covering theory of diagnosis by Reggia et al., [Peng & Reggia, 1990], it is possible to capture various sorts of interactions among defects and findings in our framework. Furthermore, where the set-covering theory is built upon a single, fixed diagnostic interpretation of causal knowledge, in our framework the fixed diagnostic interpretation is replaced by the variable 'notion of diagnosis'.

The abductive theory of diagnosis by Console and Torasso, [Console & Torasso, 1991], is more expressive than the set-covering theory of diagnosis, but cannot express several desirable properties of diagnostic knowledge. The main limitation of the theory is the strong connection between the interpretation of domain knowledge for the purpose of diagnosis, and the definition of abductive diagnosis. An entailment relation is used as a foundation for capturing the meaning of diagnostic knowledge as well as for defining a spectrum of logical definitions of diagnosis. Similar remarks can be made with respect to the related work by K. Konolige, [Konolige, 1994]. When this entailment relation is monotonic, as in earlier papers by Console and Torasso, certain interactions between defects cannot be expressed. The inability to express certain types of interaction can be overcome by resorting tot nonmonotonic logics, but at the cost of a significant increase in technical complexity of the theory. In [Console et al., 1991], deduction with negation as failure is used for this purpose. However, the assumption of negative information by negation as failure affects the meaning of diagnostic knowledge; rules such as negation as failure carry with them the danger of keeping these assumptions implicit. Consequences of nonmonotonic interactions can be studied readily in terms of our framework. Another proposal to express nonmonotonic interactions is by D. Poole, [Poole, 1994], although the latter work goes beyond diagnostic problem solving, as diagnosis is embedded in a general framework for default reasoning in terms of hypothesis formation (cf. [Poole, 1989; Poole, 1990b]).

Consistency-based diagnosis as introduced by R. Reiter, [Reiter, 1987], and later extended by De Kleer et al., [De Kleer et al., 1992], has been identified as a specific notion of diagnosis in terms of the framework. It is interesting to note that, in contrast to consistency-based diagnosis, where refinements, such as dealing with 'fault masking', were the subject of several studies (e.g. [Konolige, 1992]), the definition of diagnosis in terms of our framework puts much more emphasis on identifying possible interactions among defects. Consequently, interactions, such as those resulting from fault masking are not easily omitted in a domain model.

The work in this thesis clearly shows that there are many different definitions of diagnosis, in fact many more than have been identified in the other literature on diagnosis, as exemplified by the new notions of refinement diagnosis introduced in Chapter 5.

9.1.3 Limitations

Ironically, the major limitation of the framework of diagnosis developed in this thesis lies in its generality. The framework imposes very few constraints with respect to notions of diagnosis, and it is not difficult to come up with a notion of diagnosis that nobody would find acceptable. Extra constraints may be required before the framework will be a truly general framework of diagnosis. Such constraints are not easily identified, except with respect to specific notions of diagnosis, such as strong-causality and weak-causality diagnosis.

Another observed limitation is that, as a tool for the semantical analysis of diagnosis, our framework is rather extensional in nature. This is in contrast to the more intensional nature of logic-based techniques for the analysis of diagnosis, such as used in defining consistency-based and abductive diagnosis. Thus, the separation of structure and function, as advocated by proponents of model-based reasoning (cf. [Davis, 1984]), possesses no clear analogue in the framework.

9.1.4 Further research

There are various aspects of the framework of diagnosis that need further investigation, that will be discussed in the following paragraphs.

Semantical characterization

In Chapter 3, several meaningful instances of evidence functions have been identified, such as evidence functions corresponding to a causal relation, and evidence functions expressing augmentation or cancellation. Other characterizations of diagnosis may provide further insight into the nature of diagnosis, and may also yield notions of diagnosis not previously described. In particular, the relationships between problem-domain properties, such as semantics, accuracy and completeness of domain knowledge, and the purpose of diagnosis, on the one hand, and suitable notions of diagnosis in the domain, on the other hand, must be elucidated.

Computational properties

In this thesis, the computational properties of the notions of diagnosis described, have only been touched upon. Yet, it is known that even the determination of an irredundant diagnosis in the set-covering theory of diagnosis is $\mathcal{NP}$ hard, [Bylander et al., 1992], even though this is a rather straightforward notion of diagnosis. Moreover, an evidence function is defined in terms of an exponential number of interactions among defects. Hence, in order to adopt the framework as a foundation for building practical diagnostic expert systems, research into the time and space complexity of these notions of diagnosis must be addressed. The fact that an evidence function can also be specified partially, and that a partial specification can be employed in diagnostic problem solving for some notions of diagnosis, could be a starting point for this research. The mathematical intricacy of the definition of an evidence function and a notion of diagnosis says little about its computational complexity. For example, the notion of associational diagnosis for interaction-free evidence functions admits computation of a single diagnosis in polynomial time and space, because only elements $e(d), d \in H$, need be considered.

It is well-known that the application of strategic, diagnostic knowledge can be important to cut down computational expenses (cf. [De Kleer & Williams, 1987; Mozetič, 1992; Treur, 1993]).

Distinguishing between various sorts of knowledge

In the framework for diagnosis, many of the subtleties in a domain, such as the relevance of observable findings in establishing a diagnosis, have not been considered. The framework could be easily enriched by making a distinction between typical and atypical findings, or findings that are relevant or irrelevant, e.g. in terms of information gain, in establishing a diagnosis, etcetera. These aspects should also be addressed as part of further research.

Information gathering

Little attention has been devoted to the dynamic aspects of diagnostic problem solving, in particular to the selective gathering of information in the course of solving a diagnostic problem has not been dealt with. Yet, these dynamic aspects are clearly important, as was also emphasized in the description of the structure of the HEPAR system. Information gathering has received some attention in the literature on consistency-based diagnosis (cf. [De Kleer et al., 1992; Hou, 1994; McIlraith & Reiter, 1992]), and also in the literature on abductive diagnosis, (cf. for example [Peng & Reggia, 1990]). Many notions of diagnosis described in the literature exploit the fact that the more defects are considered the more findings can be observed, i.e. they assume the evidence function to be monotonic, making things a lot easier. Furthermore, possible satisfaction of the independence assumption for the notion of diagnosis employed, may also be relevant in this respect. A framework for explicit reasoning about information-gathering strategies for diagnosis is proposed in [Treur, 1993].

Implementation

In the second part of this thesis, only the notion of associational diagnosis has been investigated at the level of practical implementation. It would be interesting to perform a comparative performance analysis of implementations of various notions of diagnosis for the same problem domain.

9.2 Medical decision support

The contributions made by the development of the HEPAR systems are next considered.

9.2.1 Associational diagnosis

As discussed in Chapter 1, developing diagnostic expert systems in terms of empirical associations is not new. Early expert systems such as MYCIN, [Shortliffe, 1976], were built along that line, but also the more recent PLEXUS system is based on this principle, [Jaspers, 1990]. Some domains can be readily represented as models of normal or abnormal behaviour for the purpose of diagnostic problem solving. As has been argued, hepatology is not one of those domains. Functional models of the liver and biliary tract would be either too complicated or unavailable. The only practical alternative seems to be to develop a model in terms of causal relations similar to those employed in the set-covering theory of diagnosis. It is not very likely that such simple causal relations will be sufficiently expressive for building a high-performance system in the domain of hepatology. Actually, the LIED system suggests that a much richer representation language is needed than simple causal relations [Console et al., 1992].

