Computational Intelligence, Volume 17, Number 1, 2001

ABDUCTIVE DIAGNOSIS USING TIME-OBJECTS: CRITERIA FOR THE EVALUATION OF SOLUTIONS

ELPIDA T. KERAVNOU
Department of Computer Science, University of Cyprus

JOHN WASHBROOK
Department of Computer Science, University College London

Diagnostic problem solving aims to account for, or explain, a malfunction of a system (human or other). Any plausible potential diagnostic solution must satisfy some minimum criteria relevant to the application. Often there will be several plausible solutions, and further criteria will be required to select the "best" explanation. Expert diagnosticians may employ different, complex criteria at different stages of their reasoning. These criteria may be combinations of some more primitive criteria, which therefore should be represented separately and explicitly to permit their flexible and transparent combined usage.

In diagnostic reasoning there is a tight coupling between the formation of potential solutions and their evaluation. This is the essence of abductive reasoning. This article presents an abductive framework for diagnostic problem solving. *Time-objects*, an association of a property and an existence, are used as the representation formalism and a number of primitive, general evaluation criteria into which time has been integrated are defined. Each criterion provides an intuitive yardstick for evaluating the space of potential solutions. The criteria can be combined as appropriate for particular applications to define plausible and best explanations.

The central principle is that when time is diagnostically significant, it should be modeled explicitly to enable a more accurate formulation and evaluation of diagnostic solutions. The integration of time and primitive evaluation criteria is illustrated through the Skeletal Dysplasias Diagnostician (SDD) system, a diagnostic expert system for a real-life medical domain. SDD's notions of plausible and best explanation are reviewed so as to show the difficulties in formalizing such notions. Although we illustrate our work by medical problems, it has been motivated by consideration of problems in a number of other domains (fermentation monitoring, air and ground traffic control, power distribution) and is intended to be of wide applicability.

Key words: diagnostic problem solving, temporal abductive diagnosis, diagnostic solution, time-object, evaluation criteria.

1. INTRODUCTION

1.1. Significance of Time

Time is often of great significance in diagnostic problem solving. We illustrate this using examples from the medical domain of skeletal dysplasias and malformation syndromes, developmental disorders that affect the skeletal system to varying degrees. A simplified description of the skeletal dysplasia spondyloepiphyseal dysplasia congenita (SEDC) reads as follows:

> SEDC *presents from birth* and may be lethal. It *persists throughout life*. Symptoms can include: short stature, owing to short limbs, *from birth*; mild platyspondyly *from birth*; absent ossification of knee epiphyses *at birth*; bilateral severe coxa-vara *from birth, worsening with age*; scoliosis, *worsening with age*; wide tri-radiate cartilage *up to about the age of 11 years*; pear-shaped vertebral-bodies *under the age of 15 years*; variable-size vertebral-bodies *up to the age of 1 year*; and *retarded ossification* of the cervical spine, epiphyses, and pubic bones [italics added].

The italicized text refers to time, directly or indirectly. The references to time points are absolute and specified with respect to some origin that here is birth. Absolute durations are specified explicitly or implicitly; e.g., property "SEDC present" persists throughout a patient's life, however long that might be. Since SEDC can be lethal, this duration

Address correspondence to Elpida T. Keravnou, Department of Computer Science, University of Cyprus, P. O. Box 20537, CY-1678 Nicosia, Cyprus. e-mail: *elpida@ucy.ac.cy.*

could be zero (events birth and death coincide). The occurrences (and hence durations) of properties “wide tri-radiate cartilage” and “pear-shaped vertebral-bodies,” at the granularity of years, are approximated through the qualitative expressions “up to about the age of ... ” and “under the age of . . . ,” respectively. We refer to this characteristic as *absolute vagueness* (Keravnou 1995b, 1996c). Some manifestations express temporal abnormalities or temporal trends, as in *retarded* ossification and *worsening* scoliosis, respectively. Manifestations like these, in which time constitutes an integral aspect, are *time-objects*, associations between properties and existences, e.g., ⟨platyspondyly, from-birth⟩, where “platyspondyly” is the property and “from-birth” the existence.

As another example, consider a simplified description of the dysmorphic syndrome morquio:

> Morquio *presents from the age of 1 year* and *persists throughout life*. Symptoms can include: short trunk; sloping acetabulae; generalized platyspondyly *from the age of 1 year*; thoraco-lumbar kyphosis *from the age of 4 years*; *progressive* resorption of the femoral-capital epiphyses *from the age of 2 years onwards*; more specifically flatness of the femoral-capital epiphyses appears *at the age of 2 years and persists up to an age between 8 and 15 years, and from then onwards* the ossification of femoral-capital epiphyses is absent [italics added].

There is a temporal trend in this description also. Its abstract expression is “progressive resorption of the femoral-capital epiphyses” that starts at the age of 2 years and terminates with death. At a finer level of description, the trend is divided into two phases, one of flatness and one of absence of ossification. The exact meeting point between the two phases (or change point from flatness to absence) is uncertain and may be at any time between the ages of 8 and 15 years.

The preceding descriptions of SEDC and morquio give the overall models for these disorders. Such models need to be temporally adapted to the case under consideration. For example, morquio presents a different picture for a 1-year-old, a 3-year-old, and a 17-year-old.

The diagnostic task for this domain (Keravnou et al. 1994) is to determine which skeletal dysplasia or malformation syndrome best accounts for the patient’s condition. Patient data are largely obtained from radiographs that give discrete snapshots of the development of the patient’s skeleton. For example, the following data are for a patient for whom the available radiographs were for the pelvis and the lateral spine at the ages of 2 and 7 years old and for the hands and the lateral skull at the age of 10 years:

> Carpal-bones small *at the age of 10 years*; femoral-capital-epiphyses abnormal *at the age of 2 years*; femoral-capital-epiphyses flat and irregular *at the age of 7 years*; vertebral-end-plates irregular *at the age of 7 years* [italics added].

The patient information is point-based in contrast to the medical knowledge, which is mainly interval-based. Patient information tends to be grossly temporally incomplete, so a competent, knowledge-based diagnostic system must be able to process the available data in an intelligent way. This usually requires an ability to derive abstractions from the given data that fill in the gaps and can be matched against the model of a disorder for a patient of that age. For example, a system should be able to conclude morquio as the explanation of the abnormal observations given earlier even though most do not appear as such in the model for morquio.

Abstractions for which time plays a central role are called *temporal abstractions* and are attracting considerable research interest as a fundamental intermediate reasoning process for the intelligent interpretation of temporal data in support of tasks such as diagnosis and monitoring (Haimowitz and Kohane 1996; Kahn et al. 1991; Keravnou 1997; Larizza et al. 1997; Lavrač et al. 1997; Miksch et al. 1996; Nejdl and Gamper 1994; Russ 1995; Shahar 1994; Shahar and Musen 1996; Shahar et al. 1992). Background domain knowledge (Keravnou et al. 1992) may be in the form of temporal abstractions; for example,

> The ossification of the cervical-spine *normally* begins at the eighth *week of gestation* and terminates by the 25th *year* [italics added].

This gives a high-level description of the normal ossification process, which spans a chain of temporal contexts (developmental periods); this becomes apparent once the process is decomposed into subprocesses at finer levels of description (see Section 3.2). Knowledge of normality serves several purposes in a diagnostic context. First, such knowledge can be used for establishing whether some observation describes an abnormal situation and therefore warrants an explanation. The ability to make such discriminations is a prime requirement of a competent system. For example, the earliest age at which the primary centers of ossification of the first cervical vertebra are expected to appear is 12 months and the latest 15 months. The nonappearance of these centers for a child of 10 months is not abnormal. Second, knowledge of normality can be used to further qualify (abstract) observations of abnormality. For example, the nonappearance of the centers at the age of 2 years would be abnormal, more specifically a delay in the ossification of the vertebra. Similarly, knowledge of normality can be used to refine expectations of disorder hypotheses, such as the expectation of retardation in the ossification of the cervical spine for SEDC, to potential observations of abnormality.[1]

1.2. Article Overview

We present an abductive framework for diagnostic problem solving. *Time-objects* are used as the representation formalism, and a number of primitive, general, evaluation criteria into which *time* has been integrated are defined.

To be considered, a potential solution to a problem must provide a *plausible* explanation. To be selected, a solution must be the "best" of the plausible solutions. Therefore, the questions that should concern developers of abductive diagnostic systems are "What is the meaning of explanation plausibility?", "How are plausible solutions formed?", and "How is the best solution selected?" These questions are the essence of abductive reasoning. In real-life diagnostic problems, the answers are not at all obvious. Explanation plausibility is defined through necessary evaluation criteria (or constraints). A potential solution that does not satisfy these minimum requirements, or *hard evaluation constraints*, is not plausible and should be rejected. Other *soft* evaluation constraints, specified over and above the minimum constraints, function to rank plausible solutions from different perspectives. When no one plausible solution fares best under the different soft evaluation perspectives, the problem is to determine the overall best plausible solution, which may include "no diagnosis."

The contributions of this article are twofold: (1) the integration of time within an abductive diagnostic framework and (2) the definition of a number of primitive, general evaluation criteria for temporal diagnostic solutions. The proposed primitive criteria, each of which represents an intuitive yardstick for evaluating the space of potential solutions, can be combined in a multitude of ways for obtaining appropriate definitions for particular applications, and examples of this are given. However, the article cannot propose any general definitions for plausible and best explanations because these will be domain/application-specific, if not expert-specific.

[1] Retardation is, of course, an expectation that cannot be observed, only inferred from an observation made at a particular time.

In many real-life domains, as in the example domain overviewed earlier, time is of major significance and should form an explicit and integral aspect of the knowledge and reasoning of diagnostic problem solvers. It is hard to imagine a diagnostic system with only an implicit notion of time performing well for the example domain. Although much research work is reported in abductive diagnostic reasoning (e.g. Bylander et al. 1991; Console and Torasso 1990; Peng and Reggia 1990; Pople 1973; Poole 1989a, 1990; etc.), relatively little is reported in *temporal*-abductive diagnosis (Console et al. 1992; Console and Torasso 1991a; Friedrich and Lackinger 1991; Keravnou and Washbrook 1990; Long 1996). We single out the work by Console and Torasso (1991a) because of its domain-independent exposition and compare and contrast it against our proposal at many points in the ensuing discussion. A significant difference between the two approaches is that while Console and Torasso's work represents time explicitly (in the form of temporal constraints on causal relations), it does not form an integral aspect of the domain entities. In our approach, this integration is achieved by modeling the relevant concepts (failures, faults, therapeutic actions, as well as normality) as *time-objects*. A time-object (Keravnou 1998) is a dynamic entity comprising a property and an existence. Time-objects enable a uniform and natural amalgamation of temporal knowledge with other essential types of knowledge such as structural and causal knowledge. This article elaborates the modeling of diagnostic concepts as collections of time-objects to achieve the key issue of the integration of time.

The formation of potential diagnostic solutions is discussed by overviewing context-free and context-sensitive mechanisms (through primary and secondary triggers, respectively). However, a major focus of this article is the evaluation of such solutions. In agreement with other researchers (e.g., Thagard 1992; Peng and Reggia 1990; etc.), we advocate the need for the formation and evaluation of potential solutions to be tightly coupled processes. Solutions are formed incrementally, and at all times the partial solutions must satisfy the minimum criteria that define explanation plausibility. Those solutions which fare better from some perspective(s) are further investigated for extension or refinement. We further advocate that primitive evaluation criteria should be represented separately and explicitly to permit their flexible and transparent combined use at different stages of the overall reasoning.

1.3. Evaluation Criteria

Just as time has been largely ignored in diagnostic problem solving, evaluation criteria have not been given the attention they deserve. Here we comment on some other approaches to the incorporation of evaluation criteria, irrespective of whether they explicitly address temporal aspects or not.

Thagard (1992) is quite emphatic about the need for a tight coupling between the formation and evaluation of hypotheses in computational, abductive systems. More specifically, he says that there are three possible models: (1) the two processes are completely independent, and hypotheses are formed in a random fashion, a nonviable option under limited resources; (2) the processes are weakly related, and only hypotheses that explain at least something are formed; or (3) they are strongly related, and only hypotheses that constitute likely possibilities are formed.

We refer to observations that necessarily need to be explained as *hard* findings (see Section 4). These relate directly to the hard evaluation constraints. If the formation of hypotheses is guided by the hard evaluation constraints, there is necessarily a strong relation between the formation and evaluation of hypotheses, since at any stage only plausible hypotheses are retained. Thagard also points out the inability (of some AI systems) to recognize those observations in need of explanation; this is a limitation because not every observation demands explanation (observations that denote surprising, unusual, or significant events should be singled out). We agree with Thagard that "further research is needed to identify how evaluation constraints can be used more effectively to help limit the range of hypotheses that can be generated in order to lead to ones more likely to be accepted" (Thagard 1992, p. 193). It is therefore important to fully appreciate the nature of evaluation constraints and explanation plausibility before specifying the formation process.

Thagard (1978, 1991b) suggests consilience, simplicity, and analogy as general criteria for measuring the quality of explanatory hypotheses, putting the emphasis on pragmatic notions. *Consilience* is concerned not only with how much a hypothesis explains but also the variety of things it explains. Variety of types of observations is common in realistic domains, e.g., in SDD, patient data can be clinical, biochemical, or histologic as well as radiologic. A hypothesis is *dynamically* consilient if it becomes more credible over time, a notion very pertinent to the example domain. *Simplicity* is concerned with the number of supporting assumptions, the well-known Occam's razor ["What can be done with fewer assumptions is done in vain with more" (Poole 1989a, quoting P. Edwards)], and *analogy* advocates the reusability of successful explanation models in analogous situations. The preceding general notions have been incorporated in a theory of explanatory coherence (Thagard 1991a).

In early diagnostic systems, much attention was paid to the evaluation aspect in an attempt to model the corresponding heuristics of expert diagnosticians (e.g., Miller et al. 1982; Patil 1981; Pauker et al. 1976; Pople 1977, 1982; etc.). The limitation of these approaches was the merging and embedding of the different criteria in so-called scoring functions, hiding the intuitive meaning of each criterion and preventing their flexible and transparent combined use in different contexts. On the positive side, though, such scoring functions were actively used throughout the reasoning process, thus directly influencing the formation of potential solutions. More recently, the trend in abductive diagnosis has been to explore how much can be achieved with somewhat restrictive and thus nonpragmatic criteria (Thagard 1991c). In such approaches, explanation plausibility is nothing less than complete accounting (coverage) of all observations of abnormality irrespective of their relative importance, say, for therapy. In real life it is rarely the case that this evaluation constraint will be satisfied; more usually, a hypothesis explains some observations but fails to explain others and may even be in conflict with them. Two celebrated theories of abductive diagnosis, namely, Peng and Reggia's parsimonious covering theory (Peng and Reggia 1990) and Poole's logic-based theory (Poole 1988, 1989a, 1990, 1994; Poole et al. 1987), are based on this restricted notion of explanation plausibility. In these theories, neither of which addresses time, there is indeed a strong coupling between the formation and evaluation of hypotheses, since only hypotheses that satisfy the preceding hard evaluation constraint are formed. The principle used to select the best explanation from the plausible ones is that of simplicity. More specifically, Peng and Reggia suggest parsimonious criteria based on relevancy (every disorder hypothesis included in an explanation is causally related to some observation of abnormality), irredundancy (none of the proper subsets of an explanation is itself an explanation), or minimality (prefer the explanation with the minimum cardinality). Poole suggests criteria based on minimality (prefer the explanation that makes the fewest, in terms of set inclusion, assumptions), least presumption (prefer the explanation that makes the fewest, in terms of what can be implied, assumptions), or minimal abnormality (prefer the explanation that makes the fewest failure assumptions or makes the same abnormality assumptions but fewer normality assumptions). Van Harmelen and ten Teije (1994), who also define explanation plausibility as full coverage of observations, propose the use of domain knowledge in selecting the best explanation on the ground that minimality can still yield a number of best explanations. As already mentioned, we only propose a number of primitive evaluation criteria. It is possible to identify such criteria at a general level. However, it is not possible to give general definitions for *plausible* and *best explanation* that would apply to different, real-life diagnostic domains. For illustration, we discuss how SDD uses some of the proposed primitive criteria in its notions of plausible and best explanation.

In summary, we argue that time, for those diagnostic domains where it is of significance, should be represented explicitly in an integrated way. We adopt the notion of the time-object as the central representation primitive for achieving the proper integration of time. The modeling of time enables a more accurate formation of potential solutions; e.g., the presence of an abnormality may not be diagnostically significant as such, but its specific pattern of appearance is. It also enables a more accurate evaluation of the considered solutions; e.g., the expected picture of a disorder/failure is different depending on the stage of its evolution. Furthermore, we stress the significance of evaluation criteria in abductive diagnostic reasoning, irrespective of the significance of time, since the overall aim is to derive a solution that is the best explanation of the observations. We advocate that primitive evaluation criteria should be represented separately and explicitly to allow their transparent combined use in different reasoning contexts. Such flexibility in the representation of primitive evaluation criteria enables the formulation of different notions of explanation plausibility and best explanation.

The rest of this article is organized as follows: Section 2 gives a global view of diagnostic reasoning from the temporal abductive perspective. Section 3 overviews the adopted temporal ontology, presents models for failures, faults, normality, and therapeutic actions in terms of time-objects; and outlines mechanisms for the context-free (via primary triggers) and context-sensitive (via secondary triggers) formation of potential solutions. Section 4 presents the proposed primitive evaluation criteria for temporal solutions, Section 5 illustrates the application of some of these criteria in the SDD system, and finally, Section 6 concludes the discussion.

2. TEMPORAL ABDUCTIVE DIAGNOSTIC REASONING

In this section we present a high-level view of temporal abductive diagnosis.

2.1. Application Context, Diagnostic Theory, and Case Histories

A diagnostic problem solver has a theory, its knowledge per se, that is bounded by an application context. The theory is applied to the history of an actual case (human or other) when solving diagnostic problems for that case.

