Extending Model-Based Diagnosis to Medicine

Keith L. Downing† and Lawrence E. Widman‡
†Department of Computer and Information Science, Linköping University,
S-58183 Linköping, Sweden
downing@ida.liu.se
and
‡ Research Service, Albuquerque Veterans Administration Medical Center
and Division of Cardiology, University of New Mexico School of Medicine
Albuquerque, New Mexico
widman@sumex-aim.stanford.edu

April 15, 1991

1 Introduction

This research attempts to extend the conventional model-based diagnosis (MBD) paradigm [4, 13] to the domain of cardiovascular physiology, where feedback mechanisms of varied timescales are prevalent. Most MBD systems have been developed within the domain of digital electronics, where typical device models lack feedback and have well-defined inputs and outputs. However, the basic MBD paradigm, as developed for GDE [4], places no such restrictions upon a diagnosible model. GDE only requires an ability to propagate information throughout the model in order to derive contradictions between predicted and observed variable readings. Feedback loops do not interfere with this process as long as the feedback model represents only the steady-state relations between variables.

Steady-state models of physiological systems are not hard to find, however, physiological systems are rarely in steady state (even at fairly coarse observational granularities) due to the plethora of homeostatic mechanisms, which frequently alter some variables in order to stabilize others. This complicates MBD, because contradictions during constraint propagation can ambiguously indicate either faults or temporary instabilities. To ameliorate this problem, we propose two extensions to the GDE methodology. First, observations must take place at different time slices. Second, the conflicts and candidates generated at each slice must be analyzed in light of information about the general time-varying behavior of the system's component mechanisms.

2 AI in Medicine versus Model-based Diagnosis

Most AI research in automated diagnosis lies within one of two fields: AI in Medicine (AIM) or Model-based Diagnosis (MBD). Contemporary work in both areas uses "first-principle" or "deep-model" information as the basis for diagnostic reasoning. However, the stereotypical domains of each field: physiology and electronics, respectively, admit different types of models.

Electrical devices consist of well-defined components having local properties that are accurately described by physical laws. Standard MBD techniques propagate information through these component models, compare predictions to observations, and close in on the faulty component, i.e., one whose local-variable values violate a physical constraint. The constraint-based reasoning indigenous to MBD supports

*This research was supported in part by the Swedish Information Technology Program and by grants from the Research Service of the Department of Veteran Affairs and from the American Heart Association–Texas Affiliate.

dependency tracking, which in turn paves the way for assumption-maintenance machinery such as the ATMS [5]. De Kleer and Williams [4] capitalized on the powers of constraints and the ATMS to design The General Diagnostic Engine (GDE), which provides a domain-independent means of (a) diagnosing multiple faults, and (b) formally determining 'best' measurements.

Local components with well-defined behaviors are also abundant in physiology, but they predominantly occur at a microscopic granularity. The coarse-grained components of physiological systems are far more complex than those of electronics. Even a simple blood vessel is far removed from the idealized conduits of electronics and hydraulics. When blood flow increases in a vessel, its walls expand and decrease resistance, but if flow rises considerably, regulatory mechanisms cause resistance increases. A physiological system consists of many such components plus the local and global regulatory mechanisms that monitor them. This makes deep-model development significantly more difficult in physiology than in electronics.

AIM researchers have combatted this complexity by exploiting empirical associations between variables as the basis for deep models [12, 14]. Although these models often lack (a) a solid foundation in physical laws, and (b) a clear mapping between modular physical structures and behaviors, they nonetheless illustrate first-principle relationships between physiological mechanisms and behaviors.

Patil [11] argues that MBD offers little hope for medical diagnosis, because current medical knowledge provides few causal pathways from anatomical structures (i.e., physical components) to physiological behaviors. However, 'component' need not imply a physical object, just as 'structure' in qualitative physics can mean anything from 'physical topology' to 'differential equation'. Instead, components may be abstract regions or mechanisms, so long as they exhibit modularity. For instance, Ironi et. al. [9] show the utility of compartmental models in medical diagnosis. Compartments have no necessary mapping to physical space, yet they abstract physiological systems to a level of interacting modules. Hence, the extension of MBD to medicine should not be dismissed solely on the grounds given by Patil, because many standard physiological ontologies are modular yet non-anatomical. By representing these modules as constraints and by employing a suitable constraint-propagation technique, we can make physiology amenable to the MBD paradigm.

3 Extending MBD to Physiological MBD

This research uses physiological mechanisms as the prerequisite modules for MBD. A properly working mechanism satisfies constraints among its local and shared variables. After a perturbation, some mechanisms have a non-zero delay period before their constraints become resatisfied, and some may have only a finite duration, after which their constraints become violated again. Variations in the delays and durations of mechanisms account for the multiple time scales of physiological activity.