9.2.2 Development by refinement

As discussed in Chapter 7, the development of an expert system based on empirical associations, with associational diagnosis, requires experimental feedback in the course of its development. A simple computational environment has been built for the purpose of the development of HEPAR. This experience indicates that such experimental support should be added to current development methodologies of knowledge-based systems. Building a practical expert system is still too often viewed as a clean and straightforward modelling activity with little or no experimental feedback from the problem domain required. Hence, the situation for knowledge-based systems is not much different from system simulation, the field where the modelling of systems was actually first introduced [Shannon, 1975]. Experimental feedback differs from prototyping by the fact that the main purpose of prototyping is clarification of a system's requirements; the requirements for an expert system that requires experimental feedback may already be clear enough.

9.2.3 Performance and usefulness

The results for HEPAR discussed in the previous chapter indicate that the system could be of practical use in the clinic. It has been observed in the medical literature that clinicians with little experience in hepatology, i.e. every clinician not specialized in gastroenterology, have difficulty in diagnosing disorders of the liver and biliary tract at an early stage. Hence, there exists a potentially successful, clinical niche for the system.

9.2.4 Further research

There are several aspects of medical decision support that require further investigation. The most important of these are briefly discussed.

Integration into the clinical environment

The clinical usefulness of HEPAR has not yet been investigated. The experience of other researchers in studying the clinical usefulness of an expert system indicates that such studies are difficult to implement for various reasons. Possibly the most important reason may be the fact that medical expert systems are usually stand-alone systems. Such studies could be performed more easily if it were possible to integrate the system within an existing computational environment, such as an hospital information system, or World Wide Web (WWW). Unfortunately, the current computational infrastructure of hospitals is still quite primitive; even today, little or no clinical data are stored on computerized media. This problem must be addressed before considering the introduction of expert systems into the clinic.

Reasoning with uncertain knowledge

As discussed in the previous chapters, the HEPAR system deals with uncertain medical knowledge using the mathematically unsound certainty-factor model [Van der Gaag, 1990; Lucas & Van der Gaag, 1991]. However, the reasoning that is carried out by the system is mainly symbolic (logical) in nature, as was demonstrated by a reformulation of the system to first-order predicate logic [Lucas, 1993]. Although many researchers nowadays claim that the handling of uncertain knowledge requires the adoption of probability theory (cf. for example [Heckerman, 1992]), the consequence of this would be abandoning logical reasoning methods as are typically employed in knowledge-based systems, a price too high to pay. As an alternative, P. Krause et al., [Krause et al., 1995], propose a general framework, called the logic of argumentation, that is able to handle uncertainty within the setting of symbolic reasoning. The basic idea of the logic of argumentation – making decisions on the basis of arguments derived from proof-tree information – is similar to that of inference nets as proposed in [Van der Gaag, 1989; Lucas & Van der Gaag, 1991], which underly, amongst others, the certainty-factor model. It does not offer a mathematical solution to the problem, because the essential problem remains the design of a sound calculus of uncertainty that can be added to symbolic reasoning.

Decision theory and strategic reasoning

Since medical expert systems employing associational diagnosis are based on the experience of clinicians in the management of their patients, decisions as to which diagnostic procedures ought be undertaken for a given patient, are implicitly embodied in the knowledge encoded and is applied by means of some diagnostic strategy. Decision theory, being a combination of probability theory and utility theory, makes it possible to explicitly reason about such decisions, allowing the analysis of cost-benefit considerations of any action undertaken by the clinician, an application not readily accessible by systems based on associational diagnosis (however, cf. [Treur, 1993]). However, the development of such decision-theoretical systems for areas of the size that can be handled by current expert-system technology is a major undertaking. It therefore seems attractive to consider starting the development of such decision-theoretical models by taking an existing expert systems such as HEPAR as a point of departure. We have carried out an experiment in converting the HEPAR system into a belief network, looking at the notion of belief network primarily as a means for representing causal medical knowledge [Korver & Lucas, 1993]. No mapping was found that enables the automatic conversion of a rule-based expert system to a causal graph. Due to the differences in the type of knowledge represented and in the formalism used to represent uncertainties, much of the causal knowledge required in a belief network cannot be derived from an expert system based on empirical associations. Designing a belief network for the same hepatological domain required (re)consultation of the medical literature and renewed knowledge elicitation from experts. The development of a reliable belief network requires the availability of probabilities computed from a large series of real-life patients. Thus, in this study we encountered similar limitations as typically met in the construction of probabilistic systems, which were the main reasons for starting the HEPAR project in the first place.

Hybrid expert systems

In this thesis, we have tacitly assumed that either traditional statistical methods, such as logistic regression, or techniques used in the field of expert systems can be used for developing diagnostic systems. One of the nicest features of the engineering basis of expert systems is that it allows for the easy integration of various representations and reasoning techniques. Thus, it would be interesting to study the integration of the Pocket Diagnostic Chart and the HEPAR system within one system, again a subject of further research.

Bibliography

[Abu-Hanna & Jansweijer, 1994] A. Abu-Hanna and W. Jansweijer (1994). Modeling domain knowledge using explicit conceptualizations. *IEEE Expert*, **9**(5), 53–64.

[Adlassnig & Scheithauer, 1989] K.-P. Adlassnig and W. Scheithauer (1989). Performance evaluation of medical expert systems using ROC curves. *Computers and Biomedical Research*, **22**, 297–313.

[Adlassnig & Horak, 1995] K.-P. Adlassnig and W. Horak (1995). Development and retrospective evaluation of HEPAXPERT-I: a routinely used expert system for interpretive analysis of hepatitis A and B serologic findings. *Artificial Intelligence in Medicine*, **7**, 1–24.

[Aikins, 1980] J.S. Aikins (1980). *Prototypes and Production Rules: a Knowledge Representation for Computer Consultations*. PhD thesis, Report no. STAN-CS-80-814, Computer Science Department, Stanford University, Stanford, CA.

[Aikins, 1983] J.S. Aikins (1983). Prototypical knowledge for expert systems. *Artificial Intelligence*, **20**, 163–210.

[Albert & Jacques, 1993] P. Albert and G. Jacques (1993). Putting CommonKADS at work using KADStool. *Kennistechnologie'93*.

[Allemang et al., 1987] D. Allemang, M.C. Tanner, T. Bylander and J. Josephson (1987). Computational complexity of hypothesis assembly. *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 1112–1117.

[Alter et al., 1989] H.J. Alter, R.H. Russell, J.W. Shih et al. (1989). Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A non-B hepatitis. *New England Journal of Medicine*, **321**, 1494–1500.

[Andreassen et al., 1987] S. Andreassen, M. Woldbye, B. Falck and S.K. Andersen (1987). MUNIN — A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 366–372.

[Angele et al., 1994] J. Angele, D. Fensel and R. Studer (1994). The model of expertise in KARL. *Proceedings of the 2nd World Congress on Expert Systems*, Lisbon.

[Anjewierden et al., 1990] A. Anjewierden, J. Wielemaker and C. Toussaint (1990). Shelley – Computer aided knowledge engineering. In *Current Trends in Knowledge Acquisition* (B. Wielinga, J. Boose, B. Gaines, A.Th. Schreiber and M. van Someren, eds.), pp. 41–59. Amsterdam: IOS Press.

[Backhouse, 1986] R.C. Backhouse (1986). *Program Construction and Verification.* Englewood Cliffs, NJ: Prentice-Hall.

[Bankowitz et al., 1989] R.A. Bankowitz, M.A. McNeil, S.M. Challinor, R.C. Parker, W.N. Kapoor and R.A. Miller (1989). A computer-assisted medical diagnostic consultation service: implementation and prospective evaluation. *Annals of Internal Medicine*, **110**, 824–832.

[Barnett et al., 1987] G.O. Barnett, J.J. Cimino, J.A. Hupp and E.P. Hoffer (1987). DXplain – an evolving diagnostic decision-support system. *Journal of the American Medical Association*, **258**, 67–74.