The *application context (AC)* delineates the extent of the problem solver's "expertise" or competence and thus the scope of its diagnostic theory. The role of the application context is in recognizing whether a particular problem is within, at the periphery, or outside the problem solver's expertise prior to attempting to solve the problem. It therefore reflects what the problem solver is supposed to know and be able to do. The specification of an application context may appear superfluous at the outset on the ground that a system will never be consulted for a problem outside its scope. However, in real-life situations, this is possible, especially for overlapping domains. For example, it is not uncommon for a patient to be referred on by one specialist to another, either because he or she detects that the patient's problem is peripheral or outside his or her expertise or because he or she feels that a more specialized opinion is required. In multi-problem-solver frameworks where problem solvers operate in a collaborative rather than stand-alone basis, it is necessary to specify application contexts for the individual problem solvers as the means for recognizing the relevance of problems to solvers. For example, the early diagnostic system MDX (Chandrasekaran and Mittal 1983) exhibited a rudimentary collaborative architecture through a community of hierarchically organized specialists. Each specialist had knowledge of what it knew to enable it to decide whether a problem referred to it really belonged to its scope. In a sophisticated form, an application context is a kind of metadiagnostic theory. However, an application context can be expressed in a simple procedural way through a set of questions that are raised at the beginning of a session with the system for establishing the relevance of the particular problem. This is the approach taken by the SDD system, whose scope is singly occurring skeletal dysplasias and malformation syndromes. A central question for establishing whether a problem belongs to the system's scope is whether the patient exhibits a generalized skeletal problem. Generally speaking, the application context specifies the domain(s) of application (types of cases, e.g., human) and the types of failure addressed (e.g., single or multiple failures and of what sort, such as single skeletal dysplasias).

The *diagnostic theory* (*DT*) constitutes the knowledge of the diagnostic system. In this article we are interested in temporal-abductive diagnostic theories, i.e., theories with explicit notions of time whose purpose is to best explain (account for) abnormal situations.

A central component of a theory is the set of *temporal models for the distinct failures* covered by the system. The theory is complete if it includes a model for every known failure covered by the application context. Depending on the breadth and rate of growth of the application domain, it may be difficult to have a complete diagnostic theory. For example, the domain of skeletal dysplasias and malformation syndromes includes 2000+ such disorders, and this number grows due to the continual discovery of new syndromes. The diagnostic theory of the SDD system currently includes models for just 200 skeletal dysplasias.

In addition, a diagnostic theory includes *background knowledge*. For the SDD system, this is knowledge of the normal evolution of ossification processes as well as anatomic and other fundamental medical knowledge. To draw a comparison with the Theorist framework (Poole et al. 1987; Poole 1994), the failure models correspond to conjectures (abnormality assumptions that are only considered if there is evidence suggesting them), whereas background knowledge comprises both defaults, normality assumptions that are assumed to hold unless there is evidence to the contrary (e.g., normal evolution of ossification processes), and facts (e.g., anatomic knowledge). Finally, the background part of a diagnostic theory includes models of therapeutic (or other) actions of relevance to the covered failures from a diagnostic perspective. We consider the integration of therapeutic actions in a diagnostic setting another contribution of our approach. The different parts of a diagnostic theory are explained in detail and illustrated in Section 3.

A *case history* (*CH*) gives factual information on an *actual case*. It is a temporal (historical) database of the case. For medical diagnostic theories, case histories are patient records. Some of the information recorded is time-invariant (e.g., the sex and race of a patient or the decomposition of a physical device into components and subcomponents), but most of the information is temporal (e.g., dates and results of medical examinations, past failures, therapeutic actions, etc.). A case history is continuously updated with new information. The derivation of trends or periodic occurrences is not possible without a temporal account of the history of a case.

A diagnostic problem is triggered when observations suggest a malfunction. During a particular diagnostic activity, the application context and diagnostic theory remain static, but the relevant case history is dynamically updated through the incremental acquisition of more observations, the derivation of (intermediary) conclusions on the status of the patient, and the application of therapeutic actions. A competent diagnostic system should be able to evolve on the basis of its own experience, thus gradually refining and extending its diagnostic theory (i.e., knowledge) and even application context. This topic, however, is outside the scope of this article.

2.2. Temporal-Abductive Diagnosis

Diagnosis is a stepping stone to treatment, and the two processes may be integrated/interleaved in some higher-level process whose aim is to improve the status of a patient in a timely fashion (Friedrich 1993; Sadegh-Zadeh 1994). Time constitutes an integral aspect of both diagnosis and treatment. For many diagnostic problems, it is the pattern of changes that is significant in solving the problem rather than a snapshot at a particular time. A case history is stored in a temporal database that records the patient's past, current status, and possibly predicted future. The database consists of a collection of temporal assertions, associations between some property, and a span of *valid time*.[2] Such temporal assertions are *concrete time-objects* (see Section 3.1). Time-invariant properties for the case (e.g., properties describing a physical device's structural composition) are believed, at all times, to hold throughout the lifetime of the case; hence everything can be treated as a temporal assertion. A temporal assertion that is no longer believed simply can be deleted from the database.

Let CH_t be a case history at time t (recall that observations of misbehavior become part of the case history), and let S_t be a potentially abducible diagnostic solution for the particular case at time t, i.e.,

$$DT \cup CH_t \rightarrow S_t$$

where → stands for "suggests," not logical implication, because the inference is abductive and thus plausible in nature. If $S_t = \{\langle \neg f, t\rangle | f \in \textit{failures under AC}\}$, i.e., none of the covered failures is believed to hold at time t, the case is assumed to be functioning normally.[3] Otherwise, the case is malfunctioning, and S_t is an explanation. If no failure can be established to hold at time t, although there are observations of ongoing abnormality, it is possible that a transient failure has caused a persistent fault (Friedrich and Lackinger 1991). The adopted temporal ontology (see Section 3) enables the modeling of a transient or persistent failure, both of which can cause a transient or persistent fault. A competent diagnostic system ought to be able to explain an ongoing manifestation

[2] Strictly speaking, a *temporal* assertion is associated with both valid and *belief* time as used in Mylopoulos et al. (1990). In database research, valid time is also referred to as *historical time* and belief time as *transaction time* (Maiocchi and Pernici 1991). Hence an assertion that only has valid time is a historical assertion. However, we feel that the term *temporal assertion* is more representative because in the scope of this work it is the valid time that is of prime significance, while belief time is of secondary importance (once the integration of valid time is achieved, belief time is conceptually easy to incorporate). Thus, in common with other proposals on temporal diagnostic frameworks, we do not make use of belief time at this stage.

[3] A diagnostic system does not usually reason by sequentially attempting to refute each potential failure in turn. Through a triggering mechanism (see Section 3.3), likely failures are activated (i.e., diagnostic hypotheses are formed) that are subsequently explored for confirmation/refutation.

by attributing it to a transient (and hence past) failure if such is the case. Like the case history, a potential diagnostic solution S_t at time t consists of temporal assertions (i.e., time-objects), say, $\langle f, vt \rangle$. Thus S_t consists of assertions of past and/or ongoing failures that explain observations of abnormality included in the case history. Each potentially abducible diagnostic solution $S_{i,t}$ at time t, $i = 1, \ldots, n$ represents a hypothetical extension of the case history CH_t. The diagnostic problem solving terminates successfully at time $t = G$ iff at that time one of the competing diagnostic solutions is "confirmed" as the correct solution; i.e., it gets truly believed. Only a confirmed (believed) diagnostic solution can become part of the case history. Assertions about actual therapeutic actions performed on the case are entered directly into the case history.

Sadegh-Zadeh (1994) has proposed a framework for clinical reasoning called *differential indication*, where observation, diagnosis, and treatment are all seen as forms of actions, and therefore, the various reasoning processes can be integrated naturally—in Sadegh-Zadeh's argument they form a single integral process whose goal is to improve the patient situation. The following is an adaptation of part of Sadegh-Zadeh's proposal, for the purposes of this article: Let TA_t and OA_t, respectively, be therapeutic and observation actions suggested by the diagnostic-therapeutic theory DT and the case history CH_t at time t. The contents of TA_t and OA_t are temporal assertions of the same type as those included in CH_t and $S_{i,t}$. Such assertions denote the instigation of actions at some time after t. If a suggested action is actually instigated, it leaves OA_{t+}/TA_{t+} and enters (or its result, in the case of observation actions, enters) CH_{t+}, where $t+$ denotes a time after t. Similarly, a suggested action is deleted by default when its specified initiation point becomes a point in the past. Thus at time t we have

$$DT \cup CH_t \to TA_t$$

$$DT \cup CH_t \to OA_t$$

$$DT \cup CH_t \to S_{1,t} \qquad DT \cup CH_t \to S_{2,t} \qquad \cdots \qquad DT \cup CH_t \to S_{n,t}$$

If there is just one potential diagnostic solution, which says that the case is functioning properly, $TA_t = \{\}$ and presumably $OA_t = \{\}$. Otherwise, at least one therapeutic or observation action is proposed that may well be "wait for such a period of time to see how things develop." Thus the preceding inferencing is repeated at subsequent points in time, using the updated case history, until the diagnostic conclusion is reached that the case is functioning normally or that no further improvement is possible. It is still possible that some observation and therapeutic actions are suggested.

According to the preceding sequence of inferences, the derivation of actions is done independently of the derivation of potential diagnoses—"parallel" inferencing. Alternatively, the following "sequential" inferencing (which takes place in the same quantum of real time denoted by t) enables decisions on potential actions to be influenced by beliefs on potential diagnoses:

$$DT \cup CH_t \to S_{1,t} \quad DT \cup CH_t \to S_{2,t} \qquad \cdots \qquad DT \cup CH_t \to S_{n,t}$$

$$DT \cup CH_t \cup \{S_{1,t}, S_{2,t}, \ldots, S_{n,t}\} \to TA_t$$

$$DT \cup CH_t \cup \{S_{1,t}, S_{2,t}, \ldots, S_{n,t}\} \to OA_t$$

The set of all hypothesized diagnostic solutions $\{S_{1,t}, S_{2,t}, \ldots, S_{n,t}\}$ is a "conflicting" set. However, the actions are globally decided on the basis of all entertained diagnostic solutions; this mode of inferencing is especially appropriate for observation actions that aim to differentiate competing diagnostic solutions. The "parallel" inferencing is abductive in nature. The "sequential" inferencing also has a deductive aspect, the prediction of the expectations of potential solutions.

In summary, a diagnostic process involves (Figure 1) detection of a malfunction, formulation of a diagnostic problem, formation of potential diagnostic solutions, evaluation of potential diagnostic solutions, and conclusion of the "best" diagnostic solution. The detection of a malfunction is often done outside the diagnostic system. The formulation of the diagnostic problem entails a ranking of the observations of abnormality by importance, to define the misbehavior that needs to be explained. This requires event-driven reasoning, where suitable abstractions (intelligent interpretations) are drawn from the observations. Once a diagnostic problem is formulated, potential solutions are formed and evaluated. During this process, new possible solutions may be introduced. Ideally, one of these solutions is eventually concluded, although often this is not possible (or even necessary if all potential solutions lead to the same therapy). Instead, the most likely solutions are presented. In addition, the system may present solutions that, though less likely, could have serious outcomes and should be considered when deciding on a therapeutic action.

Diagnostic reasoning does not proceed sequentially through the preceding steps. First, new information about the case can be obtained throughout the diagnostic process. Thus changes in the formulation of the problem can occur. Second, as already

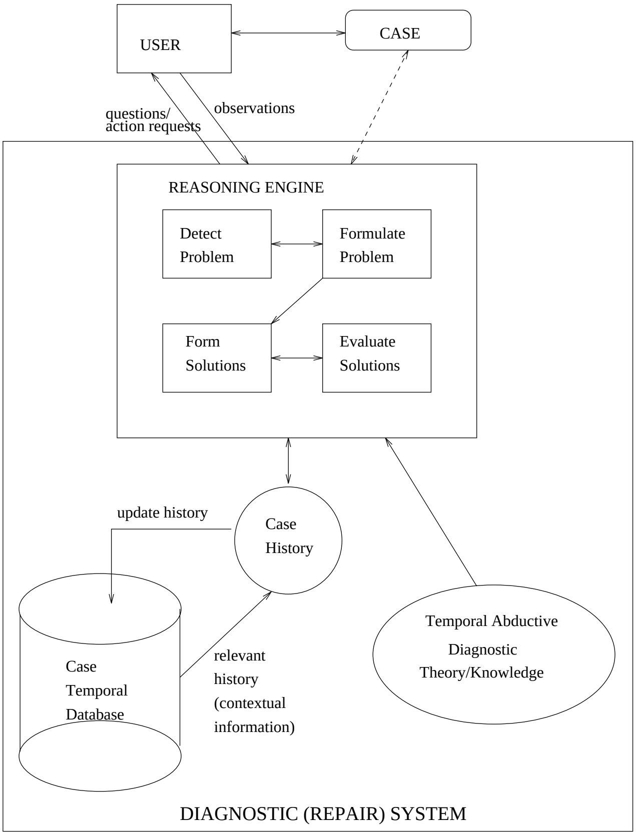


FIGURE 1. High-level view of a diagnostic system.

discussed, the formation and evaluation of potential solutions are strongly coupled processes—at any stage in the reasoning, only the most promising potential solutions, as decided by the evaluation criteria used in the particular reasoning context, are actively pursued.

3. DIAGNOSTIC CONCEPTS AS TIME-OBJECTS

3.1. Time Ontology

The principal primitives of the adopted time ontology are the *time-axis* and the *time-object* that, respectively, provide a model of time (Keravnou 1999) and a model of occurrences (Keravnou 1996a, 1996b, 1998). A time-axis α represents a period of valid time from a given conceptual perspective. It is expressed discretely as a sequence of time-values, $\text{Times}(\alpha) = \{t_1, t_2, \ldots, t_n\}$, relative to some origin. Time-axes are of two types, atomic axes and spanning axes. An *atomic axis* has a single granularity (time-unit) that defines the distance between successive pairs of time-values of the axis. Its time-values are expressed as integers. A *spanning axis* spans a chain of other time-axes. It has a hybrid granularity formed from the granularities of its components, and its time-values, also inherited from its components, are tuples (Keravnou 1999). An application can involve a single atomic axis and a single granularity or a collection of time-axes and multiple granularities where the same period of time can be modeled from different conceptual perspectives. In the examples in Section 1, relevant time-axes could be fetal-period, infancy, childhood, puberty, and maturity, the latter four collectively forming a spanning axis of lifetime. Similarly, childhood could be a spanning axis, decomposed into early, middle, and late childhood. The others could be atomic axes, where the granularity for fetal-period and infancy could be months, whereas for maturity, years, etc. If the origin for all these axes is birth, the time-values for fetal-period would be $\{-10, -9, \ldots, 0\}$, a negative value denoting a time before the origin, which is denoted by 0. These are general, or abstract, time-axes whose origin is a generic time-point. Such time-axes can be instantiated for specific cases by binding their abstract origin to an actual time point, thus obtaining concrete time-axes. The distinction between abstract and concrete, which applies to time-objects as well, is important; a diagnostic theory is expressed at an abstract level; a case history, at a concrete level.

The notion of a time-axis is an abstraction mechanism for a more efficient and conceptual organization of time-objects. However, the ultimate objective is to be able to express different types of occurrences, and hence the central notion is the time-object. For brevity, in the following discussion we assume that there is a single abstract atomic axis used in the definition of the diagnostic theory. This is instantiated to provide the concrete time-axis for the definition of the particular case history.

A time-object is a dynamic entity that has time as an integral aspect. It is an association between a *property* and an *existence*. The manifestations of SEDC and morquio given in Section 1.1 are examples of time-objects, e.g., ⟨pear-shaped vertebral-bodies, under the age of 15 years⟩, ⟨coxa-vara, from birth⟩, etc. The notion of a time-object enables the definition of different types of occurrences, such as simple (atomic) occurrences or compound occurrences (such as trend occurrences, periodic occurrences, or any other pattern of simpler occurrences). To be able to express such occurrences, the ontology of time-objects includes three types of relations between time-objects: *temporal relations* that are adapted and extended from Allen's (1984) set, *structural relations*, that enable the composition and decomposition of time-objects, and *causal relations*.

Time-objects, like time-axes, are either abstract or concrete. Disorder models consist of abstract time-objects, whereas case histories consist of concrete time-objects. The existence of abstract/concrete time-objects is given with respect to abstract/concrete time-axes. Given the multiplicity of time-axes, formally, a time-object τ is defined as a pair $\langle \pi_\tau, \varepsilon_\tau \rangle$ where π_τ is the property of τ and ε_τ is its existence function. The time-axis that provides the most appropriate conceptual context for expressing the existence of τ is referred to as the *main time-axis* for τ, and the expression of τ's existence with respect to its main time-axis is referred to as its *base existence*. The existence function ε_τ maps the base existence of τ to other conceptual contexts (time-axes). A time-object has a *valid* existence on some time-axis iff the granularity of the time-axis is meaningful to the property of the time-object (see below) and the span of time modeled by the time-axis covers (possibly partially) the base existence of the time-object. If time-object τ does not have a valid existence in the context of time-axis α, $\varepsilon_\tau(\alpha) = \perp$ (the time-object is undefined with respect to the particular temporal context). If time-object τ has a valid existence on some time-axis α, its existence on α, $\varepsilon_\tau(\alpha)$,[4] is given as

$$\varepsilon_\tau(\alpha) = \langle t_s, t_f, \zeta \rangle$$

where $t_s, t_f \in$ Times(α), $t_s \leq t_f$, and

$$\zeta \in \{\texttt{closed, open, open-from-left, open-from-right, moving}\}$$

Time-values t_s and t_f, respectively, give the (earliest) *start* and (latest) *finish* of the time-object on α. The third element of the existence expression, ζ, gives the *status* of τ on α. If the status is `closed`, the existence of the time-object, and hence its duration, is fixed. Otherwise, the status denotes openness (i.e., vagueness) on the one or both ends of the existence. In the case of openness at the start, t_s gives the earliest possible start, whereas function *left-freedom*$_\tau(\alpha)$ gives the latest possible start. Similarly, in the case of openness at the finish, t_f gives the latest possible finish, whereas function *right-freedom*$_\tau(\alpha)$ gives the earliest possible finish. The existence of an `open` time-object, on a given time-axis, is therefore defined through an initial period of uncertainty, an in-between period of certainty, and a final period of uncertainty. If the earliest finish of an `open` time-object, with respect to some time-axis, precedes or coincides with its latest start, there is no period of certainty. Thus the duration of a nonclosed existence of a time-object can only be shortened.

Hence a time-object can exist as a *point-object* on some time-axis but as an *interval-object* on another time-axis. In the former case, the temporal extent of the time-object is less than the time-unit of the particular time-axis. If a time-object is a point-object under some time-axis, it is treated as an indivisible (nondecomposable) entity under that time-axis. A special `moving` time-object is *now*, which exists as a point-object on any relevant concrete time-axis and functions to partition (concrete) time-objects into past, future, or ongoing.