When faulty, a mechanism will fail to satisfy its constraints even during its 'duration' period. So a PMBD (Physiological MBD) system should detect faulty mechanisms in the same way that GDE finds a broken adder. However, the temporal aspects of mechanisms lead to problems. To wit, if we take observations from a real physiological system, propagate them through the mechanism constraints, and derive a contradiction, then at least one of the mechanisms must either be (a) faulty, or (b) in a delay or post-duration period. Conventional MBD algorithms assume only case (a), but systems with feedback mechanisms, especially of multiple time scales, must admit both possibilities.

As a further complication, a medical diagnostician often does not know the time at which a fault first occurred. This precludes complete knowledge of which regulatory mechanisms should be active at symptom-observation time. But by taking measurements at different times, the diagnostician can gain relative temporal information which can help disambiguate faulty from inactive mechanisms.

In short, the temporal aspects of both physiological mechanisms and the diagnostic process necessitate two important extensions to the MBD paradigm: (1) observations at multiple time slices, and (2) comparative analyses of diagnoses generated in different slices, in light of temporal information associated with the physiological mechanisms. We consider these issues in more detail after first describing Widman's constraint-based modeling technique.

4 Widman's Constraint-based Modeling Technique

Widman [15, 16] uses a modeling technique based on classical systems analysis to capture the functional relationships among objects in a system whose parameters can be described with flexible precision. In this context, an "object" is a measurable or potentially measurable variable in the system. For example, a model of a car might include as objects, or variables, the position of an accelerator pedal and the rotational velocity of the axle. The functional relationships, or mechanisms, that relate the objects to each other are represented as links among the objects. In the case of a car model, for example, the output of the pedal position object might be linked by a linear relationship to the input of the axle velocity object. This relationship would then represent abstractly the mechanism by which the pedal position regulates axle velocity. In a more detailed model, this relationship would be represented by, for example, the mechanisms in the engine by which pedal position affects the axle velocity.

The fundamental notions are: (1) each object is a quantity or set of quantities (see below) whose magnitude(s) are determined by its relationship with each of its inputs. These relationships are the *external* mechanisms by which the object is affected by the inputs. (2) objects fall into characteristic types for a given system being modeled. For example, a fluid-flow system would have storage containers, valves, sources, and sinks. An electrical system would have capacitors, resistors, and inductors, as well as sources and sinks. Each object, or variable, type has characteristic relationships to its inputs. For example, storage containers sum their sources and subtract their sinks, while valves divide their fluid inputs in accordance with their valve position inputs. These characteristic relationships reflect the *internal* mechanisms of the objects being modeled.

Some object types can be complex. For example, a distensible fluid storage container would contain, at a given moment, a certain volume of fluid stored at a certain pressure associated with a certain compliance (the change in pressure resulting from an incremental change in stored volume). In the cardiovascular system, the arteries and veins require representation of yet another aspect, namely the net effect of the inputs that modify the vessel compliance independently of the contained volume. Thus, (3) objects of complex type represent several scalar quantities whose magnitudes are determined by relationships with each other. For example, in the distensible fluid-storage type object, the pressure is determined by the current volume and the current compliance. (4) the relationships between inputs and an object and the relationships within an object can be described as difference equations. And, (5) the parameters in the difference equations may be left unspecified if they are not known or may be estimated with whatever semi-quantitative precision is available. Unknown parameters are assigned default values that generate reasonable qualitative behaviors or envisionments.

The envisionment method for this model is direct numerical simulation. Unlike QSIM [10], this forward reasoning method does not permit multiple behaviors to be inferred from a given model. Rather, each model yields exactly one behavior. Multiple behaviors are obtained by creating a set of models to represent the range of reasonable parameter values. This technique proves useful in modeling complex systems with numerous feedback loops, and many parameters that are difficult to estimate precisely but that can be estimated approximately. Such complex systems have traditionally been difficult to represent and simulate by pure qualitative methods.

Even in a relatively simple heart model from cardiovascular domain [17], there are a dozen individual feedback relationships and several dozen parameters that cannot be estimated in individual patients. Perhaps because the cardiovascular system has evolved to be stable under stress [8], the model that represents it is likewise well-behaved under a variety of conditions [17].