[Benjamins, 1993] V.R. Benjamins (1993). *Problem Solving Methods for Diagnosis.* PhD thesis, University of Amsterdam.

[Benjamins & Jansweijer, 1994] V.R. Benjamins and W. Jansweijer (1994). Towards a competence theory of diagnosis. *IEEE Expert*, **9**(5), 43–52.

[Berkowitz, 1964] D. Berkowitz (1964). Pitfalls in the differential diagnosis of jaundice. *American Journal of Gastroenterology*, **41**, 488–498.

[Berner et al., 1994] E.S. Berner, G.D. Webster, A.A. Shuherman, J.R. Jackson, et al. (1994). Performance of four computer-based diagnostic systems. *The New England Journal of Medicine*, **330**(25), 1792-1796.

[Beschta et al., 1993] A. Beschta, O. Dressler, H. Freitag, M. Montag and P. Struss (1993). DP-Net – a second generation expert system for localizing faults in power transmission networks. In *Proceedings of the International Conference on Fault Diagnosis* (Tooldaig93), Toulouse, pp. 1019–1027.

[Bezem, 1988] M. Bezem (1985). Consistency of rule-based expert systems. In *Proceedings of the 9th International Conference on Automated Deduction* (E. Lusk and R. Overbeek, eds.), pp. 151–161. Berlin: Springer-Verlag.

[Birkhoff & Mac Lane, 1977] G. Birkhoff and S. Mac Lane (1977). *A Survey of Modern Algebra.* New York: Macmillan.

[Boehm, 1979] B.W. Boehm (1979). Software engineering: R & D trends and defense needs. In *Research Directions in Software Technology* (P. Wegner, ed.). Cambridge, Mass: MIT Press.

[Bratko et al., 1989] I. Bratko, I. Mozetič and N. Lavrač (1989). *KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems.* Cambridge, Massachusetts: The MIT Press.

[Breese et al., 1988] J.S. Breese, E.J. Horvitz and M. Henrion (1988). *Decision Theory in Expert Systems.* Research Report 3, Rockwell International Science Center, Palo Alto, CA.

[Breuker & Wielinga, 1989] J. Breuker and B. Wielinga (1989). Models of expertise in knowledge acquisition. In *Topics in Expert System Design* (G. Guida and C. Tasso, eds.), pp. 265–295. Amsterdam: North-Holland.

[Brown et al., 1982] J.S. Brown, D. Burton and J. de Kleer (1982). Pedagogical, natural language and engineering techniques in SOPHIE I, II and III. In *Intelligent Tutoring Systems* (D. Sleeman and J.S. Brown, eds.), pp. 227–282. New York: Academic Press.

[Buchanan & Shortliffe, 1984] B.G. Buchanan and E.H. Shortliffe (1984). *Rule-based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project.* Reading: Addison-Wesley.

[Burbank, 1969] F. Burbank (1969). A computer diagnostic system for the diagnosis of prolonged undifferentiated liver disease. *American Journal of Medicine*, **46**, 401–415.

[Bylander et al., 1992] T. Bylander, D. Allemang, M.C. Tanner and J.R. Josephson (1992). The computational complexity of abduction. In *Knowledge Representation* (R.J. Brachman, H.J. Levesque and R. Reiter, eds.), pp. 25–60. Cambridge, Massachusetts: The MIT Press.

[Campbell, 1976] E.J.M. Campbell (1976). Basic science, science and medical education. *Lancet*, **i**, 134–136.

[Carlstrom et al., 1963] E. Carlstrom, Y. Edlund, H.A. Hansen, K. Hugosson and N. Wedinius (1963). Hepatic tests in the differential diagnosis of jaundice. *Scandinavian Journal of Clinical and Laboratory Investigations*, **15**(3).

[Chandrasekaran & Mittal, 1983] B. Chandrasekaran and S. Mittal (1983). Conceptual representation of medical knowledge for diagnosis by computer: MDX and related systems. In *Advances in Computers* (M.C. Yovits, ed.), **22**. London: Academic Press.

[Clancey, 1985] W.J. Clancey (1985). Heuristic classification. *Artificial Intelligence*, **27**, 289–350.

[Clancey & Letsinger, 1984] W.J. Clancey and R. Letsinger (1984). NEOMYCIN: Reconfiguring a rule-based expert system for application to teaching. In *Readings in Medical Artificial Intelligence: the First Decade* (W.J. Clancey and E.H. Shortliffe, eds.). Reading, Massachusetts: Addison-Wesley.

[Clancey & Shortliffe, 1984] W.J. Clancey and E.H. Shortliffe, eds. (1984). *Readings in Medical Artificial Intelligence: the First Decade.* Reading, Massachusetts: Addison-Wesley.

[Clark, 1978] K.L. Clark (1978). Negation as failure. In *Logic and Databases* (H. Gallaire and J. Minker, eds.), pp. 293–322. New York: Plenum Press.

[Conn et al., 1979] H.O. Conn, A.T. Blei and M. Chojkier et al. (1979). The naked physician: the blind interpretation of liver function tests in the differential diagnosis of jaundice. In *The liver. Quantitative aspects of structure and function* (R. Preising and J. Bircher, eds.), pp. 386–394. Aulendorf: Editio Cantor.

[Console et al., 1989] L. Console, D. Theseider Dupré and P. Torasso (1989). A theory of diagnosis for incomplete causal models. *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 1311–1317.

[Console et al., 1991] L. Console, D. Theseider Dupré and P. Torasso (1991). On the relationship between abduction and deduction. *Journal of Logic and Computation*, **1**(5), 661–690.

[Console et al., 1992] L. Console, G. Molino, V. Ripa di Meana and P. Torasso (1992). LIED: liver information, education and diagnosis. *Methods of Information in Medicine*, **31**(4), 284–297.

[Console & Torasso, 1989] L. Console and P. Torasso (1989). A multi-level architecture for diagnostic problem solving. *Computational Intelligence*, **1**, 101–112.

[Console & Torasso, 1990a] L. Console and P. Torasso (1990). Hypothetical reasoning in causal models. *International Journal of Intelligent Systems*, **5**, 83–124.

[Console & Torasso, 1990b] L. Console and P. Torasso (1990). Integrating models of correct behaviour into abductive diagnosis. *Proceedings of ECAI'90*, pp. 160–166.

[Console & Torasso, 1991] L. Console and P. Torasso (1991). A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, **7**(3), 133–141.

[Cooper, 1990] G.F. Cooper (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, **42**, 393–405.

[Cox & Pietrzykowski, 1987] P.T. Cox and T. Pietrzykowski (1987). General diagnosis by abductive inference. *Proceedings of the IEEE Symposium on Logic Programming*, pp. 183–189.

[Cravetto et al., 1985] C. Cravetto, L. Lesmo, G. Molino and P. Torasso (1985). LITO2: a frame-based expert system for medical diagnosis in hepatology. In *Artificial Intelligence in Medicine* (I. de Lotto and M. Stefanelli, eds.). Amsterdam: North-Holland.

[Van Daalen, 1993] C. van Daalen (1993). *Validating Medical Knowledge Based Systems.* PhD thesis, Delft University of Technology.

[Dague, 1994] P. Dague (1994). Model-based diagnosis of analog electronic circuits. *Annals of Mathematics and Artificial Intelligence*, **11**, 439–492.

[Davis, 1980] R. Davis (1980). Meta-rules: reasoning about control. *Artificial Intelligence*, **15**, 179–222.

[Davis, 1984] R. Davis (1984). Diagnostic reasoning based on structure and behavior. *Artificial Intelligence*, **24**, 347–410.