[4] The existence function ε is in fact a two-parameter function, $\varepsilon(\tau, \alpha)$; ε_τ simply denotes the partial parameterization of the function with respect to the argument time-object τ which gives a single parameter function. Similarly, π_τ denotes the (full) parameterization, with respect to the argument time-object τ, of the single parameter function $\pi(\tau)$.

The structural relations between time-objects are `isa-component-of` and its inverse `contains` and `variant-component` and its inverse `variant-contains`; the latter two express conditional containment of optional components:

Axiom 1. $\texttt{contains}(\tau_i, \tau_j) \Leftarrow \texttt{variant-contains}(\tau_i, \tau_j, c) \wedge \textit{conds-hold}(c)$

A variant component can only be assumed in some case if the specified conditions are satisfied. This is expressed in predicate *conds-hold*. For example, aspects of the ossification processes for carpals and radial and tarsal epiphyses differ between boys and girls. These distinctions can be conveniently modeled through variant components of the particular ossification processes. A compound time-object has a valid existence under any time-axis in which at least one of its components has a valid existence, and a component time-object exists within the one that contains it. Temporal views of a compound time-object, from the perspective of specific temporal contexts, thus can be defined. Trends and periodic occurrences are modeled as compound time-objects (Keravnou 1997).

Causality is a central relationship in diagnostic problem solving. The ontology of time-objects includes relations `causes`, `causality-link` and `cause-spec` which are defined at the level of abstract time-objects, concrete time-objects, and abstract properties, respectively (Keravnou 1998). Relation $\texttt{causes}(\tau_i, \tau_j, cs, cf)$, where τ_i and τ_j are abstract time-objects, *cs* is a set of temporal and other constraints, and *cf* is a certainty factor, is used in the following Axiom for deriving a `causality-link` between a pair of concrete instances of τ_i and τ_j. A general constraint that always needs to be satisfied is that a potential effect cannot precede its potential cause:

Axiom 2. $\texttt{causality-link}(\tau_i, \tau_j, cf) \Leftarrow \texttt{causes}(\tau_i, \tau_j, cs, cf) \wedge \textit{conds-hold(cs)} \wedge$

$\neg\texttt{starts-before}(\tau_j, \tau_i)$

Predicate $\texttt{starts-before}(\tau_j, \tau_i)$ expresses that τ_j starts before τ_i. Even if all the specified conditions are satisfied, by some case, still it may not be definite that the `causality-link` actually exists owing to knowledge incompleteness. This is modeled by the certainty factor.

Properties, which constitute the other half of time-objects, are atomic or compound (negations, disjunctions, or conjunctions), passive or active, and some are time-invariant. A typical format for atomic properties is (⟨subject⟩ [, ⟨attribute 1⟩, ⟨value 1⟩, . . . , ⟨attribute *n*⟩, ⟨value *n*⟩]). Examples of properties are "sex male," "sore throat," "severe coughing," "removal of tonsils," etc. Properties have explicit temporal attributes. A property is associated with relevant granularities; e.g., "headache present" is associated with hours and days but probably not months or years. A property either has an infinite or a finite persistence. In the latter case, the following are additionally specified: whether the property can *recur* (multiple instantiations of the given property in the same case are possible) and maximum and minimum durations under any of the relevant granularities, independently of any context in which they may be instantiated, where the default is to persist indefinitely. A default margin for the initiation of some instantiation of the property, under a relevant time-axis (*earliest-init, latest-init*), is also included in the temporal attributes of properties. If not specified, this is assumed to be the entire extent of the particular time-axis. For example, "SEDC present" is an infinitely persistent property whose *earliest-init* is birth. On the other hand, "flu present" is a finitely persistent, recurring property, and "chicken pox present" is a finitely persistent but normally not a recurring property. In addition, the ontology adopts the semantic attributes of properties specified by Shoham (1987), e.g., downward hereditary, upward hereditary, solid, gestalt, etc.

Relation cause-spec between properties has six arguments, where the first two are properties, the third a granularity, the fourth and fifth sets of relative (*TRel*) and absolute (*TAbs*) temporal constraints, respectively, and the last one a certainty factor. This relation also enables the derivation of a causality-link between a pair of time-objects:

Axiom 3. $\texttt{causality-link}(\tau_i, \tau_j, cf) \Leftarrow \texttt{cause-spec}(\rho_i, \rho_j, \mu, TRel, TAbs, cf)$

$$\wedge \pi(\tau_i) = \rho_i \wedge \pi(\tau_j) = \rho_j \wedge \textit{r-satisfied}(\tau_i, \tau_j, \mu, TRel)$$

$$\wedge \textit{a-satisfied}(\tau_i, \tau_j, \mu, TAbs) \wedge \neg \texttt{starts-before}(\tau_j, \tau_i)$$

Predicates *r-satisfied* and *a-satisfied* express the satisfiability of the relative and absolute temporal constraints, respectively. Other property relations include exclusion, necessitation, etc.

The implementation of the time ontology in terms of meta, abstract, and concrete layers is discussed in Keravnou (1999). This includes a declarative assertion language combining object-oriented, functional, and logical features for the expression of the various Axioms.

3.2. Modeling Failures, Faults, Normality, and Actions as Time-Objects

In this section we explain how abductive diagnostic theories (failure models and background knowledge of normality) and case histories can be uniformly represented in terms of time-objects (Keravnou 1995a, 1996d). A specific diagnostic activity operates within a (possibly moving) window of real time that at any instant of time gives the past and future period of interest. This time window forms the concrete time-axis. The relevant history of a case is that which is covered by the concrete time-axis.

Failure Model. An abductive diagnostic theory primarily contains *failure models* for all the known failures covered by the application context. A necessary condition for the instantiation of a failure model is that the model's abstract time-axis can be mapped onto the concrete time-axis for the case.

Typically, a failure is a nonobservable malfunction whose presence in some situation is detected through its observable manifestations, its associated faults. We classify failures and faults as follows from the temporal perspective: (1) infinitely persistent, either with a fixed or a variable initiation margin (e.g., SEDC or morquio), (2) finitely persistent but not recurring, again either with a fixed or a variable initiation margin (e.g., chicken pox), and (3) finitely persistent that can recur (here the initiation margin is variable), e.g., flu. The temporal extent of a finite persistence is either indefinite or bounded (through minimum and maximum durations). Transiency is associated with finite, and usually short, duration. Thus, at any point in real time, actual failures/faults are described as either persistent (ongoing) or transient (in the past). The classification of failures/faults just given is fully covered by the semantics of properties in the adopted time ontology (see above).

A typical model of some failure is an (acyclic) causal structure comprising a number of causal paths emanating from the node denoting the failure and terminating at nodes denoting (usually observable) faults. Intermediate nodes on such paths denote internal (usually unobservable) causal states. When no internal states are included, the causal structure is reduced to a simple associational structure between the failure and its faults. Figure 2 illustrates the representation of a failure model as a causal structure. Such a causal structure is naturally expressed as a collection of abstract time-objects—each node corresponds to a time-object and each arc to the relevant instance of relation causes. If the failure Φ has a fixed initiation margin, its existence is expressed relative to the origin of the time-axis; otherwise, the failure is assumed to occur at any time, and only a (default) margin for its duration, if any, is known. The existence of an internal causal state ($CS_1, CS_2, \ldots, CS_9$) or a fault ($O_1, O_2, \ldots, O_5$) is expressed relative to either the origin of the time-axis or the existence of the failure Φ. The temporal knowledge regarding the existences of the nodes in a failure model may well be incomplete, and hence, where appropriate, default durations are used; in the worst case, every node has an indefinite persistence. In addition, the existences of the various time-objects may be constrained in a relative way by pairwise temporal relations between them.

Compound causal antecedents (at different levels of abstraction) can be expressed simply in the time-object ontology. Causal state CS_8 is an example of such a compound causal antecedent consisting of causal states CS_6 and CS_7. The existence of O_5 depends on the existence of CS_8, and hence, by implication, it depends on the existences of both CS_6 and CS_7. No restrictions are imposed on the relative existences of CS_6 and CS_7. O_5 materializes in some situation if its compound antecedent CS_8 materializes (and all the specified conditions, if any, are satisfied); CS_8's materialization depends on the materialization of its components.

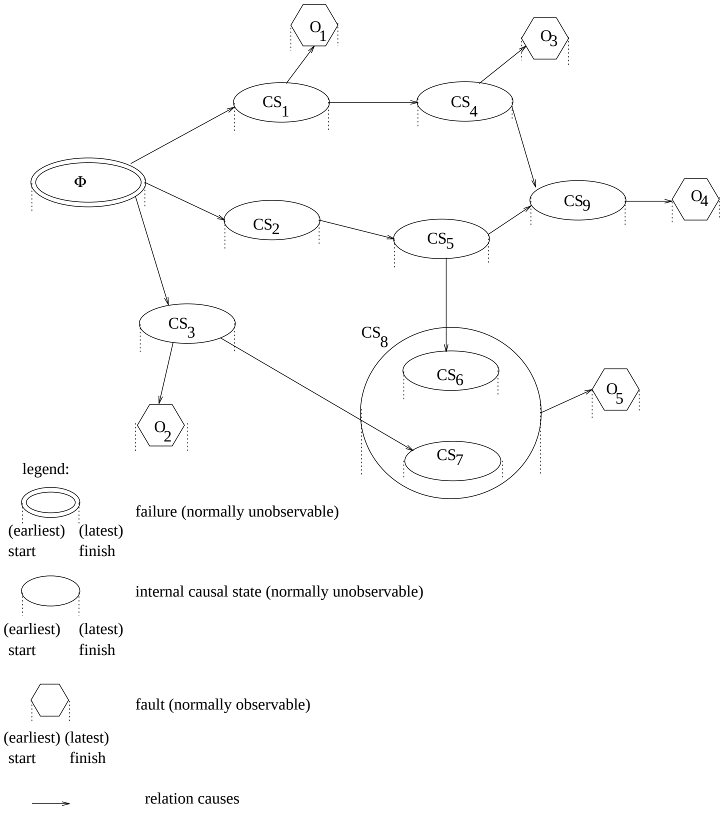


FIGURE 2. Failure model as a causal structure of time-objects.

Similarly, a node can have alternative, independent, causal antecedents, e.g., causal state CS_9 that has causal antecedents CS_4 and CS_5. The interpretation is that the temporal extent of CS_9 is the same under any of the following situations: only CS_4 materializes, only CS_5 materializes, and both CS_4 and CS_5 materialize. If this is not so, the depicted part of the model is replaced with three paths whose respective antecedents correspond to the preceeding three situations, and their consequents are three distinct time-objects sharing the same property but having different extents.

For illustration purposes, the descriptions of the skeletal dysplasia SEDC and the dysmorphic syndrome morquio given in Section 1 are represented as causal structures of time-objects (Figures 3 and 4). Each node depicts a time-object; the text inside the node gives the property of the time-object (a discussion on the format and full semantics of properties for this domain are outside the scope of this article), and the text outside gives its (base) existence [(earliest) start and (latest) finish]. These existences are expressed with respect to the single abstract time-axis (not depicted in the figures), lifetime, say, whose granularity is years and origin birth. Some of these existences are uncertain. For example, the exact termination of properties "vertebral-bodies pear-shaped" and "tri-radiate-cartilage wide" in the SEDC model cannot be specified. Similarly, the exact meeting point of the two components of the trend time-object "femoral-capital-epiphyses progressive-resorption" in the morquio model is not known, but a margin for it can be specified (ages 8 to 15 years).

The majority of nodes in these models are considered as observable. This is so because the conditions described by their properties can be observed through radiographs. The two primitive trends in SEDC are represented as *metaqualifications*, or progression patterns, ("worsening") over observable properties. The trend per se is not

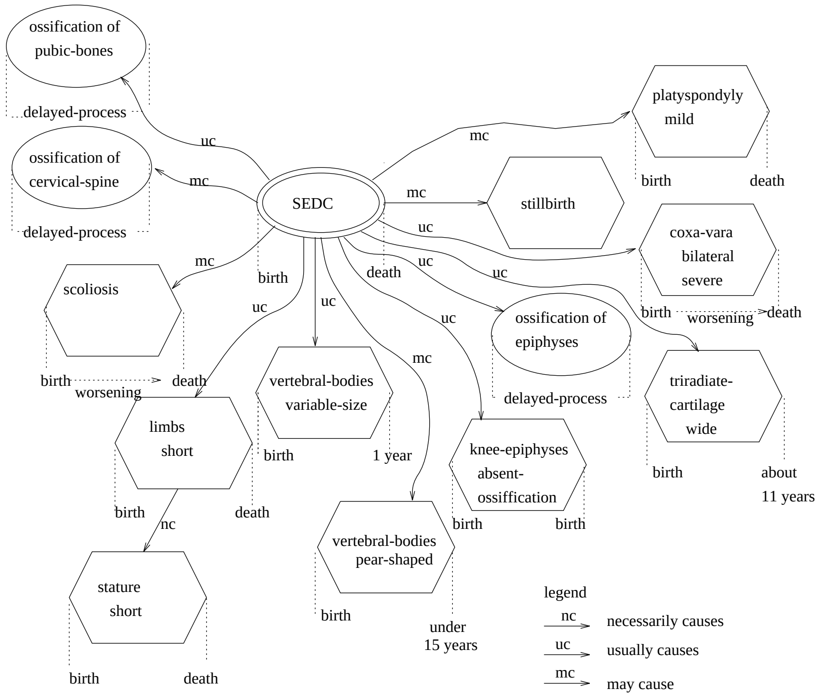


FIGURE 3. Modeling the skeletal dysplasia SEDC as a causal structure of time-objects.

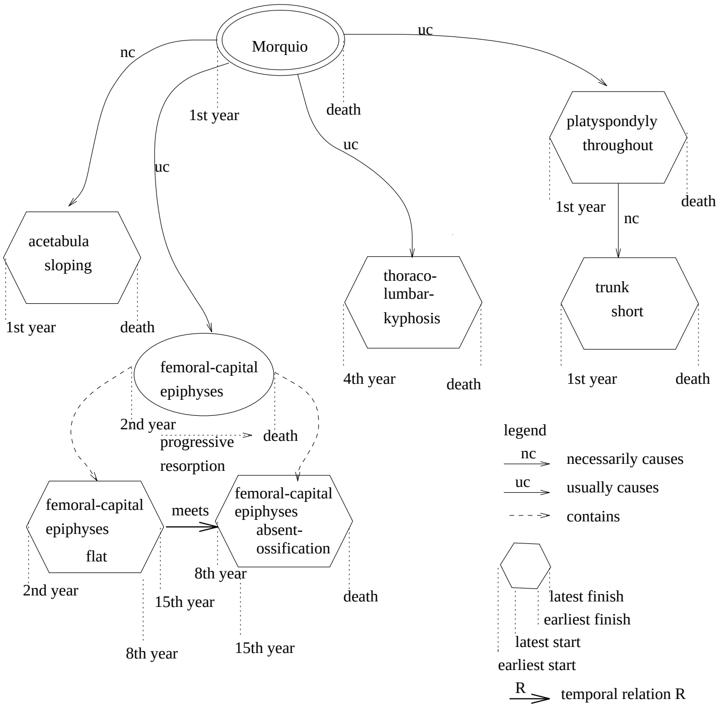


FIGURE 4. Modeling the dysmorphic syndrome morquio as a causal structure of time-objects.

observable and can only be inferred (or denied) through a temporal sequence of observations of the given property. Morquio's (compound) trend is also expressed through a metaqualification, "progressive-resorption," over property "femoral-capital-epiphyses." The given node is considered unobservable because its use is in the context of the given progression; femoral-capital-epiphyses are not observed for the sake of some snapshot recording but for the purpose of establishing the trend.

The solid arcs are instances of relation `causes`. In the depicted instances, there are no conditions, and the certainty factors are expressed in a qualitative fashion as *nc* (necessarily causes), *uc* (usually causes), and *mc* (may cause), given as labels on the arcs. Thus each arc can be expressed as `causes` $(\tau_i, \tau_j, \{\}, nc/uc/mc)$. Usually only a small proportion of the causal arcs define necessary causation; the majority of them define conditional, uncertain causation. Due to knowledge incompleteness, the set of conditions associated with a causal arc is often incomplete or unknown, and hence there is inherent uncertainty. Based on the degrees of uncertainty, the effects of some cause (and hence the faults of some failure) can be classified into necessary, common, occasional, etc. Such a classification can be used in the evaluation of potential diagnostic solutions (see Section 4).

Since most of the causal chains are of unit length, what we really have here are associational relations between the disorders (failures) and their manifestations (faults); no intermediate, internal causal states are depicted. However, in the SEDC model, some terminal nodes are in fact (unobservable) internal states. These represent delays in particular ossification processes. The existences of these time-objects are not given in absolute terms but implicitly through the metaqualification "delayed-process," which encompasses a multitude of actual possibilities. Since the significant information is the presence and not the specific form of delay, there is no need to be any more detailed.

As with a trend, a delay per se is not observable but inferable. On the basis of the background knowledge giving the normal behavior of ossification processes and the case observations at a lower level (such as information on the status of bone growth observed directly from radiographs), it can be inferred whether ossification is progressing normally or if (part of) it is exhibiting a delay or prematurity. In SDD this inferencing is sometimes done by the user of the system, who can then enter directly the abstract information that there is delay/prematurity in an ossification process.

Figure 5 illustrates a very small portion of the representation of the normal ossification process of the cervical spine. The process is represented as an unobservable compound time-object whose components are also unobservable compound time-objects. The components represent subprocesses that are gradually refined into observable events, such as the appearance of primary and secondary centers of ossification, which signify the start and/or completion of (sub)processes. Such events are point-objects at the relevant granularities, and although their occurrences cannot be specified exactly, they can be bounded. The description of this ossification process refers to three different granularities, weeks, months, and years, that implicitly refer to the conceptual temporal contexts (time-axes), fetal-period, infancy, childhood, puberty, and maturity. Informally, the following Axioms can operate on such representations of normal processes for deducing delays: If an event marking the start of a process has happened after the expected latest start, there is a delay; if an event marking the completion of a process has happened after the expected latest finish, there is a delay; and (for compound processes) if a component of the process is delayed, (part of) the process is delayed. Similar Axioms can be used for deducing prematurity. It is possible that an actual ossification process exhibits both a delay and a prematurity.