An attractive feature of this semi-quantitative technique is that the explicit mapping onto numerical difference equations of the external and internal relationships, or mechanisms, that determine each object's behavior, permits expectation-driven diagnostic reasoning. For this purpose, each of the difference equations is considered to be a constraint equation that is satisfied at steady-state but not during transient changes or in faulty states. Then, each object's behavior is equivalent to one or more constraint equations that can be used to test for consistency among the inter-object and intra-object mechanisms of the model's objects [15, 16]. As noted above, the ability to detect failure of constraints is fundamental to model-based diagnostic reasoning.

Control systems frequently include feedback loops that operate at characteristic time scales. For example, the loop by which the sympathetic nervous system regulates heart rate operates on the order of

seconds, while the loop by which volume depletion regulates fluid excretion operates on the order of a few hours. Another advantage of mapping object behaviors onto explicit mathematical equations is that these time scales, and the delays they imply, can be included in diagnostic models. Thus, the model representing the assumption that cardiac contractility fell *within the past hour* would support the inference that (i.e., include the control loops predicting that) the heart rate would increase in compensation, but would not support the inference that compensatory fluid retention could be observed. The parallel model for the assumption that cardiac contractility fell several weeks ago, in contrast, would support the second inference. It would exclude the first inference, however, because the short-term mechanism adapts to the new blood pressure within a few days, and thus is no longer active at the second time point. This flexibility of representation follows naturally from the explicit mathematical representation of the behavior of the objects that constitute the system being modeled.

Finally, the object-oriented aspect of the semi-quantitative modeling technique supports hierarchical diagnosis. A limitation of the equations describing object behavior is that they cannot describe the mechanisms underlying them. The symbolic data for each object, in contrast, can point to a submodel that, in abstract, is represented by the object.

For example, the link between blood pressure and the sympathetic nervous system can usefully be thought of as a direct causal relationship. In fact, the relationship is complex: blood pressure is sensed by mechanical stretch receptors in the carotid sinuses, located in the carotid arteries, and the aortic arch. These receptors regulate nervous system inputs to poorly understood areas in the brainstem, which then modify the activity of the sympathetic nervous system. This extended reasoning chain can be represented within the model as a submodel. Such a submodel would be useful in diagnosing, for example, patients with syncope (loss of consciousness). One consideration in these patients is excessive sensitivity of the carotid sinuses, leading to inappropriately large decreases in blood pressure in response to pressure on the carotid arteries. The model can ask about a history of neck surgery or throat cancer when exploring this diagnosis only if the deeper causal chain is available to it. In general, of course, the simpler, more abstract relationship is preferable.

5 The PMBD Diagnostic Scenario

This work addresses situations in which the diagnostic program has data from tests taken at multiple time slices during the course of the ailment. The program freely moves about the time slices gathering data and refining candidates. This "time hopping" is analogous to a doctor's ability to check a patient's history or to further analyze samples or specimens taken from the patient at earlier times. [1] Time hopping will often create situations in which the same variables have not been considered at all of the time points. Our diagnostic procedure accounts for these across-time data-set differences in comparing candidates from distinct time slices.

6 Temporal Enhancements of Model-Based Diagnosis

To handle the temporal aspects of PMBD, we enhance the basic MBD operations of conflict and candidate generation with focusing heuristics that exploit the delays and durations of mechanisms. These heuristics rely heavily on the presence of multiple time points of observation, which lead to conflict and candidate sets associated with each temporal slice.

Let Φ be the set of observation time points, where each point is defined relative to an estimated 'time zero' at which the first fault(s) occurred. [2] Next, $\forall t_i \in \Phi$ let $MV(t_i) = \{V : \mathit{measured\_variable}(V, t_i)\}$ and $PV(t_i) = \{V : \mathit{value\_predicted\_for?}(V, t_i)\}$, where predictions are derived from observations via constraint propagation. Conflicts occur when the sign of variable's normalized predicted value disagrees with the sign of its normalized observed value. Then, $\forall t_i \in \Phi$ there exist $\mathit{Conflicts}(t_i)$ and $\mathit{Candidates}(t_i)$ derived solely from $MV(t_i)$ and the mechanism constraints.

[1] This scenario also occurs during data-intensive types of diagnosis such as core-dump analysis, where the diagnostician can look at time-tagged data points but certainly does not want to look at ALL of the data.

[2] Our best estimate of time zero is the time at which the first symptoms were observed (by the physician) or reported (by the patient).

'Active' assumptions about mechanisms make up the conflicts and candidates, where an 'active' mechanism is both unfaulted and operating within its 'duration' period. Conversely, a 'dormant' mechanism is unfaulted but not within its duration period. We assume that no faults are intermittent, so a faulted mechanism will never satisfy its constraint, while a dormant mechanism will only have violated constraints during delay and post-duration periods. This section presents heuristics for determining whether an inactive mechanism is faulted or dormant.