[Davis & Hamscher, 1988] R. Davis and W. Hamscher (1988). Model-based reasoning: troubleshooting. In *Exploring Artificial Intelligence: Survey Talks from the National Conference on Artificial Intelligence* (H.E. Shrobe, ed.), pp. 297–346. San Mateo, California: Morgan Kaufmann.

[Davis & Lenat, 1982] R. Davis and D.B. Lenat (1982). *Knowledge-based Systems in Artificial Intelligence*. New York: McGraw-Hill.

[Davis & Shrobe, 1983] R. Davis and H. Shrobe (1983). Representing structure and behaviour of digital hardware. *IEEE Computer*, **16**(10), 75–82.

[Dijkstra, 1972] E.W. Dijkstra (1972). Notes on structured programming. In *Structured Programming* (O.-J. Dahl, E.W. Dijkstra and C.A.R. Hoare), pp. 1–82. London: Academic Press.

[De Dombal, 1984] F.T. de Dombal (1984). Computers and decision-making: an overview for gastroenterologists. *Frontiers in Gastrointestinal Research*, **7**, 119–133.

[De Dombal et al., 1972] F.T. de Dombal, D.J. Leaper, J.R. Staniland, A.P. McAnn and J.C. Horrocks (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal*, **ii**, 9–13.

[De Dombal et al., 1991] F.T. de Dombal, V. Dallos and W.A. McAdam (1991). Can computer-aided teaching packages improve clinical care in patients with acute abdominal pain? *British Medical Journal*, **302**, 1495–1497.

[Downing, 1993] K.L. Downing (1993). Physiological applications of consistency-based diagnosis. *Artificial Intelligence in Medicine*, **5**, 9–30.

[Duda et al., 1979] R.O. Duda, J. Gaschnig and P.E. Hart (1979). Model design in the PROSPECTOR consultant program for mineral exploration. In *Expert Systems in the Microelectronic Age* (D. Michie, ed.). Edinburgh: Edinburgh University Press.

[EURICTERUS, 1993] EURICTERUS Project Management Group (1993). Objective medical decision-making: clinical database for diagnosis of jaundice (EURICTERUS). In *Advances in Biomedical Engineering* (J.E.W. Beneken and V. Thévenin, eds.), pp. 35–64. Amsterdam: IOS Press.

[Fensel & Van Harmelen, 1994] D. Fensel and F. van Harmelen (1994). A comparison of languages which operationalise and formalise KADS models of expertise. *The Knowledge Engineering Review*, **9**(2), 105–146.

[Fensel, 1995] D. Fensel (1995). *The Knowledge Acquisition and Representation Language, KARL*. Boston: Kluwer.

[First et al., 1982] M.B. First, B.J. Weimer, S. McLinden and R.A. Miller (1982). LOCALIZE: computer-assisted localization of peripheral nervous lesions. *Computers and Biomedical Research*, **15**(6), 525–543.

[François et al., 1990] P. François, C. Robert, B. Cremilleux, C. Bucharles and J. Demongeot (1991). Variables processing in expert system building: application to the aetiological diagnosis of infantile meningitis. *Methods of Information in Medicine*, **15**(2), 115–124.

[François et al., 1992] P. François, B. Cremilleux, C. Robert and J. Demongeot (1992). MENINGE: a medical consulting system for child's meningitis: study on a series of consecutive cases. *Artificial Intelligence in Medicine*, **4**, 281–292.

[Van der Gaag, 1989] L.C. van der Gaag (1989). A conceptual model for inexact reasoning in rule-based systems. *International Journal of Approximate Reasoning*, **3**(3), 239–258.

[Van der Gaag, 1990] L.C. van der Gaag (1990). *Probability-based Models for Plausible Reasoning*. PhD thesis, University of Amsterdam.

[Van der Gaag, 1994] L.C. van der Gaag (1994). A pragmatic view of the certainty factor model. *The International Journal of Expert Systems: Research and Applications*, **7**(3), 289–300.

[Genesereth, 1984] M.R. Genesereth (1984). The use of design descriptions in automated diagnosis. *Artificial Intelligence*, **24**, 411-436.

[Genesereth & Nilsson, 1987] M.R. Genesereth and N.J. Nilsson (1987). *Logical Foundations of Artificial Intelligence*. Palo Alto, CA: Morgan Kaufmann.

[Goldberg & Ellis, 1978] D.M Goldberg and G. Ellis (1978). Mathematical and computer-assisted procedures in the diagnosis of liver and biliary tract disorders. *Advances in Clinical Chemistry*, **20**, 49–128.

[Gorry & Barnett, 1968] G.A. Gorry and G.O. Barnett (1968). Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, **1**, 490–507.

[Graham et al., 1982] S.L. Graham, P.B. Kessler and M.K. McKusick (1982). Gprof: a call graph execution profiler. *Proceedings of the SIGPLAN'82 Symposium on Compiler Construction*, SIGPLAN Notices 17, pp. 120–126.

[Guida & Tasso, 1989] G. Guida and C. Tasso (1989). Building expert systems: from life cycle to development methodology. In *Topics in Expert System Design* (G. Guida and C. Tasso, eds.). Amsterdam: North-Holland.

[Habbema et al., 1978] J.D.F. Habbema, J. Hilden. and B. Bjerregaard (1978). The measurement of performance in probabilistic diagnosis I: the problem, descriptive tools, and measures based on classification matrices. *Methods of Information in Medicine*, **17**, 217–226.

[Hamscher, 1994] W. Hamscher (1994). CROSBY: financial data interpretation as model-based diagnosis. *Annals of Mathematics and Artificial Intelligence*, **11**, 511–524.

[Van Harmelen & Balder, 1992] F. van Harmelen and J. Balder (1992). (ML)^2: a formal language for KADS models of expertise. *Knowledge Acquisition*, **4**, 127–161.

[Haubek et al., 1981] A. Haubek, J.H. Pedersen and F. Burcharth et al. (1981). Dynamic sonography in the evaluation of jaundice. *American Journal of Radiology*, **136**, 1071–1074.

[Hayes, 1977] P. Hayes (1977). In defense of logic. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pp. 559–565.

[Heckerling et al., 1991] P.S. Heckerling, A.S. Elstein, C.G. Terzian and M.S. Kushner (1991). The effect of incomplete knowledge on the diagnosis of a computer consultant system. *Medical Informatics*, **16**(4), 363–370.

[Heckerman, 1990] D.E. Heckerman (1990). An empirical comparison of three inference methods. In *Uncertainty in Artificial Intelligence 4* (R. Shachter, T. Levitt, L. Kanal and J. Lemmer, eds.), pp. 283–302. New York: North-Holland.

[Heckerman, 1992] D.E. Heckerman (1992). *Probabilistic Similarity Networks*. Cambridge, Massachusetts: The MIT Press.

[Heckerman et al., 1992] D.E. Heckerman, E.J. Horvitz and B.N. Nathwani (1990). Towards normative expert systems: part I – The Pathfinder project. *Methods of Information in Medicine*, **31**, 90–105.

[Heckerman & Nathwani, 1992] D.E. Heckerman and B.N. Nathwani (1992). Towards normative expert systems: part II – probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in Medicine*, **31**, 106–116.

[Hilden & Habbema, 1990] J. Hilden and J.D.F. Habbema (1990). Evaluation of clinical decision aids – more to think about. *Medical Informatics*, **15**, 275–284.

[Hickam et al., 1985] D.H. Hickam, E.H. Shortliffe and M.B. Bishoff et al. (1985). The treatment advice of a computer-based cancer chemotherapy protocol advisor. *Annals of Internal Medicine*, **103**, 928–936.