Figure 5 also illustrates a generic causal relation that captures the default persistence, i.e., the normal expectation, regarding the ossification status of any skeletal part: Once a part of the skeleton is ossified, it is expected to continue to be ossified for life. The causal relation is generic because both the skeletal part and the time-values involved are expressed through variables. Such (normality) causal relations are part of the background knowledge of the diagnostic theory. In addition, the theory can include generic (abnormality) `cause-spec`, at the level of properties, independently of the failure models.

It is a strength of our representation that failure models and normality can be uniformly represented. [Other approaches that integrate normality and failures are, for example, Console and Torasso (1990a) and Poole (1989b)]. At the implementation level, the causal and decomposition structures (comprised of time-objects) such as those illustrated in Figures 3 through 5 are translated into symbolic form. The symbolic representation of the SEDC model is given in Figure 6. The language used for expressing diagnostic theories and case histories is founded on the ontology of time-objects (properties, existences, openness, granularities, causal, temporal and decomposition relations, etc.) augmented with diagnostic concepts (failures/disorders, faults, observability, normality, etc.).

The overall causal network representing the diagnostic theory is partitioned into distinct causal models for the various failures. The partitioning is necessary in order to allow multiple, dynamic instantiations of the same failure, thus capturing recurring failures. There is only one model per failure; however, the same failure can appear as an ordinary causal state node in another failure's model. A *primary* cause[5] does not have any causal antecedents; primary causes are those failures which do not figure in

[5] Primary failures can be associated with relative, *a priori*, likelihoods of occurrence.

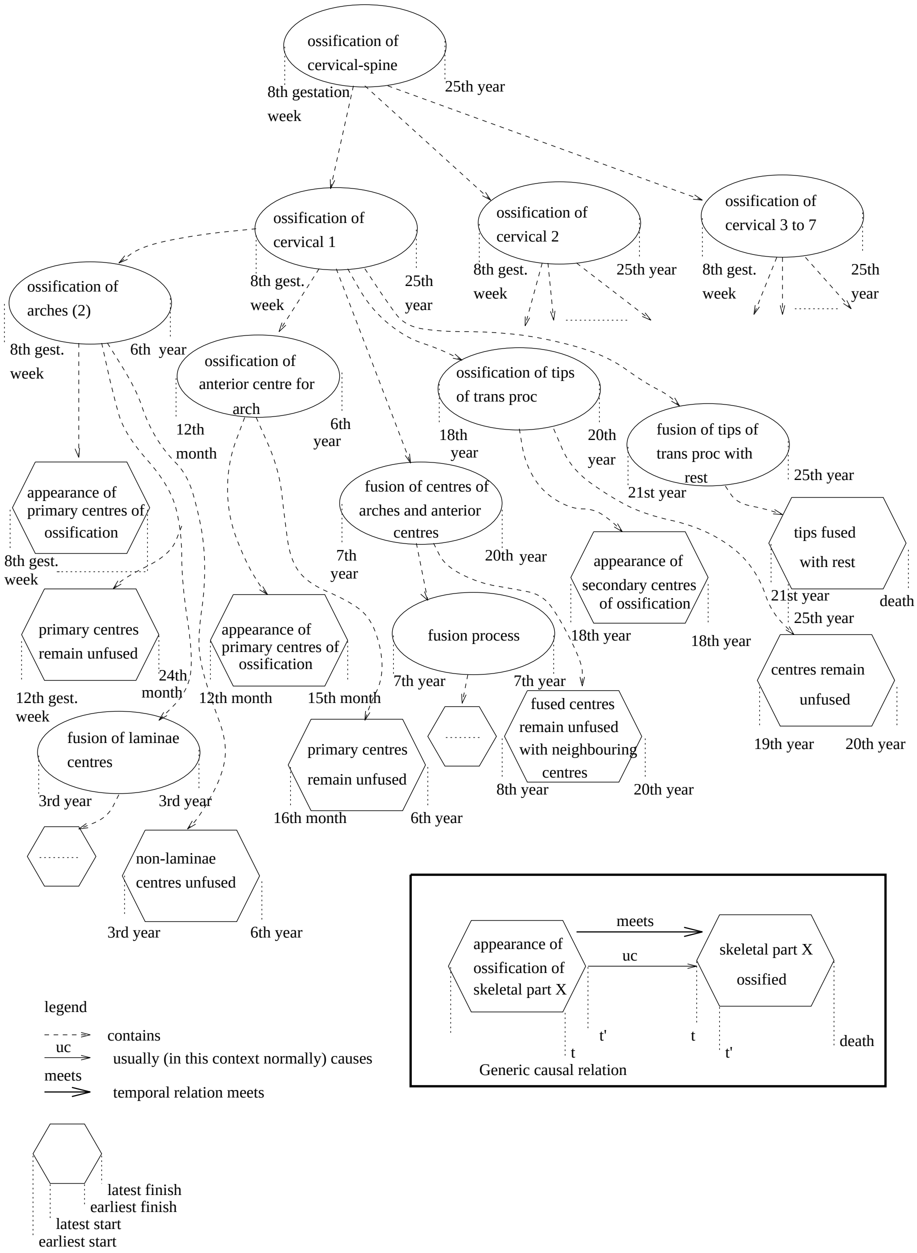

ossification of cervical-spine
8th gestation week
25th year
ossification of cervical 1
8th gest. week
25th year
ossification of cervical 2
8th gest. week
25th year
ossification of cervical 3 to 7
8th gest. week
25th year
ossification of arches (2)
8th gest. week
6th year
ossification of anterior centre for arch
12th month
6th year
ossification of tips of trans proc
18th year
20th year
fusion of tips of trans proc with rest
21st year
25th year
appearance of primary centres of ossification
8th gest. week
fusion of centres of arches and anterior centres
7th year
20th year
appearance of secondary centres of ossification
18th year
18th year
tips fused with rest
21st year
death
primary centres remain unfused
12th gest. week
24th month
appearance of primary centres of ossification
12th month
15th month
fusion process
7th year
7th year
25th year
centres remain unfused
19th year
20th year
fusion of laminae centres
3rd year
3rd year
primary centres remain unfused
16th month
6th year
fused centres remain unfused with neighbouring centres
8th year
20th year
non-laminae centres unfused
3rd year
6th year
legend
contains
uc
usually (in this context normally) causes
meets
temporal relation meets
latest finish
earliest finish
latest start
earliest start
appearance of ossification of skeletal part X
meets
uc
skeletal part X ossified
t'
t
t
t'
death
Generic causal relation

FIGURE 5. Modeling the normal behavior of an ossification process in terms of time-objects.

some other failure's model (and hence as intermediate nodes) but only in their own model. A primary failure can be observable, e.g., alcoholism, a primary failure causing cirrhosis (Console and Torasso 1991a). Different failure models are implicitly related through node sharing or explicitly related through secondary triggers (see below). We assume that a subset of failures defines *diagnoses*.

In Console and Torasso (1991a), the causal network is a fully connected structure that does not permit multiple instantiations of the same failure, and hence periodic

```
(disorder-model SEDC
   (main-exist lifetime (birth death closed))
   (unobservable)
   (usually-causes
      ((ossification-of pubic-bones)
         (meta-qualification delayed-process))
      ((limbs short)
         (main-exist lifetime (birth death closed))
         (observable)
            (necessarily-causes
               ((stature short)
                  (main-exist lifetime (birth death closed))
                  (observable))))
      ((vertebral-bodies variable-size)
         (main-exist lifetime (birth 1-year closed))
         (observable))
      ((knee-epiphyses absent-ossification)
         (main-exist lifetime (birth death closed))
         (observable))
      ((ossification-of epiphyses)
         (meta-qualification delayed-process))
      ((triradiate-cartilage wide)
         (main-exist lifetime (birth 11-years open-from-right))
         (observable))
      ((coxa-vara bilateral severe)
         (main-exist lifetime (birth death closed))
         (observable)
         (trend worsening)))
   (may-cause
      ((stillbirth) (observable) (necessarily-causes (death)))
      ((ossification-of cervical-spine)
         (meta-qualification delayed-process))
      ((scoliosis)
         (main-exist lifetime (birth death closed))
         (observable)
         (trend worsening))
      ((vertebral-bodies pear-shaped)
         (main-exist lifetime (birth 15-years open-from-right))
         (observable))
      ((platyspondyly mild)
         (main-exist lifetime (birth death closed))
         (observable))))
```

FIGURE 6. Symbolic representation of SEDC model.

failures cannot be dealt with. In addition, and of relevance to the preceding, "... one cannot deal with changing data and thus periodic findings; moreover, one cannot take into account the trend of the values of a parameter, which is usually a very important piece of information for diagnosticians" (Console and Torasso 1991a, p. 300). Temporal data abstraction is therefore not supported, nor are compound occurrences. Furthermore, the integration of models of correct behavior is not discussed in that work, the emphasis being on temporal constraint satisfaction over causal networks, whereas our emphasis is on integrating time within the objects in the diagnostic system. The temporal knowledge used specifies minimum and maximum delays for each causal arc, i.e., the minimum and maximum delay between the initiation of the cause and the initiation of its effect; the proposers argue that, in medical domains, only such temporal knowledge is usually available, and thus the temporal extents of nodes need to be dynamically derived, by reasoning backwards from the temporal extents of observations. However, on the basis of such delay information, the initiation margins for the various nodes, relative to the possibly unbound initiation of some initial node, can be predetermined, and thus in the worst case, only the termination margins are indefinite. Our time-ontology allows for such openness in the temporal extents of time-objects, as already illustrated. However, margins for the duration of nodes in some failure model *can* be available as the example domain has demonstrated, and if they are not available, the default margins for the durations of the relevant properties can be used; hence margins for the

```
(case-history
    (patient-name ... )
    (date-of-birth ... )
    (sex ... )
    (race ... )
    (consanguinity ... )
    (radiological-findings
       ((carpal-bones small)
           (main-exist case-lifetime (10-years 10-years closed)))
       ((femoral-capital-epiphyses abnormal)
           (main-exist case-lifetime (2-years 2-years closed)))
       ((femoral-capital-epiphyses flat irregular)
           (main-exist case-lifetime (7-years 7-years closed)))
       ((vertebral-end-plates irregular)
           (main-exist case-lifetime (7-years 7-years closed)))))
```

FIGURE 7. Symbolic representation of a case history.

terminations of nodes also can be predetermined. Bounding the temporal extents of nodes in a causal model reduces the computational complexity of a temporal constraint propagator; furthermore, it avoids the need to impose unrealistic temporal assumptions for reducing the computational overheads, such as a cause cannot "outlive" its effect, or there cannot be a gap between a cause and its effect, etc.

Case History. Next we explain how a case history can be represented in terms of time-objects. As discussed earlier, a case's temporal database consists of temporal assertions, associations between properties, and spans of valid time. The relevant history of a case consists those assertions whose valid time is covered by the concrete time-axis corresponding to the time window of the diagnostic activity—in this way, "irrelevant" assertions, e.g., assertions in the remote past of the case, are screened out.[6] Each selected temporal assertion is a (concrete) time-object. Thus the (relevant) case history is a collection of time-objects. We assume that the number of time-objects in a case history is kept to the minimum possible by performing appropriate merges as well as other forms of temporal data abstraction on the raw time-objects (Keravnou 1996b). Furthermore, potential causality dependencies between these time-objects are investigated through the application of Axiom 3, and where a `causality-link` is established to hold, it is appropriately instantiated. Some of the time-objects comprising the case history are *contextual*; i.e., they do not need any explanation. These usually assert past failures or past or ongoing therapeutic actions. Figure 7 gives the symbolic representation of the case history mentioned in the introduction. In this case, all the time-objects comprising the diagnostic problem are point-objects at the granularity of years, the time-unit of the particular concrete time-axis (named *case-lifetime*).

Therapeutic Actions. A pure diagnostic system is not required to plan and monitor the execution of therapeutic actions in parallel to trying to reach a diagnostic solution. Still, such a system should have an understanding of the notion of a therapeutic action or, more generally, the notion of an action. If the case history under consideration mentions past or ongoing therapeutic actions, the system should understand their effects. For example, the model of a failure may be different in the context of such actions [associated faults are nullified or accentuated (prolonged)].

Knowledge about therapeutic actions is part of the background component of the diagnostic theory. Each generic therapeutic action is represented in terms of the action

[6] For some applications dealing with monitoring rather than diagnosis, the concrete time-axis (time window) is a moving one in the sense that the position of its origin is continuously moving in a forward direction, and thus assertions whose valid time is no longer covered by the time-axis are "forgotten" (Dojat and Sayettat 1994).

per se, its preconditions and effects. All these are time-objects; the existences of the preconditions/effects are given relative to the existence (i.e., initiation) of the action (Keravnou 1996d). At a generic level, the action is related with its effects through instances of relation `causes`. Knowledge of preconditions of actions is relevant to the task of a therapy planner but not to the task of a diagnostic system that only needs to be able to understand potential interactions between (past or ongoing) actions and conjectured failures. Thus a diagnostic system only needs to know the potential effects of such actions.

For every action in a case history, Axiom 2 is applied to each of the `causes` relations between the action and its effects in order to decide which `causality-links` actually hold; the effects corresponding to these links are also recorded in the case history. Entering such time-objects (effects of therapeutic actions) may result in revoking or clipping the persistence of predicted observations in the case history, if any.

3.3. Diagnostic Solutions: Instantiating Failure Models

At any time t, there are a number of potential diagnostic solutions, or *hypothetical worlds*. (For the rest of the article, the terms *potential diagnostic solution* and *hypothetical world* are used interchangeably.) A hypothetical world consists of instantiated failure models; hence it can be abstracted to the time-objects comprising the starting states of the included failure models, the failures per se. This subsection discusses (1) mechanisms for the context-free (through primary triggers) and context-sensitive (through secondary triggers) formation and extension of hypothetical worlds and (2) the dynamic integration of the components of a hypothetical world and their tailoring against relevant actions recorded in the case history.

Temporal-Abductive Triggering Mechanisms. There are three ways to trigger (abduce) a failure model: (1) through primary triggers, (2) through secondary triggers, and (3) through another failure's instantiation that includes the given failure as a causal state node.

A failure is associated with a number of *primary triggers*. In some theories, every observable node in a causal model potentially could act as a primary trigger. However, for better focusing and higher compatibility with human diagnostician practices, the role of a primary trigger is reserved to a small subset of these nodes. A primary trigger is some cheap, easily obtainable information (e.g., striking abnormality, observable primary cause, contextual information, etc.) that directs attention to the particular failure. The primary triggers for SEDC and morquio are illustrated in Figures 8 and 9, respectively, as (abstract) time-objects. Comparing the primary triggers with the corresponding manifestations (faults) in the SEDC and morquio models (see Figures 3 and 4), it can be seen that the primary triggers are less restrictive with respect to both their properties and their existences. The existence of most of the depicted triggers is in fact the default "at any time" because most of the actual triggers are expected to be point-objects. The primary trigger "platyspondyly at any time" that is associated with both disorders is considerably less restrictive than the corresponding manifestations that are "mild platyspondyly from birth onward" for SEDC and "platyspondyly throughout from the age of 1 year onward" for morquio. Furthermore, for two of the SEDC triggers, the significant information, regarding their existence, is the start point. Thus the given triggers read "short stature from birth until any time" and "skeletal abnormality from birth until any time." Finally, the progression "worsening" associated with property

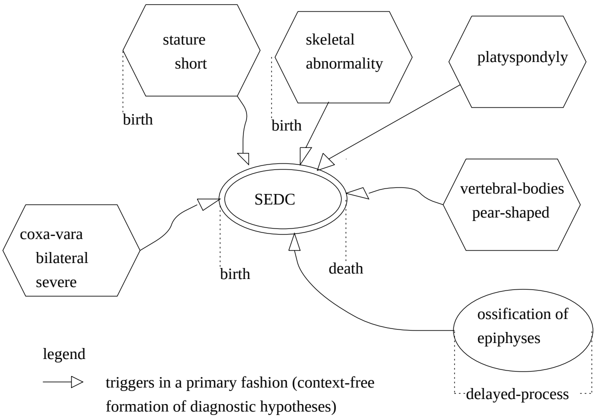


FIGURE 8. Primary triggers for SEDC.

“bilateral, severe, coxa-vara” in the SEDC model is not included in the corresponding trigger. Similarly, two of the morquio triggers relate to the disorder’s expectation regarding the progressive resorption of femoral-capital epiphyses. Once again, the triggers are considerably less restrictive. For example, the temporal constraints specifying the particular form of resorption are missing. Thus primary triggers, by virtue of being less restrictive than corresponding faults in a failure model, simply provide heuristic guidance in the generation of diagnostic hypotheses, and they are by no means infallible; after all, the same trigger can be associated with many failures. For example, the hypothesis of morquio will be triggered on the basis of “femoral-capital epiphyses absent at-birth” despite the fact that this hypothesis is in conflict with this observation. It does not matter if unlikely hypotheses are triggered, provided that the subsequent evaluation will result in their rejection. However, it does matter if likely hypotheses are not triggered, and hence there must be alternative, simultaneously applied ways of forming hypotheses. Primary triggers represent a context-free mechanism for the formation of hypotheses, since the formation is done independently of any other hypothesis.

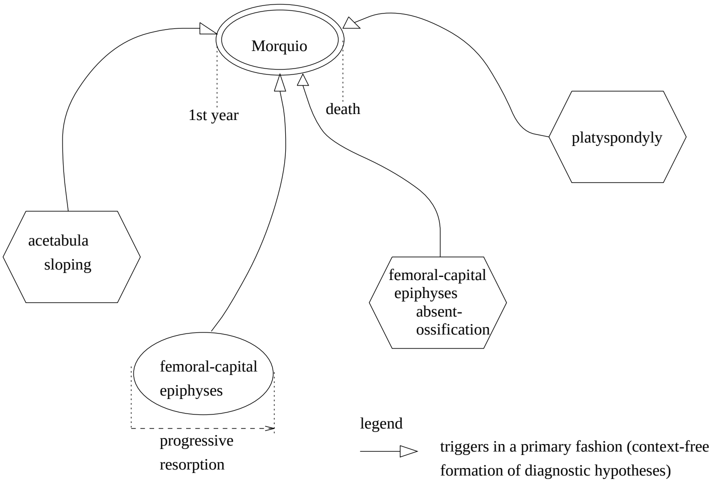


FIGURE 9. Primary triggers for morquio.