A conflict consists of a disjunction of $\neg active(M_k)$ assumptions, while a candidate is a conjunction of $\neg active(M_k)$s that logically satisfies each conflict. The 'suspect mechanisms' of a time slice are those involved in conflicts:

$$SM(M_k, t_j) \leftrightarrow \exists C \in Conflicts(t_j) \ni \neg active(M_k) \in C \tag{1}$$

Also define the 'incriminating measurements' of mechanism M_k during a particular time slice as those sets of observed variables that led to the derivation of a conflict that contains M_k:

$$\begin{aligned} IM(M_k, t_j) \equiv \{S : S \subseteq MV(t_j) \wedge \exists C \in Conflicts(t_j) \ni \\ \neg active(M_k) \in C \wedge supporting\_measurements(C) = S\} \end{aligned} \tag{2}$$

As shown by de Kleer and Williams [4], if the constraint propagator makes only deterministic predictions based on a given observation set, then the predictions derived by any superset of those observations will be a superset of the original predictions. An analogous type of monotonicity pertains to observation sets in different time slices. Namely:

$$MV(t_i) \subseteq MV(t_j) \Rightarrow PV(t_i) \subseteq PV(t_j) \tag{3}$$

Unfortunately, the above antecedent entails no strong relationships between $Conflicts(t_i)$ and $Conflicts(t_j)$, since the antecedent says nothing about the *values* of the observed variables. Those values may differ between observation times, thereby leading to completely different conflict sets, regardless of the similarity between $MV(t_i)$ and $MV(t_j)$.

However, the fact that a mechanism M_k appears in a conflict of t_i but not of t_j indicates that M_k is probably not faulty, but that it might have been dormant at t_i.

$$\begin{aligned} \forall i, j, k : [MV(t_i) \subseteq MV(t_j) \wedge SM(M_k, t_i) \wedge \neg SM(M_k, t_j)] \\ \rightsquigarrow dormant(M_k, t_i) \end{aligned} \tag{4}$$

Above, '$\rightsquigarrow$' denotes a heuristic implication. The relationship between the MVs indicates that the constraint propagator had at least as much information at t_j as at t_i, but it could not find a reason to suspect M_k. In fact, a weaker antecedent that employs only M_k's incriminating measurement sets, yields the same confidence in the conclusion:

$$\begin{aligned} \forall i, j, k : [[\forall S \in IM(M_k, t_i) : S \subseteq MV(t_j)] \wedge SM(M_k, t_i) \wedge \neg SM(M_k, t_j)] \\ \rightsquigarrow dormant(M_k, t_i) \end{aligned} \tag{5}$$

The above heuristic becomes stronger with the addition of temporal information about M_k:

$$\begin{aligned} \forall i, j, k : [[\forall S \in IM(M_k, t_i) : S \subseteq MV(t_j)] \wedge SM(M_k, t_i) \wedge \neg SM(M_k, t_j) \wedge \\ t_i \leq delay(M_k) < t_j \leq delay(M_k) + duration(M_k)] \\ \rightsquigarrow dormant(M_k, t_i) \wedge became\_active(M_k, (t_i, t_j)) \end{aligned} \tag{6}$$

where the *became-active* predicate implies that the mechanism went from dormant to active sometime between the two time points.

Conversely, the sudden appearance of a mechanism within a conflict set (that has no more incriminating information than a previous conflict set) may indicate a shift to dormancy rather than a fault:

$$\begin{aligned} \forall i, j, k : [[\forall S \in IM(M_k, t_j) : S \subseteq MV(t_i)] \wedge \neg SM(M_k, t_i) \wedge SM(M_k, t_j) \wedge \\ delay(M_k) < t_i \leq delay(M_k) + duration(M_k) < t_j] \\ \rightsquigarrow became\_inactive(M_k, (t_i, t_j)) \wedge dormant(M_k, t_j) \end{aligned} \tag{7}$$

These hypothesized transitions between dormant and active states provide grounds for pruning conflicts and candidates. Specifically, $became\_active(M_k, (t_i, t_j))$ encourages us to favor candidates containing $\neg active(M_k)$ for all observation times

$$t_b \in \Phi \ni t_b \leq delay(M_k) < t_j. \tag{8}$$

$became\_active(M_k, (t_i, t_j))$ also advocates the removal of $\neg active(M_k)$ from all conflicts at times

$$t_d \in \Phi \ni t_i \leq delay(M_k) < t_d \leq delay(M_k) + duration(M_k). \tag{9}$$

