Evaluating Case-Based Reasoning for Heart Failure Diagnosis

by

David S. Aghassi
A.B. Computer Science, Brown University (1988)

Submitted to the Department of
Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Science

at the
Massachusetts Institute of Technology
June, 1990

Signature of Author .... David S. Aghassi ....
Department of Electrical Engineering and Computer Science
May 11, 1990

Certified by .... Peter Szolovits ....
Peter Szolovits
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Arthur C. Smith, Chair
Departmental Committee on Graduate Students

Evaluating Case-Based Reasoning for Heart Failure Diagnosis

by

David S. Aghassi

Submitted to the Department of
Electrical Engineering and Computer Science
on May 11, 1990 in Partial Fulfillment of the Requirements for the Degree of
Master of Science

Abstract: In routine problem solving, people reason from experience, remembering their solutions to recurrent problems rather than reconstructing them from scratch each time. The method of case-based reasoning attempts to exploit this intuitive strategy on a computer, by maintaining a memory of precedents, and by solving a new case according to the solution of the most suitable precursor. Diverse applications of the method seem to suggest its viability, but a widespread lack of thorough evaluation questions this support. Indeed, while previous work implies that case-based reasoning is successful for a variety of domains, few papers identify the general relationships between performance and the domain characteristics and scaling factors that underlie it. Thus, researchers are left without an understanding of the method's scope or scale, and intuitions about human experience continue to be its primary justification.

This study addresses many of these open concerns in the context of heart failure diagnosis, evaluating the existing case-based reasoner CASEY with respect to a pool of 240 patients. To investigate the method's scale, I measured the effects of increasing experience on both accuracy and efficiency. I also analyzed the distribution of cases in order to quantify its intrinsic regularity, thus exposing the dependence of the system's utility on the domain and facilitating an extrapolation of this utility to other, similarly characterized applications. First, I gauged the recurrence of similar cases in varying size collections of patients; secondly, I measured the correlation between symptomatic similarity and diagnostic similarity; and finally, I appraised the absolute diagnostic homogeneity of the case pool.

Because cardiologists claim that most cases are variations on recurring, well-understood pathophysiologic themes, I expected to justify the application and verify the presumed regularity upon which its success depends. Instead, I discovered that CASEY's accuracy does not increase with experience, while its efficiency degrades with the number of available precedents. Fundamentally, similar cases and similar diagnoses were rare among the 240 patients, and moreover, symptomatic resemblances did not guarantee diagnostic correspondence. Because of the varying combination and interaction of multiple diseases, the patients were largely heterogeneous, suggesting that the regularity described by cardiologists occurs at a more detailed level of abstraction, perhaps in the recurrence of diagnostic syndromes comprised within the cases. This more fine-grain uniformity can be exploited only by analyzing precedents, rather than by applying them in their entirety.

Thesis Supervisor: Peter Szolovits
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgements

I would like to thank:

Peter Szolovits, my advisor, for his continual warm encouragement and skillful guidance, and particularly for his wholehearted commitment to my efforts at those times when I needed his support most.

William Long, whose deep support and valuable instruction were unremitting, and who never seemed to mind my interruptions, despite his busy schedule.

Phyllis Koton, whose work inspired me, and whose discussions stimulated me to think.

Ramesh Patil, who extended genuine and unreserved encouragement.

The members of the Clinical Decision Making Group, especially Tom Russ, Tom Wu, and Alex Yeh, who were always willing to teach me more about whatever I needed to know.

Joni Beshansky at Tufts New England Medical Center, for locating and collecting as many as 240 heart failure cases.

My parents, for their love throughout.

This research was supported by the National Institute of Health grant R24 RR01320 from the Division of Research Resources.

Contents

List of Figures

Chapter 1

Introduction

Reasoning from experience is, *intuitively*, a powerful alternative to problem solving from scratch. Most real world domains harbor tremendous regularity, in that similar situations recur, each time requiring similar responses. While these responses remain applicable, employing experience is effective and accurate. Experience also engenders efficiency, because it obviates the reconstruction of expensive analysis and deliberation. Rarely should a problem require as much attention the second time.

People rely upon and exploit these principles, using the solutions they construct repeatedly, often adapting them and tailoring them to fit new circumstances. When we give advice, we inevitably bring up "the time that I ... " and proceed to recount the appropriate anecdote. In fact, we are rarely considered sources of advice unless we have a reputation for possessing experience. *Experience* is thus the definitive difference between the *expert* and the novice. Not surprisingly, reasoning from cases or precedents is the paradigm for experts in fields as dissimilar as medicine, law, management, political science, and economics [25].

Medical diagnosis, in particular, relies critically on experience. Students spend two years in rotation and at least three additional years of residency to augment their basic knowledge with clinical "practice." Seasoned physicians diagnose unusual cases by extrapolating from relevant precedents, while able to solve familiar cases almost without thinking. Indeed, a substantial portion of any physician's diagnoses are routine, and accordingly, the knowledge involved is automatic.