The other mechanisms to be discussed are context-sensitive. The notion of a primary trigger, as a prime mechanism for the formation of hypotheses, has been used in early abductive diagnostic systems (Thagard 1992). The contribution of our approach is in having *temporal* primary triggers.

Formally, a primary trigger for some failure Φ is expressed as the triple $\langle \tau, \mathit{conds}, f_I \rangle$, where τ is an abstract time-object (possibly compound), *conds* is a list of conditions, and f_I is an instantiation function. For example, the symbolic representation of the "platyspondyly" trigger for SEDC is

```
(primary-trigger
    ((property platyspondyly) (exist lifetime *any-time*))()
    ((disorder SEDC) (exist lifetime (birth death closed)))
```

In this trigger, no conditions are specified. The semantics are that if τ's abstract existence can be mapped onto the concrete time-axis used in the case history, the case history accounts for concrete-τ (predicate *accounts-for* is defined in Section 3.4), and all the specified conditions are satisfied (by the case history), the instantiation function f_I is applied to the abstract model of Φ, and concrete-τ, to return the concrete instantiation of Φ i.e., to determine the particular existence of the failure on the concrete time-axis. Thus, if the abstract time-axis used in the definition of the failure model cannot be mapped onto the concrete time-axis, the particular failure cannot be triggered (none of its primary triggers can be activated), and hence it cannot be instantiated under any potential solution for the given diagnostic problem; in short, the period of time of relevance to the given failure is outside the temporal scope of the particular diagnostic activity. Consider, for example, a case for the SDD system where the first observations of abnormality are positioned around the age of 7 years. It is therefore reasonable for the concrete time-axis to start at the age of 5 years; any information preceding this time-point is considered irrelevant. As a result, the hypothesis of SEDC or morquio cannot be formed for this case because the abstract time-axes for the models are not fully mappable onto the concrete time-axis, and as a result, the initiations of these disorders cannot be positioned on the concrete time-axis.

Processing a primary trigger usually entails reasoning backwards in time; the trigger is (loosely) related to a potential effect of the failure, and based on the trigger's observed existence, the existence of (a particular instantiation of) the failure is hypothesized. Multiple primary trigger activations for the same failure model are possible. Some of these could refer to the same trigger, thus capturing recurring events.

Secondary triggers interrelate failure models. They are of two types, *complementary* and *opposing*. A complementary secondary trigger suggests the instantiation of another failure in conjunction with some instantiated failure. An opposing secondary trigger suggests the replacement of some failure with another failure. The format of a secondary trigger for some failure Φ is $\langle \tau, \mathit{conds}, f_I, \Phi' \rangle$, where τ is an abstract time-object, *conds* is a list of conditions, f_I is an instantiation function, and Φ' is a complementary/alternative failure. Its semantics is similar to that of a primary trigger. Figure 10 illustrates a secondary, opposing trigger from morquio to SEDC that expresses the fact that the two disorders may be distinguished on the grounds of their presentation time; SEDC presents from birth, while morquio presents from the age of 1 year. Thus, if morquio is hypothesized for some case and new evidence points to the presence of some skeletal abnormality from birth, the competing hypothesis of SEDC is also formed. In SDD, secondary triggers are rarely used in the formation of hypotheses for the simple reason that primary triggers rarely miss a likely hypothesis. However, secondary triggers play a very important role in this system when differentiating competing hypotheses (Keravnou et al. 1994).

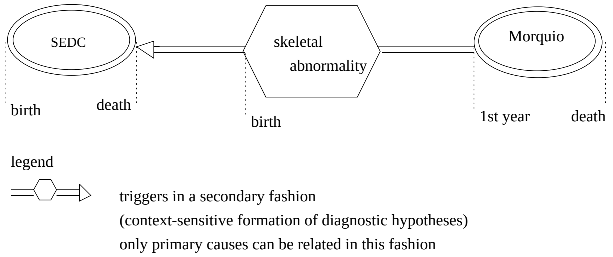

SEDC
skeletal abnormality
Morquio
birth
death
birth
1st year
death
legend
triggers in a secondary fashion
(context-sensitive formation of diagnostic hypotheses)
only primary causes can be related in this fashion

FIGURE 10. Secondary opposing trigger from morquio to SEDC.

Integrating and Tailoring a Hypothetical World. A hypothetical world consists of a number of failure model instantiations. It is *strongly integrated*, and hence the diagnostic solution it represents is coherent, if it does not consist of disconnected parts. Potential interactions between nodes of different failure (model) instantiations, in the same hypothetical world, can be determined dynamically as follows[7]: First, Axiom 3 is applied with respect to pairs of relevant nodes (of different failure instantiations) to see if any `causality-links` can be established. Second, nodes sharing the same property are investigated. If the existences of such nodes are not disjoint, the possibility of merging them into a single shared node is explored. If the existences are sufficiently disjoint to represent distinct occurrences of the given property, a loose connection may be created by collating all the distinct occurrences into a compound time-object representing the recurrence of the given property [this collating may even reveal a periodic occurrence by detecting some regularity—a recurrence pattern (Keravnou 1997)].

The investigation of the existence of links between apparently unrelated components of a hypothetical world is done with the objective of obtaining a more coherent potential solution. In the same manner, potential interactions, this time both additive and subtractive, between (therapeutic) action effects recorded in the case history and components of a hypothetical world are investigated. A potential solution is therefore tailored against contextual information in the case history. Basically, Axiom 3 is applied to see if an effect of some action causes a property shared by some abnormality state, or more simply, the effect has the same property as an abnormality state and the two overlap. These are examples of additive interactions where an action, through one of its effects, accentuates an abnormal situation. The detected accentuation from the "expected" (i.e., without external interference) evolution of a failure process should be noted in the relevant hypothetical world. A therapeutic action accentuates an abnormal situation through one of its adverse side effects; normally, however, a therapeutic action aims to nullify an abnormal situation through its targeted consequences. This is the case of a subtractive interaction between an action effect and an abnormal state; the two overlap and represent exclusive (or nullifying) properties. Here, depending on the strength of the subtractive interaction, the abnormal state and all the subsequent abnormalities that depend entirely on it are revoked or suitably reduced (their persistence is clipped).

[7] We assume that a hypothetical world is consistent; i.e., it does not include temporally overlapping nodes with mutually exclusive properties. Conflicts can be resolved if the properties concerned neutralize each other, e.g., "water retention" and "water loss." In such cases, the relevant nodes (time-objects) are replaced with (or become the components of) a single neutral node.

3.4. Accountings[8] and Conflicts

In this section we define binary predicates *accounts-for* and *in-conflict-with* that are subsequently used for the definition of the primitive evaluation criteria. Each (primitive) evaluation criterion measures, from some conceptual perspective, the goodness of fit between the case history and some potential diagnostic solution.

Predicates *accounts-for*(τ_i, τ_j) and *in-conflict-with*(τ_i, τ_j) take time-objects as arguments. Instances of these predicates are evaluated with respect to some consistent collection of time-objects and their interrelationships, the *evaluation domain*, e.g., the case history, or a hypothetical world. By default, this is taken to be the domain of the first argument, and thus we refer to ground instances:

- *accounts-for*(τ_i, τ_j). Time-object τ_i's assertion, in the given evaluation domain, can account for time-object τ_j being asserted in the same evaluation domain. The predicate is reflexive and transitive.
- *in-conflict-with*(τ_i, τ_j). Time-object τ_i's assertion, in the given evaluation domain, denies the assertion of time-object τ_j in the same evaluation domain. The predicate is symmetric.

These predicates can be instantiated for: a failure or therapeutic action accounting for some (observed) fault, a failure or therapeutic action conflicting with (observed) faults, (observed) faults satisfying/refuting expected manifestations of failure hypotheses, (observed) faults satisfying/refuting required preconditions/expected effects of hypothesized actions. They are defined through the following Axioms.

Axioms for predicates accounts-for and in-conflict-with.

4: *accounts-for* $(\tau_i, \tau_j) \Leftrightarrow ((\pi(\tau_i) \Rightarrow \pi(\tau_j)) \wedge \tau_j \subset= \tau_i)$

5: *accounts-for* $(\tau_i, \tau_j) \Leftrightarrow \exists\tau_k\ (\texttt{contains}(\tau_i, \tau_k) \wedge$ *accounts-for* $(\tau_k, \tau_j))$

6: *accounts-for* $(\tau_i, \tau_j) \Leftrightarrow \exists\tau_k\ (\texttt{causality-link}\ (\tau_i, \tau_k) \wedge$ *accounts-for* $(\tau_k, \tau_j))$

7: *accounts-for* $(\tau_i, \tau_j) \Leftrightarrow \exists\tau_k, \tau_n\ (\texttt{causality-link}\ (\tau_i, \tau_k) \wedge \texttt{contains}\ (\tau_n, \tau_k)$ $\wedge$ *accounts-for* $(\tau_n, \tau_j) \wedge$ *assumed* $(\tau_n))$

8: *in-conflict-with* $(\tau_i, \tau_j) \Leftrightarrow (\texttt{excludes}\ (\pi(\tau_i), \pi(\tau_j)) \wedge \neg\ (\tau_i\ \texttt{disjoint}\ \tau_j))$

9: *in-conflict-with* $(\tau_i, \tau_j) \Leftrightarrow \exists\tau_k\ (\texttt{contains}(\tau_i, \tau_k) \wedge$ *in-conflict-with* $(\tau_k, \tau_j))$

10: *in-conflict-with* $(\tau_i, \tau_j) \Leftrightarrow \exists\tau_k\ (\texttt{casuality-link}\ (\tau_i, \tau_k) \wedge$ *in-conflict-with* $(\tau_k, \tau_j))$

11: *in-conflict-with* $(\tau_i, \tau_j) \Leftrightarrow \exists\tau_k, \tau_n\ (\texttt{casuality-link}\ (\tau_i, \tau_k) \wedge \texttt{contains}\ (\tau_n, \tau_k) \wedge$ *in-conflict-with*$(\tau_n, \tau_j) \wedge$ *assumed*$(\tau_n))$

Predicate *assumed*, used in Axioms 7 and 11, is defined as follows:

$$\textit{assumed}(\tau_i) \Leftarrow \exists\tau_j\ (\textit{assumed}\ (\tau_j) \wedge\ \texttt{causality-link}\ (\tau_j, \tau_i))$$

$$\textit{assumed}(\tau_i) \Leftarrow \neg\ (\exists\tau_j\ (\texttt{contains}(\tau_i,\ \tau_j) \wedge \neg\ \textit{assumed}(\tau_j)))$$

$$\textit{assumed}(\tau_i) \Leftarrow \exists\tau_j\ (\texttt{contains}(\tau_j,\ \tau_i) \wedge\ \textit{assumed}(\tau_j))$$

$$\textit{assumed}(\tau_i) \Leftarrow \neg(\exists\tau_j\ \texttt{causes}(\tau_j,\ \tau_i, \_, \_))$$

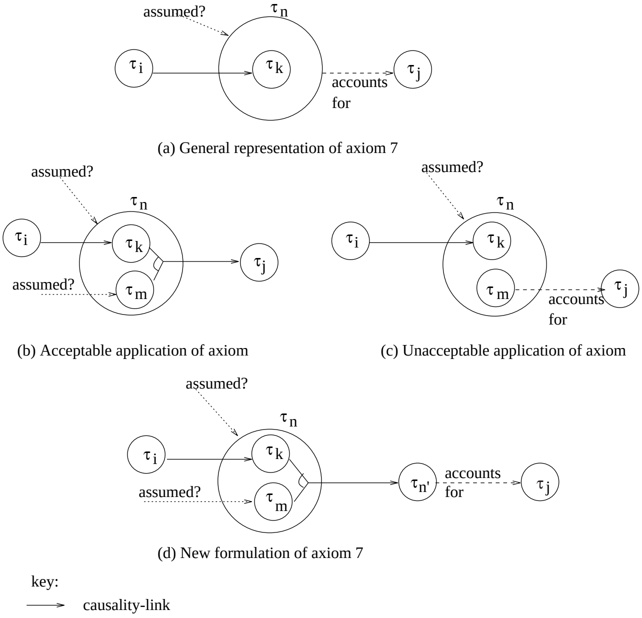
Thus a time-object τ_i accounts for another time-object τ_j either directly (Axiom 4) or indirectly through one (if any) of its component time-objects (Axiom 5) or one (if

[8] In the absence of a better term, we use *accounting* as a noun for that which accounts for something. It seems to be a better antonym for *conflict* than is *explanation*.

any) of its established causal consequent time-objects (Axiom 6). For a direct accounting, the property of τ_i implies (i.e., subsumes) the property of τ_j (recall that function π gives the property of a time-object), and the existence of τ_i covers completely the existence of τ_j (this is expressed in terms of the temporal relation $\subset=$). Partial accountings are not dealt with. Axiom 7 warrants closer inspection. The general representation of the Axiom is illustrated in Figure 11(a). The same figure gives an acceptable application of the Axiom (Figure 11b), as well as an unacceptable application of it (Figure 11c). Under the acceptable scenario, τ_i is an established causal antecedent for τ_k, which in turn constitutes part of an established compound causal antecedent for τ_j; strictly speaking, τ_i accounts only partly for τ_j. Under the unacceptable scenario, τ_i is an established causal antecedent for τ_k, but τ_k plays no part in establishing the accounting of τ_j by τ_n; this is done through another component of τ_n, τ_m (and via Axiom 5 it can be inferred that τ_n accounts for τ_j). In short, although a time-object can inherit from one of its components some accounting relation, the opposite is not true unless all its components are collectively involved in establishing the particular accounting relation; the latter can only happen in the case of compound causal antecedents. Hence, to exclude unacceptable scenarios, Axiom 7 can be reformulated as follows (see Figure 11d):

7: $\textit{accounts-for}\ (\tau_i,\tau_j) \Leftrightarrow \exists\ \tau_k,\tau_n,\tau_{n'}\ (\texttt{causality-link}(\tau_i,\tau_k) \wedge \texttt{contains}(\tau_n,\tau_k) \wedge \texttt{causality-link}(\tau_n, \tau_{n'}) \wedge \textit{accounts-for}(\tau_{n'}, \tau_j) \wedge \textit{assumed}(\tau_n))$

Predicate *in-conflict-with* is similarly defined to predicate *accounts-for*. Thus a time-object τ_i is in conflict with another time-object τ_j either directly (Axiom 8) or indirectly (Axioms 9 through 11). For a direct conflict, the properties of τ_i and τ_j are mutually exclusive (expressed in predicate `excludes`), and the existences of τ_i and τ_j are not disjoint (expressed in temporal relation `disjoint`). Axiom 11 corresponds to Axiom 7, and hence a similar problem arises to the one discussed above for Axiom 7. Thus



(a) General representation of axiom 7

(b) Acceptable application of axiom

(c) Unacceptable application of axiom

(d) New formulation of axiom 7

FIGURE 11. Axiom 7 (for the derivation of predicate *accounts-for*).

Axiom 11 can be reformulated as follows:

11: *in-conflict-with*$(\tau_i, \tau_j) \Leftrightarrow \exists\, \tau_k, \tau_n, \tau_{n'}$ (`causality-link` $(\tau_i, \tau_k) \wedge$ `contains` $(\tau_n, \tau_k) \wedge$ `causality-link` $(\tau_n, \tau_{n'}) \wedge$ *in-conflict-with* $(\tau_{n'}, \tau_j) \wedge$ *assumed* (τ_n))

Predicates *accounts-for, in-conflict-with*, and *assumed* are computationally expensive. Each of these entails branching based on causes and containment arcs. More specifically, *accounts-for*(τ_i, τ_j)/*in-conflict-with*(τ_i, τ_j) generates a search that grows forward in time with the objective of piecing together an admissible path from node τ_i to some node τ_k, say, that directly subsumes/conflicts with τ_j (τ_i could be the same as τ_k); the initial node τ_i is hypothesized or directly assumed, e.g., observed. An admissible path is defined below.

Definition 1: Admissible Path. A sequence of time-objects $\tau_1, \tau_2, \ldots, \tau_n$ forms an admissible path iff

$$\forall i = 1, \ldots, (n-1)(\texttt{causality-link}\ (\tau_i, \tau_{i+1}) \vee \texttt{contains}\ (\tau_i, \tau_{i+1}) \vee \texttt{isa-component-of}(\tau_i, \tau_{i+1}))$$

Let τ_i, τ_j, and τ_k be three consecutive time-objects on some admissible path, and let R_{ij} denote the relation from τ_i to τ_j, and let R_{jk} denote the relation from τ_j to τ_k. On the basis of the new formulations of Axioms 7 and 11, it follows that

$$(R_{ij} \neq \texttt{isa-component-of}\ \vee R_{jk} \neq \texttt{contains}\ ).$$

A `causality-link` is established through Axiom 2 or Axiom 3, each of which ensures that any conditions are satisfied (by the case history). Likewise, a conditional component is only used if the relevant conditions are satisfied, as dictated by Axiom 1.

`causality-link` can be associated with uncertainty. Thus, strictly speaking, an accounting relation between τ_i and τ_j is not categorical, but it has a degree of uncertainty given, say, by the product of all the causality uncertainties involved in the admissible path from τ_i to τ_j (similarly for a conflicting relation). Hence, if there is more than one admissible path from τ_i to τ_j, the path with the minimum uncertainty (maximum certainty) ought to be used; this entails computing all the admissible paths, which is even more expensive computationally. Since uncertainty is not the issue here, for reasons of simplicity, accountings and conflicts are considered certain.

The derivation of predicate *assumed*(τ_i) generates a search that grows backward in time with the objective of piecing together (in a backward fashion) an admissible path from some node τ_j, which may be directly assumed, to node τ_i. A node is directly assumed, in some context (say, a hypothetical world), if it does not have any potential causal antecedents. This is not to say that the particular node is necessarily a primary cause, just that it may be considered a "starting" state in the particular context. A compound node is assumed if none of its expected components is revoked (i.e., each of them may be assumed). Similarly, a node is assumed either because it is contained in an assumed node or because it has an assumed direct causal antecedent.