Similarly, $became\_inactive(M_k, (t_i, t_j))$ recommends candidates containing $\neg active(M_k)$ for time slices

$$t_a \in \Phi \ni t_a > delay(M_k) + duration(M_k) \geq t_i. \tag{10}$$

It also supports the removal of $\neg active(M_k)$ from all conflicts at observation times

$$t_d \in \Phi \ni delay(M_k) < t_d \leq delay(M_k) + duration(M_k) < t_j. \tag{11}$$

In sum, the above heuristics integrate information about conflicts, the measurements used to derive them, and the normal temporal behavior of mechanisms to differentiate dormancies from true faults. Although the grounding of these heuristics in standard GDE conflicts supports the diagnosis of multiple faults, the heuristics assume that all faults happened at time zero. Hence, all expected reaction times (i.e., delays) of regulators are measured relative to the time of the initial perturbations/faults. Improved heuristics that allow for multiple time-varying faults would require the relativization of delays to the times of particular perturbations that normally trigger each mechanism. Such an improvement would enable diagnosis of situations where regulators cause perturbations that trigger other regulators. However, for the time being, we must content ourselves with situations in which the true faults happen at time zero while dormancies can happen anytime.

7 An application of Physiological MBD

Consider a simplified version of Widman's cardiovascular model [17] in which we consider only ten variables: blood volume (BV), extracellular volume (ECV), fluid intake (FI), heart rate (HR), mean arterial pressure (MAP), pulmonary capillary wedge pressure (PCW), renal blood flow (RBF), salt retention (SR), sympathetic nervous system stimulation (SNS), and urine output (UO). The constraints in the table below represent relationships between the **normalized** values of the above variables. The normalized value of a variable is given by its current value divided by its normal value, minus one. Thus, the normalized values of the variables in a system at normal steady state will all be zero.

Mechanism	Constraint	Delay	Duration
Impulse Transmission (IT)	HR = SNS	0	∞
Baroreceptor Response (BR)	SNS = -.6*MAP	0	1 day
Tubular Retention (TR)	SR = 3*SNS - .5*PCW	2 hours	∞
Urine Production (UP)	UO = RBF - SR	3 hours	∞
Fluid Conservation (FC)	ECV = FI - UO	3 hours	∞
Renal Hydraulics (RH)	RBF = MAP - 1.75*SNS	2 mins	3 days
Capillary Fluid Balance (CFB)	BV = ECV	5 mins	∞
Vascular Compliance (VC)	MAP = .2*BV	5 mins	∞

Table 1: Circulatory Model

As shown in Table 2, each mechanism has associated faults that can often explain its violated constraint. Consider the case of an acute hemorrhage, which, in our model, implies a violated fluid-conservation (FC) constraint. A hemorrhage reduces BV, which lowers MAP. The baroreceptors sense the pressure drop and react by increasing SNS stimulation. A raised SNS increases both HR (which raises cardiac output

Mechanism	Faults
Impulse Transmission (IT)	Damaged Nerves, SA-node problems
Baroreceptor Response (BR)	Faulty Stretch Receptors
Tubular Retention (TR)	Improper potassium or protein levels, damaged JGA
Urine Production (UP)	Problems in Glomerulus, Bowman's Capsule or Tubules
Fluid Conservation (FC)	hemorrhage, excessive sweating, diarrhea
Renal Hydraulics (RH)	abnormal renal resistance to blood flow
Capillary Fluid Balance (CFB)	improper plasma protein levels
Vascular Compliance (VC)	varicose veins, hardening of arteries

Table 2: Circulatory Faults

and MAP) and renin secretion, which leads to increased salt retention, SR. This causes fluid retention, which decreases urine output, UO, which then helps raise BV and MAP up closer to normal.

Changes to HR occur almost instantly, but salt retention takes hours. Hence, the tubule-retention (TR) mechanism will have a violated constraint for the first few hours after hemorrhage onset. Also, the baroreceptor response tends to wear off after about a day of continually low MAP. Hence, the BR constraint eventually becomes violated as MAP remains low but SNS returns to normal.

Assume that we measure 5 key variables (in a human) at three different time points and come up with the **normalized** values of Table 3. The only missing data is the value of PCW at 16 hours.