[Hou, 1994] A. Hou (1994). A theory of measurement in diagnosis from first principles. *Artificial Intelligence*, **65**, 281–328.

[Indurkhya & Weiss, 1989] N. Indurkhya and S.M. Weiss (1989). Models for measuring performance of medical expert systems. *Artificial Intelligence in Medicine*, **1**, 61–70.

[Jackson et al., 1990] P. Jackson, H. Reichgelt and F. van Harmelen (1990). *Logic-based Knowledge Representation*. Cambridge, Massachusetts: The MIT Press.

[Jaspers, 1990] R. Jaspers (1990). *Medical Decision Support: An Approach in the Domain of Brachial Plexus Injuries*. PhD thesis, Delft University of Technology.

[Johnson & Keravnou, 1988] L. Johnson and E.T. Keravnou (1988). *Expert Systems Architectures.* London: Kogan Page.

[Josephson & Josephson, 1994] J.R. Josephson and S.G. Josephson (1994). *Abductive Inference: computation, philosophy, technology.* Cambridge: Cambridge University Press.

[Karbach et al., 1990] W. Karbach, M. Linster and A. Voss (1990). Models of problem-solving: one label – one idea? In *Current Trends in Knowledge Acquisition* (B. Wielinga, J. Boose, B. Gaines, A.Th. Schreiber and M. van Someren, eds.), pp. 173–189. Amsterdam: IOS Press.

[Karran et al., 1985] S. Karran, K.C. Dewburg, A.E.A. Joseph and R. Wright (1985). Investigation of the jaundiced patient. In *Liver and Biliary Disease* 2nd ed. (R. Wright, G.H. Millward-Sadler, K.G.M.M. Alberti and S. Karran, eds.), pp. 647–658. London: Saunders.

[Karran & McLaren, 1985] S. Karran and M. McLaren (1985). Physical aspects of hepatic regeneration. In *Liver and Biliary Disease* (R. Wright, G.H. Millward-Sadler, K.G.M.M. Alberti and S. Karran, eds.), pp. 233–250. London: Saunders.

[Kassirer, 1994] J.P. Kassirer (1994). A report card on computer-assisted diagnosis – the grade: C. *The New England Journal of Medicine*, **330**(25), 1824–1825.

[De Kleer, 1977] J. de Kleer (1977). Multiple representation of knowledge in mechanic problem solving. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pp. 299–304.

[De Kleer & Williams, 1987] J. de Kleer and B.C. Williams (1987). Diagnosing multiple faults. *Artificial Intelligence*, **32**, 97–130.

[De Kleer & Williams, 1989] J. de Kleer and B.C. Williams (1989). Diagnosis with behavioural modes. *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pp. 1324–1330.

[De Kleer et al., 1992] J. de Kleer, A.K. Mackworth and R. Reiter (1992). Characterizing diagnoses and systems. *Artificial Intelligence*, **52**, 197–222.

[Knill-Jones et al., 1973] R.P. Knill-Jones, R.B. Stern, D.H. Girmes, J.D. Maxwell, R.P.H. Thompson, and R. Williams (1973). Use of a sequential Bayesian model in the diagnosis of jaundice. *British Medical Journal*, **i**, 530–533.

[Konolige, 1992] K. Konolige (1992). Using default and causal reasoning in diagnosis. *Proceedings of the Workshop on Principles of Knowledge Representation and Reasoning*, Boston.

[Konolige, 1994] K. Konolige (1994). Using default and causal reasoning in diagnosis. *Annals of Mathematics and Artificial Intelligence*, **11**, 97–135.

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

[Krause et al., 1995] P. Krause, S. Ambler, M. Elvang-Coransson and J. Fox (1995). A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, **11**(1), 113–131.

[Kreyszig, 1970] E. Kreyszig (1970). *Introduction to Mathematical Statistics: Principles and Methods*. New York: John Wiley & Sons.

[Kulikowski & Weiss, 1982] C.A. Kulikowski and S.M. Weis (1982). Representation of expert knowledge for consultation: the CASNET and EXPERT projects. In *Artificial Intelligence in Medicine* (P. Szolovits, ed.), pp. 21-56. Boulder: Westview Press.

[Van Langevelde et al., 1993] I. van Langevelde, A. Philipsen and J. Treur (1993). A compositional architecture for simple design formally specified in DESIRE. In *Formal Specification of Complex Reasoning Tasks* (J. Treur and T. Wetter, eds.), pp. 143–172. New York: Ellis Horwood.

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

[Lesmo et al., 1984] L. Lesmo, M. Marzuoli, G. Molino and P. Torasso (1984). An expert system for the evaluation of liver functional assessment. *Journal of Medical Systems*, **8**, 87–101.

[Lindberg et al., 1987] G. Lindberg, C. Thomson, A. Malchow-Møller, P. Matzen and J. Hilden (1987). Differential diagnosis of jaundice: applicability of the Copenhagen Pocket Diagnostic Chart proven in Stockholm patients. *Liver*, **7**, 43–9.

[Lucas, 1986] P.J.F. Lucas (1986). *Knowledge Representation and Inference in Rule-based Systems*. Report CS-R8613, Centre for Mathematics and Computer Science, Amsterdam.

[Lucas, 1993] P.J.F. Lucas (1993). The representation of medical reasoning models in resolution-based theorem provers. *Artificial Intelligence in Medicine*, **5**(5), 395–41.

[Lucas, 1994] P.J.F. Lucas (1994). Refinement of the HEPAR expert system: tools and techniques. *Artificial Intelligence in Medicine*, **6**(2), 175–188.

[Lucas et al., 1989] P.J.F. Lucas, R.W. Segaar and A.R. Janssens (1989). HEPAR: an expert system for the diagnosis of disorders of the liver and biliary tract. *Liver*, **9**, 266–275.

[Lucas & Janssens, 1991a] P.J.F. Lucas and A.R. Janssens (1991). Development and validation of HEPAR, an expert system for the diagnosis of disorders of the liver and biliary tract. *Medical Informatics*, **16**, 259–270.

[Lucas & Janssens, 1991b] P.J.F. Lucas and A.R. Janssens (1991). Second evaluation of HEPAR, an expert system for the diagnosis of disorders of the liver and biliary tract. *Liver*, **11**, 340–346.

[Lucas & Van der Gaag, 1991] P.J.F. Lucas and L.C. van der Gaag (1991). *Principles of Expert Systems*. Wokingham: Addison-Wesley.

[Lucas & De Swaan Arons, 1987] P.J.F. Lucas and H. de Swaan Arons (1987). Extensions to the expert-system shell DELFI-2. In *Expert Systems and Artificial Intelligence in Decision Support* (H.G. Sol, C.A.Th. Takkenberg and P.F. de Vries Robbé, eds.), pp. 213–225. Reidel: Dordrecht.

[Łukaszewicz, 1990] W. Łukaszewicz (1990). *Non-monotonic reasoning*. New York: Ellis Horwood.

[Macartney, 1988] F.J. Macartney (1988). Diagnostic logic. In *Logic in Medicine*. (C. Philips, ed.). London: British Medical Journal.

[Malchow-Møller et al., 1986] A. Malchow-Møller, C. Thomson, P. Matzen et al. (1986). Computer diagnosis in jaundice: Bayes' rule founded on 1002 consecutive cases. *Journal of Hepatology*, **3**, 154–163.

[Malchow-Møller et al., 1987] A. Malchow-Møller, L. Mindeholm, H.S. Rasmussen and B. Rasmussen et al. (1987). Differential diagnosis of jaundice: junior staff experience with the Copenhagen pocket chart. *Liver*, **7**, 333–338.

[Matzen et al., 1984] P. Matzen, A. Malchow-Møller, J. Hilden et al. (1984). Differential diagnosis of jaundice: a pocket diagnostic chart. *Liver*, **4**, 360–71.