Case-based reasoning (CBR) is a form of reasoning from experience based on discrete problem-solving situations called "cases." *Computationally*, CBR resembles dynamic programming, an acceleration technique that advocates caching the results of calculations to avoid repeating them. To this end, a case-based reasoner stores previous cases with their solutions in a *case base.* Faced with a new problem, the reasoner searches this memory for appropriate *precedents* and then derives a solution from the best matched, or most similar, case retrieved. Numerous implementations substantiate the plausibility of this approach, and their success suggests broad applicability.

Methodologically, however, very little has been done to validate these optimistic indications, and few papers reach strong conclusions concerning the generality of CBR's success. Questions such as "What characteristics of the domain make CBR a useful approach?", "When is CBR better than other techniques?", and "How do accuracy and complexity scale with the number of cases?" are answered without formal or empirical support. Current research focuses on enhancing and applying the technique without yet understanding its scope or scale.

To address these open concerns, this thesis presents an evaluation of Phyllis Koton's CASEY [14], an existing application of CBR to the diagnosis of heart failure.[1] Using a pool of 240 cases provided by the Tufts New England Medical Center, I conducted a thorough empirical study of the case-based approach, investigating both its admissibility within the heart failure domain and its profitability as implemented in CASEY. Because clinical cardiology is a comparably systematic medical discipline, involving recurring presentations of well-understood pathophysiology, I expected to justify the application. A large proportion of heart failure patients are variations on the common themes of ischemic, hypertensive, and alcohol related heart disease.[2] Surprisingly, however, my results indicate that even this apparent regularity is too diffuse to make CASEY succeed.

[1] Although its design remains intact, I have rewritten and improved much of CASEY's implementation. Technically, I should refer to the new system as CASEY+, but I will continue to call it CASEY, for purposes of simplicity.

[2] William Long, personal communication.

Chapter 2

Case-Based Reasoning

2.1 Background

The origins of case-based reasoning can be traced to Schank [19], who identifies "reminding" as the motivating feature of any dynamic memory. Reminding occurs when a new experience, in the process of being integrated into memory, collides with a previous episode sharing similar situational details. In this sense, the new experience "reminds" us of the previous encounter. Because memories are organized according to their distinguishing features, an experience is destined to evoke its historical counterparts, simply by virtue of seeking its place in the proper conceptual cluster. Precedents raise expectations concerning the new situation and often become the inductive basis for generalization. Accordingly, Schank feels that "reminding is at the root of how we understand ... at the root of how we learn" [19].

Kolodner expanded Schank's initial intuitions by describing a memory indexing structure that facilitates reminding on a computer [10]. Within her scheme, memories are clustered into "organization packets," which specify normative features for the episodes they subsume. Beneath each packet, memories are indexed according to their differences from the norm, first by the features through which they differ, and then by their distinctive feature values. Thus, two situations are indexed together if they have distinguishing traits in common.

Kolodner employed this indexed memory construct, along with the notion of analogical transfer proposed by Carbonell [2], in her subsequent definition of the CBR paradigm [12]. The procedure comprises three rudimentary phases:

1. Integrate the new case into memory, retrieving all similar cases encountered.
2. Evaluate these precedents for relevance to the current case.
3. Transfer the solution of the best matched precedent to the new situation, adjusting it according to differences between the two cases (if possible).

Some case-based reasoners also include a fourth phase to address the contingency that no applicable precedents are found. In this phase, the reasoner may resort either to problem solving from first principles or to the advice of an expert.

2.2 Intuitive Support

Initially, intuition was used to justify the experiential approach. Schank draws examples from everyday life, citing, for instance, the similarity between a trip to Burger King and a trip to McDonalds. Certainly, one prepares us for the other [19]. Carbonell argues that although a person may have no truck-driving experience, (s)he could successfully employ knowledge about automobile motoring when faced with navigating the unfamiliar vehicle [2]. Kristian Hammond appeals to our sensibility with aphorisms: "If it worked, use it again" [7]. More persuasive still is his invocation of Santayana's admonishing words:

> Those who cannot remember the past are condemned to repeat it.

Unfortunately, there are circumstances under which these intuitions break down. To ensure that we do not approach CBR with unsound assumptions, we must confront the following non-obvious questions:

- How often might the past repeat?
- Under what conditions do our memories remain valid?
- To what extent do they apply to other situations within the same context?
- When is repeating the past *more* efficient than remembering it?