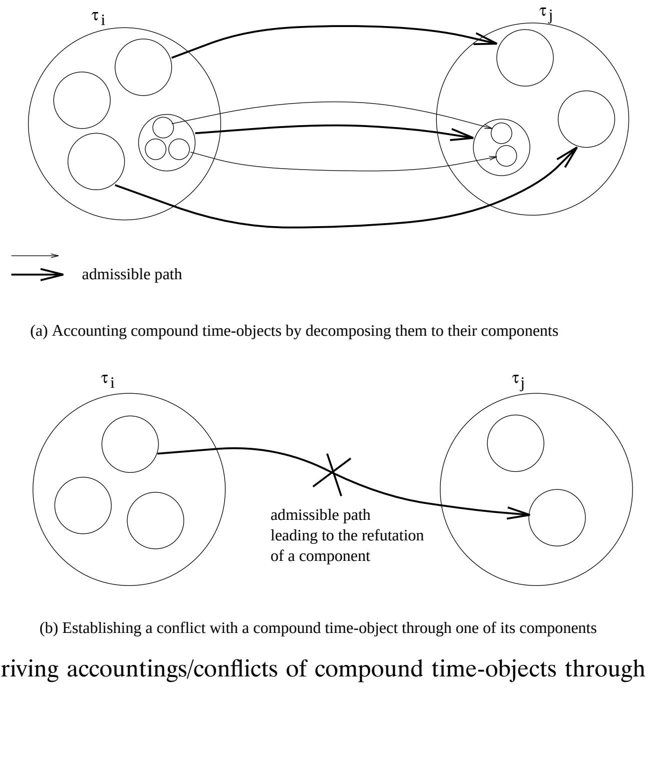
Axioms 4 through 11 treat their second argument as an atomic time-object. The following complementary Axioms apply to compound time-objects and derive an accounting/conflict through their components:

Axiom 12. *accounts-for*$(\tau_i, \tau_j) \Leftrightarrow \forall\tau_k$ s.t. `contains`(τ_j, τ_k) {accounts-for(τ_i, τ_k)}

Axiom 13. *in-conflict-with*$(\tau_i, \tau_j) \Leftrightarrow \exists\ \tau_k$ s.t. `contains`(τ_j, τ_k) {in-conflict-with (τ_i, τ_k)}

The application of Axiom 12 results in constructing an admissible path to each component of τ_j that emanates from τ_i (or some component of it). If a component of τ_j is itself a compound time-object, the application of Axiom 12 can be repeated with respect to that component (see Figure 12a). The application of Axiom 13 results in constructing at least one admissible path emanating from (some component of) τ_i and leading to some time-object (in the evaluation domain) that directly conflicts with a component of τ_j (see Figure 12b). The case history *CH* or some potential diagnostic solution S can be viewed as compound time-objects containing all the time-objects that comprise them. Hence we can formulate, in a simple way, compound queries such as "*accounts-for*(*S*, *O*)?" or "*in-conflict-with*(*CH*, *S*)?" where *O* is the compound time-object comprising all the abnormal observations. The evaluation domains for these queries consist of the single time-objects *S* and *CH*, respectively.

Figure 13 gives a high-level layered organization of the various Axioms that play a part in the evaluation of the potential diagnostic solutions. The processing elements are time-objects. At the bottom layer we have the Axioms (1–3) for deriving relations `causality-link` and `contains`, both of which involve temporal constraints. The role of these Axioms is in integrating the components of a solution. At the next layer we have the Axioms (4–13) that define predicates *accounts-for* and *in-conflict-with* (and the auxiliary predicate *assumed*). Relations `causality-link` and `contains` play a central role in the derivation of accountings and conflicts between time-objects. The latter form the basis for the definition of the primitive evaluation criteria that occupy the next layer of this organization. As we move up this hierarchy, the handling of time becomes less visible as it gets hidden behind the axioms at the lower layers. Temporal reasoning is clearly visible in the derivation of `causality-links` and in the basic axioms for *accounts-for* and *in-conflict-with*. However, the definitions of the primitive criteria do not address time directly. This abstraction is a strength, since the same



(a) Accounting compound time-objects by decomposing them to their components

(b) Establishing a conflict with a compound time-object through one of its components

FIGURE 12. Deriving accountings/conflicts of compound time-objects through their components.

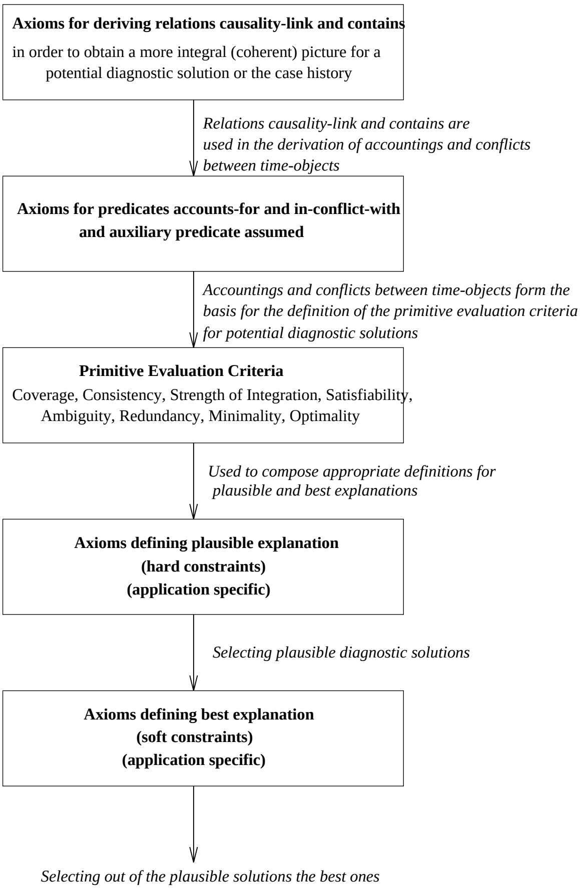

Axioms for deriving relations causality-link and contains
in order to obtain a more integral (coherent) picture for a potential diagnostic solution or the case history
Relations causality-link and contains are used in the derivation of accountings and conflicts between time-objects
Axioms for predicates accounts-for and in-conflict-with and auxiliary predicate assumed
Accountings and conflicts between time-objects form the basis for the definition of the primitive evaluation criteria for potential diagnostic solutions
Primitive Evaluation Criteria
Coverage, Consistency, Strength of Integration, Satisfiability, Ambiguity, Redundancy, Minimality, Optimality
Used to compose appropriate definitions for plausible and best explanations
Axioms defining plausible explanation (hard constraints) (application specific)
Selecting plausible diagnostic solutions
Axioms defining best explanation (soft constraints) (application specific)
Selecting out of the plausible solutions the best ones

FIGURE 13. Layered organization of axioms.

definitions can apply to atemporal solutions. The higher two layers, built on top of the primitive criteria, are application-specific; they specify the appropriate notions for plausible and best explanations.

3.5. Deduction within Abduction

In concluding this section we explain how, in the proposed framework, deductive reasoning is employed in the context of abductive reasoning. The overall reasoning performed is abductive because the goal is to generate the best explanation of some collection of observations of abnormality. In achieving this goal, though, deductive inferencing is necessary. As Pople (1973) points out in his classic paper on the mechanization of abduction, to which the interest of the AI community in abductive inference is attributed, in a deduction, the objective is to determine *whether* some statement is true. In an abduction, the objective is to determine *why* something is true (i.e., why the observed abnormalities hold). In answering the *why* question, it is obviously important to be able to determine *whether*; thus deduction may be considered to be a process subordinate to abduction (Pople 1973, p. 150).

Abduction is far more complicated than deduction. A queried statement may be deduced (derived) in a multitude of ways, and any of these suffices; effective deductive systems are able to follow the simplest derivation paths, but this is an implementation rather than a conceptual issue. In abduction, it is not sufficient just to generate one plausible explanation of the observed situation; instead, all plausible explanations need to be compared and contrasted. An explanation is usually not deducible, and so once an explanation is hypothesized, it is not possible to deduce it.

In this section we have presented the basic mechanisms for the formation and integration of potential diagnostic solutions (explanations). The processing of primary and secondary triggers is clearly abductive in nature. However, the overall instantiation of a failure model and its integration with other failure instantiations makes extensive use of Axioms 1 to 3, which are used deductively. In this section we have also defined predicates *accounts-for* and *in-conflict-with* (Axioms 4–13) that form the foundations for the definition of the primitive evaluation criteria to be discussed in the next section. While the notion of "evaluation of competitors" is not relevant to deductive reasoning, the evaluation of competing explanations is a critical aspect of abductive reasoning. The primitive evaluation criteria represent constituents of higher-level abductive inferences. Within these criteria, though, Axioms 4 to 13 are used in a deductive manner.

4. EVALUATION CRITERIA FOR DIAGNOSTIC SOLUTIONS

In this section we present a number of primitive, general criteria for the evaluation of potential diagnostic solutions (hypothetical worlds), as summarized in Table 1. In our framework, a potential diagnostic solution is a collection of interrelated time-objects (corresponding to failure model instantiations) some of which are designated as *diagnoses*—these correspond to failure time-objects. A central evaluation question, in an abductive context, is "Does this diagnosis account for the observation of this fault?" Hence, for evaluation purposes, a potential diagnostic solution is abstracted to its set of diagnoses, or *hypotheses*. A potential diagnostic solution is a dynamic entity; the contents of some solution S_i at time t is denoted by $S_{i,t}$.

The case history CH is also a dynamic entity. Its contents at time t is denoted by CH_t. This is partitioned into the dynamic subsets of *focus-abnormalities* (faults to be accounted for) F_t and *contextual information* C_t; contextual information, including normal observations, does not need to be explained. F_t and C_t are disjoint sets. The elements of F_t are assumed to be mutually independent. Thus,

$$CH_t = F_t \cup C_t.$$

Focus-abnormalities can be classified into *hard* abnormalities and *soft* abnormalities. Hard abnormalities are serious abnormalities, whereas soft abnormalities are moderate or mild abnormalities, some of which may in fact be attributable to "natural" causes. Whether a time-object represents a hard or a soft abnormality depends on its property and/or duration. For example, headache for an hour is a soft abnormality, but continuous headache for days is a hard abnormality. A competent diagnostic system should be able to decide by itself whether a focus-abnormality is soft or hard. Furthermore, F_t can be partitioned into current and past focus-abnormalities. Thus,

$$F_t = \text{hard}_t \cup \ \text{soft}_t$$

$$F_t = \text{past}_t \cup \ \text{current}_t.$$

In the following discussion, any reference to "a diagnostic solution" means "a potential diagnostic solution." Also, for reasons of simplicity, we use the symbol S_t to mean $S_{i,t}$.

TABLE 1. Primitive Evaluation Criteria For Potential Diagnostic Solutions

SUPPLY TABLE

4.1. Coverage and Consistency

Definition 2. A diagnostic solution S_t has *focus-coverage* iff it accounts for every focus-abnormality:

$$\textit{focus-coverage}(S_t) \Leftrightarrow \forall f_i \in F_t \; \exists h_j \in S_t \text{ s.t. } \textit{accounts-for}(h_j, f_i).$$

By treating S_t and F_t as compound time-objects, the definition of focus-coverage is simplified to

$$\textit{focus-coverage}(S_t) \Leftrightarrow \textit{accounts-for}(S_t, F_t).$$

Definition 3. A diagnostic solution S_t has *hard-coverage* iff it accounts for every hard focus-abnormality:

$$\textit{hard-coverage}(S_t) \Leftrightarrow \forall f_i \in \textit{hard}_t \; \exists h_j \in S_t \text{ s.t. } \textit{accounts-for}(h_j, f_i).$$

Definition 4. A diagnostic solution S_t has *current-coverage* iff it accounts for every current, hard focus-abnormality:

$$\textit{current-coverage}(S_t) \Leftrightarrow \forall f_i \in (\textit{current}_t \cap \textit{hard}_t) \; \exists h_j \in S_t \text{ s.t. } \textit{accounts-for } (h_j, f_i).$$

Axiom 14. focus-coverage $(S_t) \Rightarrow$ hard-coverage(S_t)

Axiom 15. hard-coverage$(S_t) \Rightarrow$ current-coverage(S_t)

Definition 5. A diagnostic solution S_t is *case-consistent* iff it is not in conflict with any of the data in the case history:

$$\textit{case-consistent}(S_t) \Leftrightarrow \forall d_i \in CH_t \; \neg(\exists h_j \in S_t \text{ s.t. } \textit{in-conflict-with}(h_j, d_i))$$

or

$$\textit{case-consistent}(S_t) \Leftrightarrow \neg \textit{in-conflict-with}(S_t, CH_t)$$

by treating S_t and CH_t as compound time-objects.

Thus a datum is assumed, by default, to be consistent with some diagnostic solution, unless the latter entails a direct conflict with that datum. This applies to both observations of normality and abnormality. Although this is the standard approach with respect to observations of normality, it is not so for observations of abnormality if the closed-world assumption is applied (see Section 4.5). However, by assuming consistency, in a sense we acknowledge the inherent incompleteness of our diagnostic theories.

Definitions 6 to 9. A diagnostic solution S_t can be similarly defined to be consistent with the focus-abnormalities (*focus-consistent*), the contextual data (*context-consistent*), the hard focus-abnormalities (*hard-consistent*), or the hard current focus-abnormalities (*current-consistent*).

The most restrictive form of coverage is focus-coverage; hard-coverage and current-coverage represent successively more relaxed forms of coverage (Axioms 14 and 15), and in fact, other forms of coverage may be defined. Similarly, the most restrictive form of consistency is case-consistent, whereas focus-consistent, hard-consistent, etc. represent reduced forms of consistency. The ideal expectation is for the concluded diagnostic solution to have focus-coverage and to be case-consistent. In real life, especially for medical diagnostic problems, this expectation is rarely attained, and thus explanation plausibility may not require focus-coverage. Console and Torasso (1991b) propose a way of integrating abductive and consistency-based approaches to diagnosis by combining the criteria of coverage and consistency. More specifically, they propose that a subset of the overall set of case observations be selected for coverage, while the complement of the closure of the observations, with the observations, is required to be consistent with the solution. In this proposal, time is abstracted out, presumably to simplify the computation of the closure of the observations. Depending on the choice of the subset of the observations to be covered, which the proposers describe as the critical step, different notions of coverage and thus explanation plausibility can be formulated, varying from the very restrictive (everything has to be covered, both observations of normality and abnormality) to the very relaxed (nothing needs to be covered). For the choice of the "best" explanation, the proposers suggest the criterion of minimality of abnormality assumptions (smallest number of assumed failures) or minimality by implication (E_i is better than E_j if E_j implies E_i).

4.2. Points of Failure and Strength of Integration

Definition 10. A diagnostic system functions under a *single-point-of-failure* assumption if, at any time, any diagnostic solution includes at most one diagnostic antecedent. This means that focus-abnormalities have a single cause.

Definition 11. A diagnostic system functions under a *multiple-points-of-failure* assumption if, at any time, any diagnostic solution can include more than one, possibly independent diagnostic elements, i.e., more than one independent causal antecedents. Thus, under this assumption, focus-abnormalities can be partitioned between separate, independent causes.

Definition 12. A diagnostic solution S_t is *strongly-integrated* (or *coherent*) if it is not split into disconnected parts.

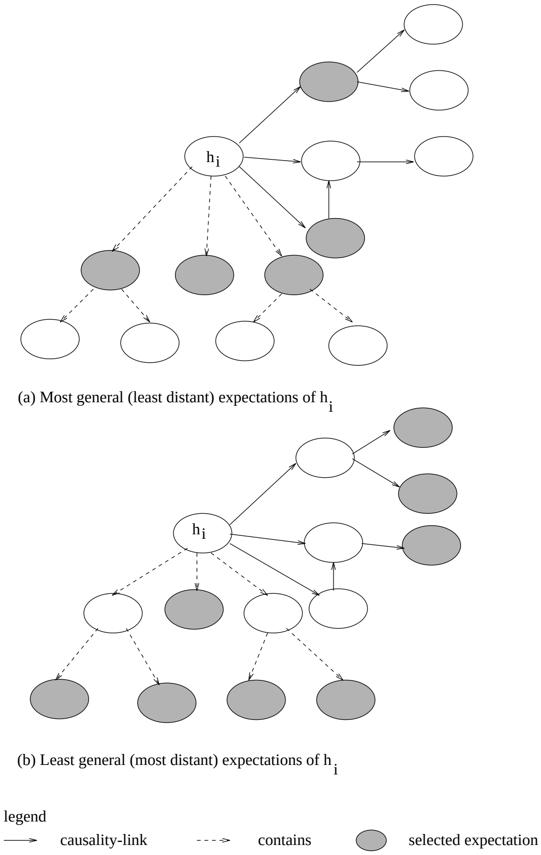
Definition 13. A diagnostic solution is *loosely-integrated* (or *incoherent*) if it consists of disconnected parts.

Axiom 16. single-point-of-failure $(S_t) \Rightarrow$ strongly-integrated(S_t)

Axiom 17. loosely-integrated $(S_t) \Rightarrow$ multiple-points-of-failure(S_t)

4.3. Satisfiability

The coverage of some solution is a measure of how well it accounts for the focus abnormalities. The *satisfiability* of some solution is a measure of how well the expectations of the diagnostic elements are satisfiable by the case history. Let S_t be some (abstracted) diagnostic solution consisting of a number of diagnostic (hypothesis) elements h_i. Function *refined*, applied to S_t, gives the entire set of time-objects comprising the diagnostic solution, not just the hypothesis elements.



(a) Most general (least distant) expectations of h_i

(b) Least general (most distant) expectations of h_i

FIGURE 14. Mutually independent expectations of hypothesis elements.

Definition 14. The expectations of hypothesis element $h_i \in S_t$ are defined to be

$$\textit{expectations}(h_i) = \{e_j \in (\textit{refined}(S_t) \backslash \{h_i\}) | \textit{accounts-for}(h_i,\ e_j)\}.$$

Thus the expectations of h_i are its components and consequents. Such elements are not necessarily mutually independent. Definition 15 gives the most general (least distant), mutually independent expectations of some hypothesis element (see Figure 14a for an illustration), whereas Definition 16 gives the least general (most distant) expectations (see Figure 14b for an illustration). Which of these definitions (or some other definition) for independent expectations is actually employed in the scope of some evaluation would depend on the particular circumstances. Time is relevant in this choice because the evaluation can only consider past and ongoing (rather than future) expectations. The observability or not of expectations is also of relevance. The most general (least distant) expectations are more likely to be past or ongoing but less likely to be observable, whereas the opposite is true for the least general (most distant) expectations.