Variable	1 hour	16 hours	1.5 days
Fluid Intake (FI)	0	0	0
Mean Arterial Pressure (MAP)	-.1	-.06	-.05
Heart Rate (HR)	.06	.04	0
Pulmonary Capillary Wedge Pressure (PCW)	-.5	**	-.2
Urine Output (UO)	0	-.3	-.3

Table 3: Observations

At each time slice, constraint propagation through the model leads to contradictions between the signs of predicted and observed values. Hence, PMBD determines the minimal conflicts and candidates for each slice, as shown in Table 4. [3]

Time	Minimal Conflicts	Minimal Candidates
1 hour	(BR,RH,TR,UP), (IT,RH,TR,UP), (CFB,FC,VC)	(RH,CFB),(RH,FC),(RH,VC),(TR,CFB),(TR,FC) (TR,VC),(UP,CFB),(UP,FC),(UP,VC), (BR,IT,CFB),(BR,IT,FC),(BR,IT,VC)
16 hours	(CFB,FC,VC)	(CFB),(FC),(VC)
1.5 days	(BR,IT), (CFB,FC,VC)	(BR,CFB),(BR,FC),(BR,VC) (IT,CFB), (IT,FC), (IT,VC)

Table 4: Conflicts and Candidates

In Table 4, note that the 16-hour slice holds only one conflict and hence singleton candidates. So the PMBD system might conclude that the extra conflicts in the first and third slice are clearly due to dormant mechanisms. Hence, one of CFB,FC or VC must be the faulted mechanism. Although this is true, the above reasoning is incorrect due to an unmeasured PCW at 16 hours. PCW is a necessary observation in

[3] In Table 4, the notation $(M_i, M_j, ..)$ is shorthand for $\neg active(M_i) \diamond \neg active(M_j) \diamond ..$, where $\diamond$ denotes 'or' for conflicts and 'and' for candidates.

the derivation of the first two 1-hour conflicts. The absence of a PCW value at 16 hours indicates that more conflicts could be derived in its presence. Thus, the BR, RH, TR, UP, and IT mechanisms are not exonerated at the 16-hour slice, since each has PCW in its 'incriminating measurements' set at 1 hour. Instead, we must apply the heuristics of the previous section to follow a more logically safe path to a diagnosis.

Notice that TR and UP disappear from conflicts after the first measurement. This indicates that heuristic (6) might be useful, but it does not apply between the 1-hour and 16-hour slices, since [4]

$IM(TR, 1\_hour) \equiv IM(UP, 1\_hour) \equiv (MAP, HR, PCW, UO) \not\subseteq MV(16\_hours)$

However, it does apply between the first and third time slices, since (a) $MV(1\_hour) \subseteq MV(1.5\_days)$, and (b) TR and UP both have delays greater than 1 hour. The successful applications of (6) yield:

$became\_active(TR, (1\_hour, 1.5\_days)) \wedge became\_active(UP, (1\_hour, 1.5\_days))$

This information, along with the satisfaction of relation (8), entails a preference, among 1-hour candidates, for those containing TR or UP.

Conversely, the reappearance of BR in the 1.5-day conflicts (after an absence in the 16-hour conflicts) indicates that the baroreceptors may have deactivated. The additional fact that $IM(BR, 1.5\_days) \subseteq MV(16\_hours)$ permits the application of heuristic (7) to yield $became\_inactive(BR, (16\_hours, 1.5\_days))$. Furthermore, knowledge of BR's 0-delay advocates (via relation (11)) the removal of BR from all 1-hour conflicts. This change removes all triplet 1-hour candidates, and it also indirectly exonerates the IT mechanism, which no longer appears in any minimal candidates. Additionally, the 1.5-day candidates containing BR become preferred. Table 5 shows the updated conflicts and candidates, as well as the preferred (starred) candidates at each time slice.

Time	Minimal Conflicts	Minimal Candidates
1 hour	(RH,TR,UP), (CFB,FC,VC)	(RH,CFB),(RH,FC),(RH,VC),*(TR,CFB)*,*(TR,FC)* *(TR,VC)*,*(UP,CFB)*,*(UP,FC)*,*(UP,VC)*,
16 hours	(CFB,FC,VC)	(CFB),(FC),(VC)
1.5 days	(BR,IT), (CFB,FC,VC)	*(BR,CFB)*,*(BR,FC)*,*(BR,VC)* (IT,CFB), (IT,FC), (IT,VC)

Table 5: Revised Conflicts and Candidates

At the first and third time slices, the PMBD system can focus on the preferred candidates, each of which can be reduced via the removal of the presumably dormant mechanisms. This reduction yields the same three singleton candidates at each of the three slices: (CFB), (FC), and (VC). PMBD must then determine the best time slice and variable at which to take the next measurement.