[Martin et al., 1960] W.B. Martin, P.C. Apostolakos and H. Roazen (1960). Clinical versus actuarial prediction in the differential diagnosis of jaundice. *American Journal of Medical Science*, **240**, 571–578.

[McCarthy, 1986] J. McCarthy (1986). Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, **28**, 89–116.

[McIlraith & Reiter, 1992] S. McIlraith and R. Reiter (1992). On tests for hypothetical reasoning. In *Readings in Model-based Diagnosis* (W. Hamscher, L. Console and J. de Kleer, eds.), pp. 89–96. San Mateo: Morgan Kaufmann.

[McIntyre, 1986] N. McIntyre (1986). Computer-aided diagnosis in jaundice and liver disease. *Journal of Hepatology*, **3**, 269–272.

[Van Melle, 1980] W. van Melle (1980). *A Domain Independent System that Aids in Constructing Knowledge-based Consultation Programs*. PhD thesis, Report STAN-CS-80-820, Computer Science Department, Stanford University.

[Van Melle et al., 1981] W. van Melle, A.C. Scott, J.S. Bennett and M. Eairs (1981). *The EMYCIN Manual.* Report STAN-CS-81-16, Computer Science Department, Stanford University.

[Milanese & Bona, 1984] M. Milanese, B. Bona (1984). A sequential approach for the optimization of diagnostic procedures in hepatology. *Journal of Medical Systems*, **8**, 73–85.

[Miller et al., 1982] R.A. Miller, H.E. Pople and J.D. Myers (1982). INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, **307**, 468–476.

[Miller & Masarie, 1990] R.A. Miller and F.E. Masarie (1990). The demise of the "Greek Oracle" model for medical diagnostic systems. *Methods of Information in Medicine*, **29**(1), 1–2.

[Motta et al., 1989] E. Motta, T. Rajan and M. Eisenstadt (1989). A methodology and tool for knowledge acquisition in KEATS-2. In *Topics in Expert System Design* (G. Guida and C. Tasso, eds.), pp. 297-322. Amsterdam: North-Holland.

[Motta et al., 1990] E. Motta, T. Rajan, J. Domingue and M. Eisenstadt (1990). Methodological foundation of KEATS, the knowledge engineer's assistant. In *Current Trends in Knowledge Acquisition* (B. Wielinga, J. Boose, B. Gaines, A.Th. Schreiber and M. van Somere, eds.), pp. 257-275. Amsterdam: IOS Press.

[Mozetič, 1992] I. Mozetič (1992). Hierarchical model-based diagnosis. In *Readings in Model-based Diagnosis* (W. Hamscher, L. Console and J. de Kleer, eds.), pp. 354–372. San Mateo: Morgan Kaufmann.

[Newell & Simon, 1972] A. Newell and H.A. Simon (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

[Ng, 1991] H.T. Ng (1991). Model-based, multiple fault diagnosis of dynamic, continuous physical devices. *IEEE Expert*, **6**(6), 38–43.

[O'Keefe et al., 1987] R.M. O'Keefe, O. Balci and E.P. Smith (1987). Validating expert system performance. *IEEE Expert*, **4**, 81–89.

[Patil, 1981] R.S. Patil (1981). *Causal representation of patient illness for electrolyte and acid-base diagnosis.* Technical Report MIT/LCS/TR-267, Massachusetts Institute of Technology.

[Patil et al., 1982] R.S. Patil, P. Szolovits and W.B. Schwartz (1982). Modeling knowledge of the patient in acid-base and electrolyte disorders. In *Artificial Intelligence in Medicine* (P. Szolovits, ed.). Boulder, CO: Westview Press.

[Pearl, 1988] J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

[Peng, 1985] Y. Peng (1985). *A Formalization of Parsimonious Covering and Probability Inference.* PhD thesis, Department of Computer Science, University of Maryland.

[Peng & Reggia, 1990] Y. Peng and J.A. Reggia (1990). *Abductive Inference Models for Diagnostic Problem Solving.* New York: Springer-Verlag.

[Pereira et al., 1991] B.J. Pereira, E.L. Milford, R.L. Kirkman and A.S. Levey (1991). Transmission of hepatitis C virus by organ transplantation. *New England Journal of Medicine*, **325**, 454–460.

[Politakis, 1985] P.G. Politakis (1985). *Empirical Analysis for Expert Systems.* London: Pitman.

[Poole et al., 1987] D. Poole, R. Goebel and R. Aleliunas (1987). Theorist: a logical reasoning system for defaults and diagnosis. In *The Knowledge Frontier* (N. Cercone and G. Mc Calla, eds.), pp. 331–352. Berlin: Springer-Verlag.

[Poole, 1988] D. Poole (1988). Representing knowledge for logic-based diagnosis. *Proceedings of the International Conference on Fifth Generation Computer Systems 1988*, pp. 1282–1290. ICOT.

[Poole, 1989] D. Poole (1989). Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, **5**(2), 97–110.

[Poole, 1990a] D. Poole (1990). Normality and faults in logic-based diagnosis. *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pp. 1304–1310.

[Poole, 1990b] D. Poole (1990). A methodology for using a default and abductive reasoning system. *International Journal of Intelligent Systems*, **5**(5), 521–548.

[Poole, 1994] D. Poole (1994). Representing diagnosis knowledge. *Annals of Mathematics and Artificial Intelligence*, **11**, 33–50.

[Pople, 1973] H.E. Pople (1973). On the mechanization of abductive logic. *Proceedings of the 3rd International Joint Conference on Artificial Intelligence.*

[Pople, 1977] H.E. Pople (1977). The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pp. 1030–1037.

[Popper, 1959] K.R. Popper (1959). *The Logic of Scientific Discovery.* London: Hutchingson.

[Preist et al., 1994] C. Preist, K. Eshghi and B. Bertolino (1994). Consistency-based and abductive diagnosis as generalized stable models. *Annals of Mathematics and Artificial Intelligence*, **11**, 51–74.

[Price & Alberti, 1985] P.C. Price and K.G.M.M. Alberti (1985). Biochemical assessment of liver function. In *Liver and Biliary Disease* 2nd ed. (R. Wright, G.H. Millward-Sadler, K.G.M.M. Alberti and S. Karran, eds.), pp. 455–493. London: Saunders.

[Punch III et al., 1990] W.F. Punch III, M.C. Tanner, J.R. Josephson and J.W. Smith (1990). PEIRCE: a tool for experimenting with abduction. *IEEE Expert*, **5**(5), 34–44.

[Quaglini et al., 1988] S. Quaglini, M. Stefanelli, G. Barosi and A. Berzuini (1988). A performance evaluation of the expert system ANEMIA. *Computer and Biomedical Rsearch*, **21**, 307–323.

[Reggia et al., 1983] J.A. Reggia, D.S. Nau and Y. Wang (1983). Diagnostic expert systems based on a set-covering model. *International Journal of Man-Machine Studies*, **19**, 437–460.

[Reiter, 1977] R. Reiter (1977). On closed world databases. In *Logic and Databases* (H. Gallaire and J. Minker, eds.). Berlin: Springer-Verlag.

[Reiter, 1980] R. Reiter (1980). A logic for default reasoning. *Artificial Intelligence*, **13**, 81–132.

[Reiter, 1987] R. Reiter (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, **32**, 57–95.

[Robbins et al., 1994] S.L. Robbins, R.S. Cotran and V. Kumar (1994). *Robbins Pathologic Basis of Disease*. Philadelphia: W.B. Saunders.

[Sauthier & Faltings, 1992] E. Sauthier and B. Faltings (1992). Model-based traffic control. *Artificial Intelligence in Engineering*, **7**, 139–151.