Definition 15. The most general (least distant), mutually independent expectations of $h_i \in S_t$, *indep-exps*(h_i), are defined to be

$$\textit{indep-exps}(h_i) = \{e_j \in \textit{expectations}(h_i) | \neg \exists e_k \in (\textit{expectations}(h_i) \backslash \{e_j\}) \\ \text{s.t. } \textit{accounts-for}\ (e_k,\ e_j)\}$$

Definition 16. The least general (most distant), mutually independent expectations of $h_i \in S_t$, *indep-exps*(h_i), are defined to be

$$\begin{aligned} \textit{indep-exps}(h_i) = \{e_j \in \textit{expectations}(h_i) | \neg\exists e_k \in (\textit{expectations}(h_i) \backslash \{e_j\}) \\ \text{s.t. } \textit{accounts-for}\ (e_j,\ e_k)\} \end{aligned}$$

Definition 17. Depending on the choice of definition for *indep-exps*, the mutually independent expectations of some diagnostic solution $S_t = \{h_1, h_2, \ldots, h_n\}$, *Exps*$(S_t)$, are defined to be

$$\textit{Exps}(S_t) = \textit{indep-exps}(h_1) \cup \textit{indep-exps}(h_2) \cup \cdots \cup \textit{indep-exps}(h_n)$$

These expectations can be classified temporally into past, ongoing, or future. Furthermore, they can be classified qualitatively into *typical, necessary, common, occasional*, etc. A *typical* expectation has a "pathognomonic" association with some hypothesis element, i.e., observing the expectation establishes the hypothesis. A *necessary* expectation is one whose refutation denies the particular hypothesis.[9] A *common* expectation of some hypothesis element occurs with most actual occurrences of the given failure; its observation supports the hypothesis, and its refutation may mildly discount the hypothesis. An *occasional* expectation only occurs sometimes; its observation supports the hypothesis, but its refutation does not need to be explained. Based on the temporal and qualitative classifications of the expectations of some solution, different satisfiability criteria can be defined, as illustrated in the following definitions.

Definitions 18 to 20. A diagnostic solution S_t is *N/T/C-satisfiable* iff all its observable, past and ongoing, necessary/typical/common expectations are satisfiable (accounted for) by the case history CH_t (which is treated as a compound time-object):

$$\begin{aligned} \textit{N-satisfiable}(S_t) &\Leftrightarrow \forall e_i \in \textit{Nexps}(\textit{accounts-for}(CH_t,\ e_i) \vee (\textit{negative}(e_i) \\ &\qquad \wedge\ \neg\textit{in-conflict-with}(CH_t,\ e_i))) \\ \textit{T-satisfiable}(S_t) &\Leftrightarrow \exists e_i \in \textit{Texps}(\textit{accounts-for}(CH_t,\ e_i)) \\ \textit{C-satisfiable}(S_t) &\Leftrightarrow \forall e_i \in \textit{Cexps}(\textit{accounts-for}(CH_t,\ e_i) \vee (\textit{negative}(e_i) \\ &\qquad \wedge\ \neg\textit{in-conflict-with}(CH_t,\ e_i))) \end{aligned}$$

$$\begin{aligned} \text{where } \textit{Nexps} &= \textit{observable}(\textit{Exps}(S_t)) \cap \textit{necessary}(\textit{Exps}(S_t)) \cap (\textit{past}(\textit{Exps}(S_t)) \\ &\quad \cup\ \textit{ongoing}(\textit{Exps}(S_t))) \\ \textit{Texps} &= \textit{observable}(\textit{Exps}(S_t)) \cap \textit{typical}(\textit{Exps}(S_t)) \cap (\textit{past}(\textit{Exps}(S_t)) \\ &\quad \cup\ \textit{ongoing}(\textit{Exps}(S_t))) \\ \textit{Cexps} &= \textit{observable}(\textit{Exps}(S_t)) \cap \textit{common}(\textit{Exps}(S_t)) \cap (\textit{past}(\textit{Exps}(S_t)) \\ &\quad \cup\ \textit{ongoing}(\textit{Exps}(S_t))) \end{aligned}$$

[9] Failures with pathognomonic or necessary expectations (faults) can be deduced. Let f be a failure, p a pathognomonic fault of f, and n a necessary fault of f. Thus we have: *holds(p)* $\Rightarrow$ *holds(f)* and $\neg$*holds*$(n) \Rightarrow \neg$*holds*(f).

Negative expectations (those involving normal properties) are satisfiable by default unless they conflict directly with the case history. Typical expectations cannot be negative. If S_t is not *N-satisfiable*, some of its diagnostic elements need to be revoked. Similarly, if S_t is *T-satisfiable*, some of its diagnostic elements need to be concluded, and in fact, if some other potential diagnostic solution (hypothetical world) is in conflict with such conclusions, it would need to be (partially) revoked. Typical expectations are not necessarily common expectations, and hence $C\text{-}satisfiable(S_t)$ does not entail $T\text{-}satisfiable(S_t)$. If S_t is *C-satisfiable*, it has a good chance of being the final solution. Often though, complete *C-satisfiability* is not attained (or even possible to establish or refute due to missing information), and hence what is considered is the degree of *C-satisfiability* strengthened by the number of the occasional expectations that are satisfiable, i.e., $(NC + NO)/(TNC + NO)$, where TNC is the total number of relevant common expectations, NC is the number of the relevant common expectations that are satisfiable, and NO is the number of the relevant occasional expectations that are satisfiable. Moreover, the common expectations can be further classified, thus generating even more measures of satisfiability. In order to compute some satisfiability measure for a pursued diagnostic solution, the diagnostic system would need to ask its user various questions concerning unknown expectations (this is a deductive aspect of the system). In the ideal situation, the concluded solution is completely satisfiable.

Definition 21. A diagnostic solution S_t is *completely satisfiable* iff all its relevant necessary expectations are satisfiable and either all its relevant common expectations are satisfiable or any of its relevant typical expectations are satisfiable:

$$satisfiable(S_t) \Leftrightarrow (N\text{-}satisfiable(S_t) \wedge (C\text{-}satisfiable(S_t) \vee T\text{-}satisfiable(S_t)))$$

A knowledge conflict arises if S_t is not *N-satisfiable* but is *T-satisfiable*.

4.4. Ambiguity, Redundancy, and Minimality

Definition 22. The explanation of finding $f_i \in F_t$ at time t, with respect to some diagnostic solution S_t, is defined to be

$$explanation_t(f_i) = \{h_j \in S_t \mid accounts\text{-}for(h_j, f_i)\}.$$

If $explanation_t(f_i) = \{\,\}$, no explanation is on offer by S_t for abnormality f_i at time t. If, on the other hand, $|explanation_t(f_i)| > 1$, two or more explanations are on offer; S_t's explanation for f_i is *ambiguous*.

Definition 23. A diagnostic solution S_t is *ambiguous* iff it contains ambiguous explanations for focus-abnormalities; otherwise, it is *crisp*:

$$ambiguous(S_t) \Leftrightarrow f_i \in F_t \text{ s.t. } |explanation_t(f_i)| > 1$$

$$crisp(S_t) \Leftrightarrow \neg ambiguous(S_t).$$

The ideal requirement is for crisp solutions.

Definition 24. The *ambiguity* of a diagnostic solution S_t is defined as

$$ambiguity(S_t) = \{\langle f, exp\rangle \in exp\text{-}power(S_t) \| exp| > 1\}.$$

where $exp\text{-}power(S_t) = \{\langle f, explanation_t(f)\rangle \mid f \in F_t\}$. An alternative definition for a crisp solution is

$$crisp(S_t) \Leftrightarrow (ambiguity(S_t) = \{\,\}).$$

Thus an empty solution is a crisp solution.

Definition 25. A diagnostic solution S_t is *redundant* iff a strict subset of it has the same coverage:

$$\textit{redundant}(S_t) \Leftrightarrow \exists S_t' \subset S_t \text{ s.t. } \textit{coverage}(S_t') = \textit{coverage}(S_t).$$

where *coverage* $(S_t) = \{f \mid \langle f,\ exp\rangle \in \textit{exp-power}(S_t) \wedge |exp| \neq 0\}$. The coverage of a diagnostic solution is the subset of focus-abnormalities explained by it (the quality of coverage depends on whether it is mostly hard or soft abnormalities that are covered). Definitions 2 to 4 can be formalized alternatively in terms of coverage:

$$\begin{aligned} \textit{focus-coverage}(S_t) &\Leftrightarrow (F_t = \textit{coverage}(S_t)) \\ \textit{hard-coverage}(S_t) &\Leftrightarrow (\textit{hard}_t \subseteq \textit{coverage}(S_t)) \\ \textit{current-coverage}(S_t) &\Leftrightarrow ((\textit{hard}_t \cap \textit{current}_t) \subseteq \textit{coverage}(S_t)). \end{aligned}$$

A single-point-of failure solution cannot be redundant.

Let Λ_t be the subset of S_t each element of which constitutes the sole explanation for some element of F_t, and let Ω_t be the subset of S_t such that for each element of Ω_t there is some element of F_t for which it constitutes an alternative explanation:

$$\begin{aligned} \Lambda_t &= \cup\, exp_i \ s.t.\ \langle f_i,\ exp_i\rangle \in \ \textit{exp-power}(S_t) \wedge |exp_i| = 1 \\ \Omega_t &= \cup\, exp_i \ s.t.\ \langle f_i,\ exp_i\rangle \in \ \textit{exp-power}(S_t) \wedge |exp_i| > 1. \end{aligned}$$

If $\Omega_t = \{\}$, S_t is crisp; otherwise, it is ambiguous (alternative formalization for Definition 23).

Theorem 1. An ambiguous diagnostic solution S_t is nonredundant if $\Omega_t \subseteq \Lambda_t$.

Proof. If $\Omega_t \subseteq \Lambda_t$, then every element of Ω_t is, in addition to being an alternative explanation for some focus-abnormality, the sole explanation for some other focus-abnormality. Thus, by removing any element of Ω_t, the coverage of S_t is reduced. Hence Ω_t is necessary for maintaining the given coverage. □

Theorem 2. An ambiguous diagnostic solution S_t is redundant if $\Omega_t \not\subseteq \Lambda_t$.

Proof. The nonempty set $(\Omega_t - \Lambda_t)$ gives elements of S_t none of which constitutes the sole explanation for some focus-abnormality. Let $\Omega_t' \subseteq (\Omega_t - \Lambda_t)$ be defined as

$$\Omega_t' = \{\omega \in (\Omega_t - \Lambda_t) \mid \forall f_i \in F_t \text{ s.t. } \textit{accounts-for}(\omega,\ f_i)(\exists\, \lambda \in \Lambda_t \text{ s.t. } \textit{accounts-for}(\lambda,\ f_i))\}.$$

A procedural derivation for Ω_t' is given below. $\Omega_t' \neq \{\}$ (see proof below). Thus, by removing Ω_t' from S_t, its coverage is not affected:

$$\textit{coverage}(S_t - \Omega_t') = \textit{coverage}(S_t)$$

Procedural Derivation for Ω'_t.

```
let $\Lambda'_t \leftarrow \Lambda_t$ and $\Omega'_t \leftarrow \{\}$
repeat for every $f_i \in F_t$
   if $explanation_t(f_i) \subseteq (\Omega_t - \Lambda'_t)$
   then let $\omega$ be some element of $explanation_t(f_i)$
      $\Lambda'_t \leftarrow (\Lambda'_t \cup \{\omega\})$
   end if
   $\Omega'_t \leftarrow (\Omega'_t \cup (explanation_t(f_i) - \Lambda'_t))$
end repeat;
```

Proof that $\Omega'_t \neq \{\}$. By definition, $(\Omega_t - \Lambda_t) \neq \{\}$. The base case is that $(\Omega_t - \Lambda_t)$ consists of a single element, say, ω. Since ω is not the sole explanation for any of the focus abnormalities, it means that every focus abnormality has an element of Λ_t as its (alternative) explanation, and thus $\Omega'_t = \{\omega\}$. The general case is that $(\Omega_t - \Lambda_t)$ consists of more than one element, say, two elements, ω_1 and ω_2. In this case it is possible that there is some focus abnormality f_i such that $explanation_t(f_i) = \{\omega_1, \omega_2\}$. Thus either ω_1 or ω_2 can become a member of Ω'_t, say, ω_1, and f_i would still have an explanation, ω_2. If ω_1 constitutes an explanation for another focus-abnormality, say, f_j, ω_1 cannot be the sole explanation for f_j; thus f_j still has an explanation that also could be ω_2.

Theorem 3. A crisp diagnostic solution S_t is nonredundant.

Proof. Since $\Omega_t = \{\}$, every element of S_t is necessary to maintain the given coverage. □

Theorem 4. A redundant diagnostic solution S_t is ambiguous.

Lemma. For a redundant solution, $\Omega_t \neq \{\}$ and $\Omega_t \not\subset \Lambda_t$.

Proof of Lemma. By Definition 25, since at least one element of S_t is not necessary and thus cannot constitute the sole explanation of some focus abnormality.

Proof of Theorem 4. Through the lemma, $\Omega_t \neq \{\}$, and thus S_t is ambiguous.

Definition 26. A diagnostic solution S_t is *minimal* iff, at the given time, there is no other solution with lower cardinality and at least the same coverage:

$$minimal(S_t) \Leftrightarrow \neg(\exists Q_t \text{ s.t. } |Q_t| < |S_t| \wedge coverage(Q_t) \supseteq coverage(S_t)).$$

Axiom 18. $redundant(S_t) \Rightarrow \neg minimal(S_t)$

Axiom 19. $minimal(S_t) \Rightarrow \neg redundant(S_t)$ (contrapositive of Axiom 18).

Definition 27. The concluded diagnostic solution $S_{i,G}$ is *optimal* iff it has focus-coverage and it is case-consistent, satisfiable, strongly-integrated, and minimal:

$$optimal(S_{i,G}) \Leftrightarrow (focus\text{-}coverage(S_{i,G}) \wedge case\text{-}consistent(S_{i,G}) \wedge satisfiable(S_{i,G}) \wedge strongly\text{-}integrated(S_{i,G}) \wedge minimal(S_{i,G})).$$

4.5. A Word on the Closed-World Assumption

In this section we discuss the viability of the closed-world assumption (CWA) with respect to dynamic diagnostic problems. First, we explain the application of the CWA to a case history or a hypothetical world. (It is assumed that relevant temporal data abstractions are applied to the case history so that maximal persistences of the time-objects comprising the history are derived.)

Let P be the universe of discourse of properties (sentences) under the particular application context, and let $N \subset P$ be the subset of negative properties (those asserting normal situations). To each $n_i \in N$ there corresponds a set of positive properties (asserting abnormal situations) $ps_i \subset P$, where $\forall p_{i,j} \in ps_i$ $\{\texttt{excludes}\ (n_i, p_{i,j})\}$; n_i is mutually exclusive with respect to every element of ps_i. All clusters of positive properties ps_i are disjoint, and the elements of N are mutually independent.

Let α be the concrete time-axis that defines the span of valid time of relevance to the diagnostic activity, and let $Times(\alpha) = \{t_1, t_2, \ldots, t_n\}$. We assume that $Times(\alpha)$ is a finite set.

Applying the CWA to the case history and hypothetical worlds. Let CH_t be the believed history of the case at real time t. The CWA essentially dictates the addition of various *normality assumptions* to the case history. The algorithm for this is given below:

Algorithm for applying the CWA to the case history at time t.

```
repeat for each $t_i \in Times(\alpha)$
   repeat for each $n_j \in N$
      let $\tau_j$ be a time-object such that $\pi(\tau_j) = n_j$ and $\varepsilon(\tau_j, \alpha) = \langle t_i, t_i, \texttt{closed}\rangle$
      /* $\tau_j$ has property $n_j$ and exists as a point-object on $\alpha$*/
      if $\neg accounts\text{-}for(CH_t, \tau_j) \wedge\neg\ in\text{-}conflict\text{-}with(CH_t, \tau_j)$
      then add $\tau_j$ to $CH_t$ as an assumed time-object (normality assumption)
   end repeat
   if $n_j$ is a concatenable property
   then apply a merge operation on the newly added time-objects
      to derive maximal persistence
end repeat
```

The normality assumptions are, of course, revocable, since the case history is dynamic (the beliefs about the particular case can change). Hence, at a subsequent point in time, $t+$, say, these assumptions are automatically revoked and recomputed on the basis of CH_{t+}. Thus the relevant normality assumptions can be added and deleted many times during the diagnostic activity. The same algorithm just given applies to a hypothetical world $S_{i,t}$. Again, since a hypothetical world is dynamic, the normality assumptions need to be continuously revoked and recomputed.

It therefore appears that the viability of the CWA in dynamic situations depends on whether the benefits (of more accurate evaluation of potential solutions) accruing from its use outweigh its computational overheads. The only evaluation measure that is affected by the CWA is the consistency measure because the inclusion of the normality assumptions is likely to increase the number of inconsistencies with a potential diagnostic solution. The coverage measures are not affected because these consider only abnormality observations. Similarly, the satisfiability measures are not affected because the normality assumptions included in a hypothetical world do not participate in the relevant sets of expectations—an expectation that is satisfiable without the application of the CWA is still satisfiable when the CWA is applied and an expectation that was not satisfiable (either false or unknown) is still unsatisfiable (although an unknown expectation can become a refuted expectation under the CWA). Thus the CWA appears to be viable only in situations where all the (abnormality) information about a case is known and given prior to the start of the diagnostic activity and nothing changes during the progress of the diagnostic activity; the beliefs about the case remain constant during the period $[O, G]$, i.e., $CH_O = \cdots = CH_t = \cdots = CH_G$. Under such a scenario, the hypothetical worlds can be completely computed on the basis of the *a priori* information (there is no need for any dynamic interactions with the external world—the user of the system—for the acquisition of new information that possibly revokes current beliefs). For many real-life diagnostic problems, however, the case information is initially incomplete; new observations, referring to the past or the present, are continuously acquired during the diagnostic activity. Further, human diagnosticians do not appear to employ the CWA. These characteristics tell against the CWA for dynamic diagnostic problems.

5. EVALUATION CRITERIA IN THE SDD SYSTEM

In this section, for illustration purposes, we briefly discuss the primitive evaluation criteria used in the SDD system and how they are combined to give the specific notions of plausible and best explanation. The algorithmic details of the overall diagnostic logic of this system are outside the scope of this article, but the interested reader is referred to Keravnou et al. (1994). SDD is a system that helps general radiologists, who are not expert in the domain of skeletal dysplasias and malformation syndromes, achieve the diagnostic performance of domain experts. SDD recently underwent a second phase of clinical trials whose results are quite favorable for the system; more specifically, it has been shown that general radiologists can perform better when using SDD than when using the standard method of diagnosis (through textbooks) (Washbrook et al. 1997).