It pays off to measure variables within time slices at which the suspect mechanisms should be active. Otherwise, PMBD is likely to derive more candidates with dormant mechanisms. Thus, PMBD prefers to take measurements at time slices that have a minimum ratio $\Omega(t_i)$, of potentially-dormant mechanisms among the slice's preferred suspects:

$$\Omega(t_i) = \frac{\| \{M_k : PSM(M_k, t_i) \wedge PDM(M_k, t_i)\} \|}{\| \{M_k : PSM(M_k, t_i)\} \|} \tag{12}$$

where$\|$ denotes set cardinality, and PSM stands for a 'preferred suspect mechanism': any mechanism that is (a) within the most preferrable candidates of a time slice, and (b) not assumed dormant. Finally, a 'potentially dormant mechanism', PDM, is defined as:

$$PDM(M_k, t_i) \leftrightarrow t_i < delay(M_k) \vee t_i > delay(M_k) + duration(M_k) \tag{13}$$

In the above example, $\Omega(1\_hour) = 1/3$, while $\Omega(16\_hours) = \Omega(1.5\_days) = 0$. The problem with the first time slice is that fluid conservation (FC), a prime suspect, is potentially dormant there. PMBD

[4] Since, in this example, each mechanisms slice-dependent incriminating-measurements (IMs) consist of a single set, we simplify the IM predicate to denote that set, instead of the more cumbersome set-of-sets used in the above general heuristics.

must now choose between the latter two time slices. The 1.5-day slice seems most appropriate since it has the same candidates (once PMBD reduces the preferred 1.5-day candidates) as the 16-hour slice, but the 1.5-day candidates are based on more measurements. Also, the choice of best measurement within the 1.5-day slice should be easier, since fewer unmeasured variables remain.

Within the 1.5-day slice, the five unmeasured variables are SNS, RBF, SR, BV, and ECV. Looking at the constraints in Table 1 for the three suspect mechanisms (FC, CFB, and VC), note that only two unmeasured variables participate in those three constraints: BV and ECV. Measuring BV would exonerate either VC, or both FC and CFB (i.e., come up with a definitive diagnosis that VC was faulted). Similarly, measuring ECV would exonerate either FC, or VC and CFB. None of the other 3 unmeasured variables can exonerate any of the prime suspects. So, after one or two measurements, the PMBD system would know that ECV = -.25, and that clearly contradicts the positive value predicted by VC (given FI = 0 and UO = -.3, the earlier observations), but it agrees with the ECV value predicted by the conjunction of VC and CFB and the observation that MAP = -.05. Hence, PMBD diagnoses the fluid conservation mechanism as faulty, and potential causes such as hemorrhage, excessive sweating, or diarrhea must be considered. A differentiation among these possibilities might require another round of MBD at a different granularity and with a new set of mechanisms and constraints.

8 Related Work

A few recent projects [1, 7] have integrated temporal information into model-based diagnosis in order to detect faults that only become evident during dynamic behavior. These approaches use dynamic models for behavior prediction and then generate candidates based on the values of system variables over time. However, the candidates themselves lack temporality: they hold at all times. Conversely, we perform constraint propagation across steady-state models at different time slices to derive time-tagged candidates.

Our approach is domain driven, since physiological systems often pass through a series of stages in adjusting to perturbations, where each stage has characteristically violated constraints. Therefore, we stick to the simpler (and more readily available in physiology) steady-state equations and apply them at (hopefully) significant time slices.

This time-slice orientation ressembles recent work in data interpretation [2], since both projects produce slice-dependent explanations (i.e., p-interps for Decoste, candidates for us) and then compare them across slices to provide global interpretations or diagnoses, respectively. However, DeCoste uses qualitative dynamic models, requires an a-priori complete envisionment, and diagnoses faulty measurements rather than mechanisms.

Finally, if we relax the distinction between faults and dormancies, our time-dependent-candidate approach is at least peripherally related to the intermittent-fault problem. The aforementionned stages of physiological adjustment are often defined by perturbations caused internally by regulators. Whether these count as intermittent faults is a matter of definition, but from a human diagnostician's point of view, they are symptom-causing factors whose causal origin (i.e., external or internal) is often irrelevant.

9 Discussion

We have shown how conventional model-based diagnosis can be applied to a medical domain. The important prerequisistes to this extension are (a) a dissection of physiological systems into modular mechanisms, (b) the representation of these mechanisms in terms of steady-state constraint equations, (c) the ability to propagate observations through these constraints in order to derive contradictions between predictions and observations, (d) the analysis of conflicts and candidates across multiple time slices in order to account for the varied timescales of physiological mechanisms.