[Schenker et al., 1962] S. Schenker, J. Balint, L. Schiff (1962). Differential diagnosis of jaundice: Report of a prospective study of 61 proved cases. *American Journal of Digestive Disease*, **7**, 449–463.

[Schmitz, 1986] P.I.M. Schmitz (1986). *Logistic Regression in Medical Decision Making and Epidemiology.* PhD thesis, Eramus University Rotterdam.

[Schreiber et al., 1994] A.Th. Schreiber, B. Wielinga, R. de Hoog, H. Akkermans and W. van der Velde (1994). CommonKADS: a comprehensive methodology for KBS development. *IEEE Expert*, **9**(6), 28–37.

[Segaar, 1991] R.W. Segaar (1991). *Decision Support for the Differential Diagnosis of Jaundice.* PhD thesis, Erasmus University Rotterdam.

[Segaar et al., 1988] R.W. Segaar, J.H.P. Wilson, J.D.F. Habbema, A. Malchow-Møller, J. Hilden and P.J. van der Maas (1988). Transferring a diagnostic aid for jaundice. *Netherlands Journal of Medicine*, **4**, 360–71.

[Shannon, 1975] R.E. Shannon (1975). *Systems Simulation: the art and the science.* Englewood Cliffs, NJ: Prentice-Hall.

[Shortliffe, 1976] E.H. Shortliffe (1976). *Computer-based Medical Consultations: MYCIN*. New York: Elsevier.

[Shortliffe, 1989] E.H. Shortliffe (1989). Testing reality: the introduction of decision-support technologies for physicians. *Methods of Information in Medicine*, **28**, 1–5.

[Shortliffe, 1991] E.H. Shortliffe (1991). Knowledge-based systems in medicine. In *Medical Informatics Europe 1991* (K.-P. Adlassnig, G. Grabner, S. Bengtsson and R. Hansen, eds.), pp. 5–9. Berlin: Springer-Verlag.

[Silverman, 1975] H. Silverman (1975). *A Digitalis Therapy Advisor, Project MAC.* Technical Report MIT/MAC/TR-143, Massachusetts Institute of Technology.

[Smith et al., 1985] J.W. Smith, J.R. Svirbely, C.A. Evans, P. Strohm, J.R. Josephson and M.C. Tanner (1985). RED: a red-cell antibody identification expert module. *Journal of Medical Systems*, **9**(3), 121–138.

[Sommerville, 1992] I. Sommerville (1992). *Software Engineering.* Wokingham: Addison-Wesley.

[Spiegelhalter & Knill-Jones, 1984] D.J. Spiegelhalter and R.P. Knill-Jones (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society*, **147**, 35–77.

[Spiegelhalter et al., 1990] D.J. Spiegelhalter, R.C.G. Franklin and K. Bull (1990). Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. In *Uncertainty in Artificial Intelligence 5* (M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds.), pp. 285–294. Amsterdam: North-Holland.

[Stefanini et al., 1993] A. Stefanini, G. Tornielli and S. Cermignani (1993). An interval-based model for fault diagnosis in power transmission nets. *Proceedings of the International Conference on Fault Diagnosis* (Tooldaig93), pp. 1000–1008. Toulouse.

[Stern et al., 1974] R.B. Stern, R.P. Knill-Jones and R. Williams (1974). Clinician versus computer in the choice of 11 differential diagnoses of jaundice based on formalised data. *Methods of Information in Medicine*, **13**, 79–82.

[Stern et al., 1975] R.B. Stern, R.P. Knill-Jones and R. Williams (1975). Use of computer program for diagnosing jaundice in district hospitals and specialized liver units. *British Medical Journal*, **ii**, 659–662.

[Sticklen & Chandrasekaran, 1985] J. Sticklen and B. Chandrasekaran (1985). Control issues in classificatory diagnosis. *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pp. 300–306.

[Sticklen & Chandrasekaran, 1988] J. Sticklen and B. Chandrasekaran (1988). MDX2: an integrated medical diagnostic system. *Proceedings of the AAAI Symposium on Artificial Intelligence in Medicine*, pp. 90–95.

[Struss, 1992] P. Struss (1992). What is in SD? Towards a theory of modelling for diagnosis. In *Readings in Model-based Diagnosis* (W. Hamscher, L. Console and J. de Kleer, eds.), pp. 419–449. San Mateo: Morgan Kaufmann.

[Struss & Dressler, 1989] P. Struss and O. Dressler (1989). Physical negation: integrating fault models into the general diagnostic engine. *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pp. 1318–1223.

[Szolovits, 1982] P. Szolovits (1982). *Artificial Intelligence in Medicine*. Boulder, CO: Westview Press.

[Szolovits et al., 1988] P. Szolovits, R.S. Patil and W.B. Schwartz (1988). Artificial intelligence in medical diagnosis. *Annals of Internal Medicine*, **108**, 80–87.

[Szolovits & Pauker, 1978] P. Szolovits and S.G. Pauker (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, **11**, 115–144.

[Ten Teije & Van Harmelen, 1994] A. ten Teije and F. van Harmelen (1994). An extended spectrum of logical definitions for diagnostic systems. In *DX-94, 5th International Workshop on Principles of Diagnosis* (G.M. Provan, ed.), pp. 334–342.

[Theodossi et al., 1981] A. Theodossi, R.P. Knill-Jones, A. Skene and G. Lindberg et al.. (1981). Inter-observer variation of symptoms and signs in jaundice. *Liver*, **1**, 21–32.

[Theodossi et al., 1983] A. Theodossi, D.J. Spiegelhalter and B. Portmann et al. (1983). The value of clinical, biochemical, ultrasound and liver biopsy data in assessing patients with liver disease. *Liver*, **3**, 315–326.

[Theodossi, 1986] A. Theodossi (1986). *An Assessment of the Value of Diagnostic Techniques in Hepatobiliary Disease*. M.D. Thesis, University of London.

[Todd & Stamper, 1993] B.S. Todd and R. Stamper (1993). *The Formal Design and Evaluation of a Variety of Medical Diagnostic Programs*. Technical Monograph PRG-109, Oxford University Computing Laboratory, Oxford University.

[Torasso & Console, 1989] P. Torasso and L. Console (1989). *Diagnostic Problem Solving*. London: North Oxford Academic Publishers.

[Treur, 1993] J. Treur (1993). Heuristic reasoning and relative incompleteness. *International Journal of Approximate Reasoning*, **8**, 51–87.

[Treur & Wetter, 1993] J. Treur and T. Wetter (eds.) (1993). *Formal Specification of Complex Reasoning Tasks.* New York: Ellis Horwood.

[Tuhrim et al., 1991] S. Tuhrim, J. Reggia and S. Goodall (1991). An experimental study of criteria for hypothesis plausibility. *Journal of Experimental and Theoretical Artificial Intelligence*, **3**, 129–144.

[Warner, 1989] H.R. Warner (1989). Iliad: moving medical decision making into new frontiers. *Methods of Information in Medicine*, **28**, 370–372.

[Waxman & Worley, 1990] H.S. Waxman and W.E. Worley (1990). Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. *Medicine*, **69**, 125–136.

[Weiss et al., 1978] S. Weiss, C. Kulikowski and A. Safir (1978). A model-based method for computer-aided medical decision making. *Artificial Intelligence*, **11**, 145–172.

[Wielinga et al., 1992] B.J. Wielinga, A.Th. Schreiber and J.A. Breuker (1992). KADS: a modelling approach to knowledge engineering. *Knowledge Acquisition*, **4**, pp. 5–53.