SDD diagnoses under the single-disorder assumption, as has been discussed in Section 3. Most skeletal dysplasias and malformation syndromes are infinitely persistent with fixed, sometimes relatively narrow initiation margins. However, their expectations can be finitely persistent, recurring, etc., and for most of the finitely persistent expectations, margins for their expected durations are known. Apart from their temporal classification (infinitely or finitely persistent, etc.), expectations are qualitatively classified into typical, necessary, common, and occasional. In addition, various subsets of the common expectations are singled out as especially significant and named *sufficient groups*. Disorder models are abduced via the mechanisms of primary and secondary triggers, and secondary triggers play an important role as differentiators of potential solutions. A patient history consists of temporal information (clinical, radiologic, biochemical, etc. findings on the patient) that is classified into hard abnormalities, other abnormalities, and contextual information.

Since the system operates under the single-disorder assumption, any potential solution, at any time, consists of a single-disorder instantiation, i.e., a single diagnostic element. Thus each potential solution is at all times strongly integrated (coherent), crisp, nonredundant, and minimal. The primitive evaluation criteria that are used are hard-coverage, focus-coverage, T-satisfiable, N-satisfiable, C-satisfiable (augmented with established occasional expectations), and S-satisfiable; the latter criterion concerns the satisfiability of sufficient groups. Thus there is no need for data consistency criteria.

Next we outline the combinations of evaluation criteria that are used in different reasoning contexts. The term *differential* is used to denote the subset of hypothetical worlds (potential solutions) that are being actively considered. The initial differential consists of a subset of the disorder instantiations triggered on the basis of the initial contents of the patient history (CH_O). More specifically, the initially triggered disorders are evaluated from the perspective of focus-coverage and S-satisfiability. Owing to data and knowledge incompleteness, it is rare for a triggered disorder to attain complete focus-coverage and S-satisfiability, and hence the *degree* of attainment is computed. For each evaluation criterion, the topmost disorder and those close to it are selected. Subsequent differentials are drawn from the subset of most promising disorder instantiations where the promise is evaluated on the basis of hard-coverage. The elements of these differentials are again evaluated on the basis of the combined criteria focus-coverage and S-satisfiable, but in this context S-satisfiability is used as the primary criterion and focus-coverage as the secondary criterion. Throughout the diagnostic activity (i.e., initially and every time new information on the patient is acquired), all instantiated disorders, whether or not they belong to the current differential, are evaluated on the basis of criteria T-satisfiable and N-satisfiable. T-satisfiability concludes the particular disorder instantiation, whereas the negation of N-satisfiability revokes it. Finally, in the context of summarizing and presenting to the user the system decisions, criteria focus-coverage and C-satisfiable are used.

The notions of plausible and best explanation employed in any diagnostic system are quite critical to its overall performance. Our experience in developing SDD shows that eliciting such notions from the domain experts is difficult and that the notion of best explanation is much harder to formalize than that of explanation plausibility. This elicitation task for a new diagnostic domain can be aided substantially if the various primitive, general evaluation criteria can be used as the starting point. In SDD it was not possible to formalize best explanation in terms of a simple, context-free formula. Instead, we have arrived at a set of rules giving, in a declarative way, the various context-sensitive interpretations of this notion. Such rules are very transparent and can be modified easily; they are given next.

A potential diagnosis constitutes a plausible explanation if it attains a focus-coverage of at least 50 percent or it is T-satisfiable. This is the minimum requirement.

The rules that define best explanation express further requirements, over and above the minimum requirement. They are prioritized and applied in descending order of priority as listed below:

1. The plausible diagnosis is T-satisfiable.
2. The plausible diagnosis is the only one; it attains a focus-coverage of at least 60 percent and an S-satisfiability of at least 60 percent.
3. The plausible diagnosis attains a focus-coverage and an S-satisfiability of at least 60 percent, and both these measures are the highest among the plausible diagnoses.
4. The plausible diagnosis attains an S-satisfiability of at least 60 percent that is the highest among the plausible ones, it attains a focus-coverage of at least 60 percent (not the highest among the plausible ones), and it is significantly better than the plausible diagnosis with the next highest S-satisfiability; where by *significantly better* we mean that either its S-satisfiability or its focus-coverage is at least 20 points higher.
5. The plausible diagnosis attains either an S-satisfiability of at least 70 percent or a focus-coverage of at least 50 percent, and a close opponent suggests it. An opponent *suggests* a hypothesis when it has it as a secondary trigger. An opponent is *close*

when it has either an S-satisfiability or a focus-coverage within at least 90 percent of the corresponding measure. (Clusters of dysplasias that are mutual opponents are predefined and related through secondary triggers.)

The cutoff percentages that appear in the preceding rules may seem ad hoc, and to a certain extent they are. They attempt to quantify qualitative expressions from the experts regarding the differentiation of plausible diagnoses and have resulted from a number of elicitation and testing sessions of the system. Their values are of secondary importance to yielding the correct differentiations and hence selections.

The first rule that applies determines the best explanation, which becomes the concluded diagnosis. If two or more plausible diagnoses are considered best explanations, it might be that the particular patient has a multidysplasia problem. If no rule applies to any of the plausible diagnoses, no diagnostic conclusion can (yet) be reached (and this may be the "diagnosis" that an expert would conclude). A concluded diagnosis is truly the best explanation only if it is shown that it is the correct diagnosis for the particular patient.

6. CONCLUSIONS

The mechanization of diagnostic reasoning as a special case of abductive inference has received considerable attention within the AI community (Hamscher et al. 1992; Struss 1992) and continues to provide fertile ground for research. In this article we have focused on two aspects of (abductive) diagnostic reasoning that we believe are not adequately addressed at present. These are *time* and *evaluation*.

Time is intrinsically relevant in many diagnostic problems, and as such, temporal reasoning plays a central role in the formation and evaluation of potential diagnostic solutions. Time therefore should be an integral aspect of the knowledge and reasoning of diagnostic systems for these domains. This integration can be achieved by treating time as an integral aspect of the entities that constitute the processing elements of the systems. The notion of a time-object captures this requirement, and we have shown how models of failures (case histories) and normal processes can be modeled in terms of time-objects.

The essence of any abductive diagnostic system is the generation of the best explanation of some observations suggesting abnormal functioning. The formation and evaluation of potential explanations are tightly coupled processes. We advocate that primitive evaluation criteria should be separately and explicitly represented to allow their flexible and transparent combined use in different reasoning contexts and have presented a number of such criteria. This allows different notions of plausible explanation (minimum requirements for accepting a potential explanation as plausible) and best explanation (selection requirements from among plausible competitors) to be clearly formulated.

Early abductive diagnostic systems concealed their evaluation criteria through opaque scoring functions. In more recent approaches, the widely adopted definition of plausible explanation as full coverage of (abnormal) observations is restrictive in a pragmatic sense. We believe that the evaluation aspects of abductive diagnostic reasoning warrant further investigation that doubtless will reveal aspects of the formation of potential explanations.

The ongoing work in the domain of skeletal dysplasias and malformation syndromes presented in this article has been our main source and test-bed of the proposed ideas, and the practical results obtained so far (through the SDD system) are very encouraging.

ACKNOWLEDGMENTS

The research reported here has been partly supported by the University of Cyprus, the Leverhulme Trust (UK) (under projects SDD and MSD), and the CEC (under AIM Project A2034 GAMES II). We also would like to thank the other members of the SDD team, Christine Hall, Felicity Dams, and Don Shaw, for the numerous, thoroughly enjoyable, and insightful research discussions, as well as the three anonymous referees for their very useful comments on earlier drafts of this article.

REFERENCES

ALLEN, J. F. 1984. Towards a general theory of action and time. Artificial Intelligence, **23**:123–154.

BYLANDER, T., D. ALLEMANG, M. C. TANNER, and J. R. JOSEPHSON. 1991. The computational complexity of abduction. Artificial Intelligence, **49**:25–60.

CHANDRASEKARAN, B. and S. MITTAL. 1983. Conceptual representation of medical knowledge for diagnosis by computer: MDX and related systems. Advances in Computers, **22**:217–293.

CONSOLE, L., L. PORTINALE, D. THESEIDER DUPRE, and P. TORASSO. 1992. Diagnostic reasoning across different time points. *In* Proceedings ECAI-92, pp. 369–373.

CONSOLE, L., and P. TORASSO. 1990a. Integrating models of the correct behavior into abductive diagnosis. *In* Proceedings ECAI-90, pp. 160–166.

CONSOLE, L., and P. TORASSO. 1990b. Hypothetical reasoning in causal models. International Journal of Intelligent Systems, **5**(1):83–124.

CONSOLE, L., and P. TORASSO. 1991a. On the co-operation between abductive and temporal reasoning in medical diagnosis. Artificial Intelligence in Medicine, **3**:291–311.

CONSOLE, L., and P. TORASSO. 1991b. A spectrum of logical definitions of model-based diagnosis. Computational Intelligence, **7**:133–141.

DOJAT, M., and C. SAYETTAT. 1994. Aggregation and forgetting: Two key mechanisms for across-time reasoning in patient monitoring. *In* Proceedings AAAI Spring Symposium, pp. 33–36. AAAI Technical Report SS-94-01, Stanford University.

FRIEDRICH, G. 1993. Model-based diagnosis and repair. AICOM, **6**:187–206.

FRIEDRICH, G., and F. LACKINGER. 1991. Diagnosing temporal misbehaviour. *In* Proceedings IJCAI'91, pp. 1116–1122.

HAIMOWITZ, I. J., and I. S. KOHANE. 1996. Managing temporal worlds for medical trend diagnosis. Artificial Intelligence in Medicine, **8**(3):299–321.

HAMSCHER, W., L. CONSOLE, and J. DE KLEER (eds.). 1992. Readings in Model-Based Diagnosis. Morgan Kaufmann Publishers, San Mateo, Calif.

VAN HARMELEN, F., and A. TEN TEIJE. 1994. Using domain knowledge to select solutions in abductive diagnosis. *In* Proceedings ECAI-94, pp. 652–656.

KAHN, G. M., L. M. FAGAN, and L. B. SHEINER. 1991. Combining physiologic models and symbolic methods to interpret time-varying patient data. Methods of Information in Medicine, **30**:167–178.

KERAVNOU, E. T. 1995a. Modelling medical concepts as time-objects. *In* Proceedings AIME-95, Lecture Notes in Artificial Intelligence, Vol. 935, pp. 67–78. Springer, Berlin.

KERAVNOU, E. T. 1995b. Temporal vagueness in medical reasoning. International Journal of Systems Research and Information Science, **7**:3-28.

KERAVNOU, E. T. 1996a. Engineering time in medical knowledge-based systems through time-axes and time-objects. *In* Proceedings TIME-96, pp. 160–167. IEEE Computer Society Press, New York.

KERAVNOU, E. T. 1996b. An ontology of time using time-axes and time-objects as primitives. Technical Report TR-96-9, Department of Computer Science, University of Cyprus.

KERAVNOU, E. T. (ed.). 1996c. Temporal reasoning in medicine (editorial). Artificial Intelligence in Medicine, **8**(3):187–191.

KERAVNOU, E. T. 1996d. Temporal diagnostic reasoning based on time-objects. Artificial Intelligence in Medicine, **8**(3):235–265.

KERAVNOU, E. T. 1997. Temporal abstraction of medical data: deriving periodicity. *In*, Intelligent Data Analysis in Medicine and Pharmacology, pp. 61–79. *Edited by* N. Lavrac, E. T. Keravnou, and B. Zupan. Kluwer Academic Publishers, Boston.

KERAVNOU, E. T. 1998. A time ontology for medical knowledge-based systems. *In* Proceedings EMCSR '98, pp. 831–835.

KERAVNOU, E. T. 1999. A multidimensional and multigranular model of time for medical knowledge-based systems. Journal of Intelligent Information Systems, **13**:73–120.

KERAVNOU, E. T., and J. WASHBROOK. 1990. A temporal reasoning framework used in the diagnosis of skeletal dysplasias. Artificial Intelligence in Medicine, **2**:239–265.

KERAVNOU, E. T., F. DAMS, J. WASHBROOK, C. M. HALL, R. M. DAWOOD, and D. SHAW. 1992. Background knowledge in diagnosis. Artificial Intelligence in Medicine, **4**:263–279.

KERAVNOU, E. T., F. DAMS, J. WASHBROOK, C. M. HALL, R. M. DAWOOD, and D. SHAW. 1994. Modelling diagnostic skills in the domain of skeletal dysplasias. Computer Methods and Programs in Biomedicine, **45**:239–260.

LARIZZA, C., R. BELLAZZI, and A. RIVA. 1997. Temporal abstractions for diabetic patients management. *In* Proceedings AIME-97. Lecture Notes in Artificial Intelligence, Vol. 1211, pp. 319–330. Springer, Berlin.

LAVRAC, N., E. T. KERAVNOU, and B. ZUPAN (eds.). 1997. Intelligent Data Analysis in Medicine and Pharmacology. Kluwer Academic Publishers, Boston.

LONG, W. 1996. Temporal reasoning for diagnosis in a causal probabilistic knowledge base. Artificial Intelligence in Medicine, **8**(3):193–215.

MAIOCCHI, R., and B. PERNICI. 1991. Temporal data management: a comparative view. IEEE Transactions on Knowledge and Data Engineering, **3**:504–523.

MIKSCH, S., W. HORN, C. POPOW, and F. PAKY. 1996. Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. Artificial Intelligence in Medicine, **8**:543–576.

MILLER, R. A., 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.

MYLOPOULOS, J., M. BORGIDA, M. JARKE, and M. KOUBARAKIS. 1990. Telos: A language for representing knowledge about information systems (revised). Technical Report KRR-TR-89-1, Department of Computer Science, University of Toronto.

NEJDL, W., and J. GAMPER. 1994. Harnessing the power of temporal abstraction in model-based diagnosis of dynamic systems. *In* Proceedings ECAI-94, pp. 667–671.

PATIL, R. S. 1981. Causal representation of patient illness for electrolyte and acid-based diagnosis. MIT Laboratory for Computer Science, Technical Memo MIT/LCS/TR-267.

PAUKER, S. G., G. A. GORRY, J. P. KASSIRER, and W. B. SCHWARTZ. 1976. Toward the simulation of clinical cognition: taking a present illness by computer. American Journal of Medicine, **60**: 981–995.

PENG, Y., and J. A. REGGIA, 1990. Abductive Inference Models for Diagnostic Problem-Solving. Springer-Verlag, Berlin.

POOLE, D. 1988. A logical framework for default reasoning. Artificial Intelligence, **36**:27–47.

POOLE, D. 1989a. Explanation and prediction: An architecture for default and abductive reasoning. Computational Intelligence, **5**(2):97–110.

POOLE, D. 1989b. Normality and faults in logic-based diagnosis. *In* Proceedings IJCAI-89, pp. 1304–1310.

POOLE, D. 1990. A methodology for using a default and abductive reasoning system. International Journal of Intelligent Systems, **5**(5):521–548.

POOLE, D. 1994. Representing diagnosis knowledge. Annals of Mathematics and Artificial Intelligence, **11**:33–50.

POOLE, D., R. GOEBEL, and R. ALELIUNAS. 1987. THEORIST: A logical reasoning system for defaults and diagnosis. *In* The Knowledge Frontier, pp. 331–352. *Edited by* N. Cercone and G. McCalla. Springer-Verlag, Berlin.

POPLE, H. E. 1973. On the mechanization of abductive logic. *In* Proceedings IJCAI-73, pp. 147–152.

Pople, H. E. 1977. The formation of composite hypotheses in diagnostic problem solving: An exercise in synthetic reasoning. *In* Proceedings IJCAI-97, pp. 144–147.

Pople, H. E. 1982. Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. *In* Artificial Intelligence in Medicine, *Edited by* P. Szolovits. pp. 119–190. Westview Press, Boulder, Colo.

Russ, T. A. 1995. Use of data abstraction methods to simplify monitoring. Artificial Intelligence in Medicine, **7**:497–514.

Sadegh-Zadeh, K. 1994. Fundamentals of clinical methodology: I. Differential indication. Artificial Intelligence in Medicine, **6**:83–102.

Shahar, Y. 1994. A knowledge-based method for temporal abstraction of clinical data. Ph.D. thesis, Department of Computer Science, Stanford University (Report No. KSL-94-64 or STAN-CS-TR-94-1529).

Shahar, Y., and M. A. Musen. 1996. Knowledge-based temporal abstraction in clinical domains. Artificial Intelligence in Medicine, **8**(3):267–298.

Shahar, Y., S. W. Tu, A. K. Das, and M. A. Musen. 1992. A problem-solving architecture for managing temporal data and their abstractions. *In* Proceedings Workshop on Implementing Temporal Reasoning (AAAI-92).

Shoham, Y. 1987. Temporal logics in AI: Semantical and ontological considerations. Artificial Intelligence, **33**:89–104.

Struss, P. 1992. Knowledge-based diagnosis: An important challenge and touchstone for AI. *In* Proc. ECAI-92, pp. 863–874.

Thagard, P. 1978. The best explanation: Criteria for theory choice. Journal of Philosophy, **75**:76–92.

Thagard, P. 1991a. The dinosaur debate: Explanatory coherence and the problem of competing hypotheses. *In* Philosophy and AI: Essays at the Interface, pp. 279–300. *Edited by* J. Pollock and R. Cummins. MIT Press/Bradford Books, Cambridge, Mass.

Thagard, P. 1991b. Philosophical and computational models of explanation. Philosophical Studies, **64**:87–104.

Thagard, P. 1991c. Review of Y. Peng and J. Reggia, “Abductive Inference Models for Diagnostic Problem Solving.” SIGART Bulletin, **2**(1):72–75.

Thagard, P. 1992. Hypothesis formation. *In* Advances in the Psychology of Thinking, Vol. 1, pp. 177–201. *Edited by* M. Keane, and K. Gilhooly. Harvester Wheatsheaf, Hemel Hempstead, UK.

Washbrook, J., F. S. Dams, C. M. Hall, D. Shaw, and E. T. Keravnou. 1997. Malformation syndromes diagnostician project: Final report to the Leverhulme Trustees (available on request).