Widman's semi-quantitative modeling approach provides a modular view of physiological mechanisms along with constraint equations summarizing their normal behaviors. This representation unifies the envisionment model used for predicting future behavior, and the diagnostic model for reasoning backwards in time from symptoms to faults. This paves the way for multi-directional constraint propagation through the many feedback loops of physiological models. The contradictions derived via constraint propagation then enable standard MBD over the mechanisms involved in the derivation. However, the non-zero delays

and finite durations of physiological mechanisms necessitate MBD at multiple time points in order to differentiate faulted from dormant mechanisms. We provide a few of the necessary extensions to MBD such as (a) heuristics for assuming dormancy based upon temporal knowledge of mechanisms and comparative analyses of conflicts and candidates across time, (b) criteria for focusing diagnosis based upon dormancy assumptions, and (c) guidelines for selecting measurements across temporal slices.

As future work, we hope to integrate the implemented aspects of this research with the theoretical developments discussed above. Although Widman's semi-quantitative simulator has been implemented and employed in cardiovascular diagnosis, we have yet to use it for incremental model-based diagnosis; that is, as a generator of both diagnostic hypotheses and next-best measurements. To this end, we must also formalize measurement-making decisions across multiple time slices, hopefully via a temporal extension of the popular information-theoretic approach of [4].

Finally, we hope to integrate fault models [3, 13, 6] into the PMBD framework, since some faults in physiological systems are more easily described as abnormal variable settings than as violated constraints. A combination of both representations appears advantageous for hierarchical diagnosis, since a violated constraint at one level is often explained by the value of a lower-level variable. For instance, a violated capillary fluid balance (CFB) constraint might point to a low plasma protein level. In addition, fault modes will assist in intermittent-fault diagnosis, since many regulator-induced faults involve alterations to common variables such as arterial resistance and venous compliance.

References

[1] P. Dague, P. Devès, P. Luciani, and P. Taillibert. Analog Systems Diagnosis *Proceedings ECAI*, 173–178, Stockholm, Sweden, 1990.

[2] D. DeCoste. Dynamic across-time measurement interpretation. *Proceedings AAAI*, 373–379, 1990.

[3] J. deKleer and B. Williams. Diagnosis with behavioral modes. *Proceedings IJCAI*, 1324–1330, 1989.

[4] J. deKleer and B. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 31(1):97–130, 1987.

[5] J. deKleer. An assumption-based truth maintenance system. *Artificial Intelligence*, 28(2):127–162, 1986.

[6] K. Downing, and J. Shrager. Causes to clauses: Managing assumptions in qualitative medical diagnosis. *International Journal of AI in Engineering* 3(4), pages 192–199, 1988.

[7] T. Guckenbiehl, and G. Schäfer-Richter. SIDIA - Extending prediction-based diagnosis to dynamic models. *International Workshop on Principles of Diagnosis*, Stanford University, 1990.

[8] A. Guyton, C. Coleman, R. Manning, and J. Hall. Some problems and solutions for modeling overall cardiovascular regulation. *Mathematical Biosciences*, 72:141–155, 1984.

[9] Ironi, L., Stefanelli, M., and Lanzola, G. Qualitative models in medical diagnosis. Pubblicazioni N. 737, Institute Di Analisis Numerica, Pavia, Italy, 1990.

[10] B.J. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29(3), pages 289–338, 1986.

[11] R. Patil. Artificial Intelligence for diagnostic reasoning in medicine. In H. Shrobe, editor, *Explorations in Artificial Intelligence*, pages 347–379, Morgan Kaufmann Publishers, 1988.

[12] R. Patil, P. Szolovits, and W. Schwartz. Modeling knowledge of the patient in acid-base and electrolyte disorders. In P. Szolovits, editor, *Artificial Intelligence in Medicine*, pages 191–226, Westview Press, 1982.

[13] P. Struss and O. Dressler. Physical negation - Integrating fault models into the general diagnostic engine. *Proceedings IJCAI*, 1318–1323, 1989.

[14] S.M. Weiss, C.A. Kulikowski, S. Amarel, and A. Safir. A model-based method for computer-aided medical decision-making. *Artificial Intelligence*, 11:145–172, 1978.

[15] L.E. Widman. Semi-quantitative "close enough" systems dynamics models: An alternative to qualitative simulation. In L.E. Widman, K.A. Loparo, and N.R. Nielsen, editors, *Artificial Intelligence, Simulation and Modeling*, pages 159–188, John Wiley & Sons, New York, 1989.

[16] L.E. Widman, Y.-B. Lee, and Y.H. Pao. Towards the diagnosis of medical causal models by semi-quantitative reasoning. In P.L. Miller, editor, *Topics in Medical Artificial Intelligence*, pages 55–70, Springer-Verlag, New York, 1988.

[17] L.E. Widman. Expert system reasoning about dynamic systems by semi-quantitative simulation. *Computer Methods and Programs in Biomedicine*, 29:95–113, 1989.