[Wielinga et al., 1993] B.J. Wielinga, A.Th. Schreiber and J.A. Breuker (1993). Towards a unification of knowledge-modeling approaches. In *Second Generation Expert Systems* (J.-M. David, J.-P. Krivine and R. Simmons, eds.), pp. 299–335. New York: Springer Verlag.

[Wirth, 1971] N. Wirth (1971). Program development by stepwise refinement. *Communications of the ACM*, **14**(4), 221–227.

[Wright et al., 1985a] R. Wright, G.H. Millward-Sadler, K.G.M.M. Alberti, and S. Karran, eds. (1985). *Liver and Biliary Disease* 2nd ed. London: Saunders.

[Wright et al., 1985b] R. Wright, G.H. Millward-Sadler and F.G. Bull (1985). Acute viral hepatitis. In *Liver and Biliary Disease* 2nd ed. (R. Wright, G.H. Millward-Sadler, K.G.M.M. Alberti and S. Karran, eds.), pp. 679–763. London: Saunders.

[Wu, 1991] T.D. Wu (1991). A problem decomposition method for efficient diagnosis and interpretation of multiple disorders. *Computer Methods and Programs in Biomedicine*, **35**, 239–250.

[Wulff, 1981] H.R. Wulff (1981). *Rational Diagnosis and Treatment: an Introduction to Clinical Decision-making* 2nd ed. Oxford: Blackwell.

[Wyatt & Spiegelhalter, 1990] J. Wyatt and D.J. Spiegelhalter (1990). Evaluating medical expert systems: what to test for and how? *Medical Informatics*, **15**(3), 205–217.

[Yu et al., 1979a] V.L. Yu, L.M. Fagan and S.M. Wraiht et al. (1979). Antimicrobial selection by a computer: a blinded evaluation by infectious disease experts. *JAMA*, **242**, 1279–1282.

[Yu et al., 1979b] V.L. Yu, B.G. Buchanan, E.H. Shortliffe, et al. (1979). An evaluation of the performance of a computer-based consultant. *Computer Programs in Biomedicine*, **9**, 95–102.

[Zadrozny, 1993] W. Zadrozny (1993). On rules of abduction. *Annals of Mathematics and Artificial Intelligence*, **9**, 387–419.

Samenvatting

Diagnose kan worden opgevat als een proces van informatievergaring, dat tot doel heeft vast te stellen welke aandoening een patiënt heeft, of welke component van een technisch apparaat defect is. Informatievergaring wordt wel als het *dynamische* aspect van diagnose beschouwd; de interpretatie van de vergaarde informatie voor het stellen van een diagnose wordt opgevat als het *statische* aspect van diagnose. In het eerste deel van dit proefschrift worden deze statische aspecten van diagnose bestudeerd. De dynamische aspecten van diagnose komen slechts zijdelings in dit proefschrift aan de orde. In het tweede deel van het proefschrift wordt de ontwikkeling, implementatie en toetsing van het HEPAR-systeem, een expertsysteem voor de diagnose van aandoeningen van de lever en galwegen, besproken.

In Hoofdstuk 2 wordt ingegaan op de belangrijkste conceptuele modellen en de hieruit voortgekomen formele theorieën van diagnose. Diagnostische systemen worden vaak ontwikkeld op grond van ervaringskennis, maar ook modellen van de structuur en het gedrag van fysieke systemen kunnen als basis voor de ontwikkeling dienen. Drie conceptuele modellen van diagnose kunnen nu worden onderscheiden. Bij DNSB-diagnose (diagnose op basis van 'Deviation from Normal Structure and Behaviour') verklaart een diagnose een geobserveerd verschil tussen voorspeld correct gedrag en geobserveerd gedrag. Een MAB-diagnose (een diagnose op grond van 'Matching Abnormal Behaviour') verklaart een overeenstemming tussen voorspeld, foutief gedrag en geobserveerd gedrag. Tenslotte, AC-diagnose (diagnose op grond van 'Abnormality Classification') geeft een verklaring van geobserveerde feiten in termen van een classificatie. Een formeel analogon van DNSB-diagnose is 'consistentie-gebaseerde diagnose': het verschil tussen voorspeld correct gedrag en geobserveerd gedrag wordt geformaliseerd met behulp van het logische begrip 'onvervulbaarheid'. MAB-diagnose wordt formeel beschreven in termen van abductie: overeenstemming tussen voorspeld, afwijkend gedrag en geobserveerd gedrag wordt geformaliseerd met behulp van de 'logisch gevolg' operator. Een iets directere, maar minder expressieve formalisering wordt geboden door de 'set-covering' theorie. AC-diagnose kan ook worden geformaliseerd met behulp van de 'logisch gevolg' operator. Eerder onderzoek heeft uitgewezen dat de uitdrukkingskracht van de diverse logische theorieën voldoende groot is voor het tot uitdrukking brengen van andere conceptuele modellen dan waar ze oorspronkelijk voor ontworpen zijn. Dit levert echter onnatuurlijk overkomende formaliseringen op, waarin de oorspronkelijke diagnosemodellen prominent aanwezig blijven. De slotsom van de analyse in het proefschrift is dat de bovengenoemde formele theorieën van diagnose maar beperkt bruikbaar zijn als algemene semantische raamwerken voor diagnose.

In Hoofdstuk 3 wordt een algemeen verzameling-theoretisch, semantisch raamwerk voor diagnose ontwikkeld. In dit raamwerk wordt een onderscheid gemaakt tussen: (1) de interpretatie van kennis als diagnose-kennis; (2) de interpretatie van geobserveerde feiten, samen met diagnose-kennis, volgens een gegeven diagnosebegrip; (3) de selectie van diagnosen op grond van bepaalde criteria. Verschillende eigenschappen van deze aspecten worden bestudeerd, en in Hoofdstuk 4 toegepast op uit de literatuur bekende diagnosebegrippen.

De diagnosebegrippen in de literatuur kunnen meestal alleen worden toegepast als een kennismodel een volledige, en correcte representatie van de realiteit biedt. Bij de ontwikkeling van een expertsysteem is het niet altijd mogelijk aan deze voorwaarden te voldoen. In Hoofdstuk 5 worden enkele diagnosebegrippen gedefinieerd die tot doel hebben ook een diagnose te kunnen stellen bij imperfecte kennis.

In het tweede deel van het proefschrift wordt de ontwikkeling van het HEPAR-systeem besproken. Dit systeem kan worden beschouwd als een praktische realisering van de in het eerste deel van het proefschrift besproken AC-diagnosevorm. In Hoofdstuk 6 wordt een overzicht gepresenteerd van de medisch-biologische kennis die een rol speelt bij het stellen van een diagnose op het gebied van leverziekten. De door de clinicus gevolgde strategie wordt besproken, alsmede enkele typische ziektebeelden. In Hoofdstuk 7 wordt de structuur van de kennisbank van het HEPAR-systeem besproken. Bij de ontwikkeling van het systeem is gebruik gemaakt van een iteratieve methode van verfijning van de kennisbank. De hiertoe ontwikkelde hulpmiddelen hebben ook een rol gespeeld bij de toetsing van het systeem, zoals besproken in Hoofdstuk 8. Bij de toetsing is gebruik gemaakt van twee bestanden met patiëntgegevens. De voornaamste conclusie is dat het systeem niet onderdoet voor de beste beslissingsondersteunende computersystemen op het gebied van leverziekten, maar een veel specifiekere, en dus informatievere diagnose kan stellen dan deze andere systemen.

Tenslotte, in Hoofdstuk 9 worden de belangrijkste beperkingen en wenselijke uitbreidingen van het diagnostisch raamwerk geïdentificeerd. Enkele onderwerpen voor vervolgonderzoek worden gesuggereerd.