2.3 Methodology and Case-Based Reasoning

Many of the seminal ideas for incorporating methodology into case-based reasoning research were presented at the 1989 DARPA sponsored Case-Based Reasoning Workshop. Phyllis Koton and Paul Cohen express several pressing reasons for the importance of this undertaking. First, CBR is no longer solely a vehicle for psychological modeling. Applied increasingly to real-world problems, it is obliged to acquire justifications more substantial than "plausibility." In Koton's words, "With increased visibility comes increased responsibility" [15]. Secondly, the emerging field must appraise itself critically before it allows its opponents to do so: indeed, by upholding standards within its realm, it will earn the respect of the general research community. Most crucially, research must be directed to be effective. Relying on demonstration and exposition, researchers may forfeit many of the insights they could derive from evaluating their efforts, empirically or otherwise. Without a methodological framework, they can not attribute definite meaning to their results, and are destined to "do clumsy, inconclusive, redundant research!" [4].

Cohen maintains that productive research addresses itself to certain fundamental questions. To investigate a method such as CBR, according to his agenda[4], one might inquire:

- What criteria should be used to judge the method's success?
- What assumptions does the method presume?
- Why is the method correct?
- What is the scale and complexity of the method?
- How is the method more effective or efficient than other techniques?
- For what class of tasks is the method general?

Furthermore, the experiments that answer these questions should be guided themselves by more specific questions [3], such as:

- How representative are the test cases?
- What limitations do they illustrate?
- How robust is the implementation?
- Is its performance predictable?
- Is it efficient?
- What aspects of the program are crucial or incidental to its success?
- Why are results confirming or unexpected?
- Can these results be generalized?

2.4 Previous Work in Validation

Although substantial research has been devoted to implementing case-based reasoners and demonstrating their capabilities in various domains [12,6] (see also the 1988 and 1989 *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*), few papers address these questions convincingly. Typically, evaluating case-based reasoning has involved measuring the percentage of test case solutions that meet some standard of accuracy. This standard may be defined either by the judgement of an expert, by comparison to a solution entailed by first principles, or by the real-world success of the case-based solution itself. Paul Cohen laments, though, that "We never learn why the author believes his or her scheme is a good idea, only that it works for a handful of examples in a program which serves only to produce demonstrations" [4].

Several researchers, however, have moved somewhat beyond the prescription. Marc Goodman, who developed a case-based reasoner for battle planning, correlated the accuracy of his system's predictions with the number of cases it retrieved [5]. His work confirms that (in battle planning, at least) the number of potential precedents increases the likelihood that a good match will be found. William Mark notes that "the value of our case based approach will be measured by the ease with which it can be applied to other problems" [18]. He anticipates expanding the use of his autoclave layout manager, Clavier, to different autoclave applications and to other configuration problems.

Ray Bareiss devised and executed a diversity of tests for Protos, his case-based classifier of hearing disorders [1]. After comparing its accuracy to that of experts, intermediates, and novices, as well as other case-based and inductive approaches, Bareiss concluded that the case-based method was more effective than all but the experts (for his domain). Moreover, he determined that, as the number of cases increased, Protos' knowledge base experienced linear growth at most. Thus, (in clinical audiology) CBR attains some measure of efficiency. Finally, he identified the case base as his reasoner's "source of power" through ablation studies in which he eliminated indexing and matching knowledge.

In conclusion, however, Bareiss acknowledged that "even the results of a battery of experiments as comprehensive as those described cannot be taken as validation of an approach to learning and problem solving. In particular, our results are generalizable only insofar as the audiology task is representative of a larger class" [1]. Thus, while researchers are attempting to broaden their evaluations, they are limited by the dependence of their results on the domain. Indeed, merely exhibiting a domain for which CBR succeeds does not entail general applicability when the factors responsible for success are not identified. In response to this concern, Phyllis Koton makes an initial attempt to characterize domains that admit CBR [16]. Her criteria are as follows:

1. "Similar problems recur." If history repeats often, then remembering the past is useful. The benefits of CBR outweigh the costs if:

$$\begin{array}{rc} & \text{the probability of finding a good match} \\ \times & \text{the time gained by not having to recompute the solution} \\ & > \\ & \text{the probability that no match is found} \\ \times & \text{the time wasted searching and examining precedents}^{1} \end{array}$$

2. "The domain is stable under perturbations." If similar problems require similar solutions, then memories of these solutions are broadly applicable. Conversely, when almost all aspects of a precedent solution must be modified, it is more efficient to start from scratch.

3. "Interactions are limited." For solution transfer to be efficient, the number of modifications required should be proportional to the number of differences between the old and new cases. Thus, modifications can be

[1] This equation, paraphrased from Koton's paper, does not account for the possibility of error. Specifically, the solution derived from a precedent might be unsuitable despite the fact that the precedent was judged to be a "good match." Certainly, criteria 2 and 3 assure us that this possibility is rare, but they do not rule it out entirely.

made locally and incrementally. (Although stated as a separate criterion, this is—I believe—a qualification of 2.)

Koton's criteria make a significant step towards delineating the scope of CBR. However, as her work is recent, no confirming empirical studies have yet been attempted.

Chapter 3

CASEY

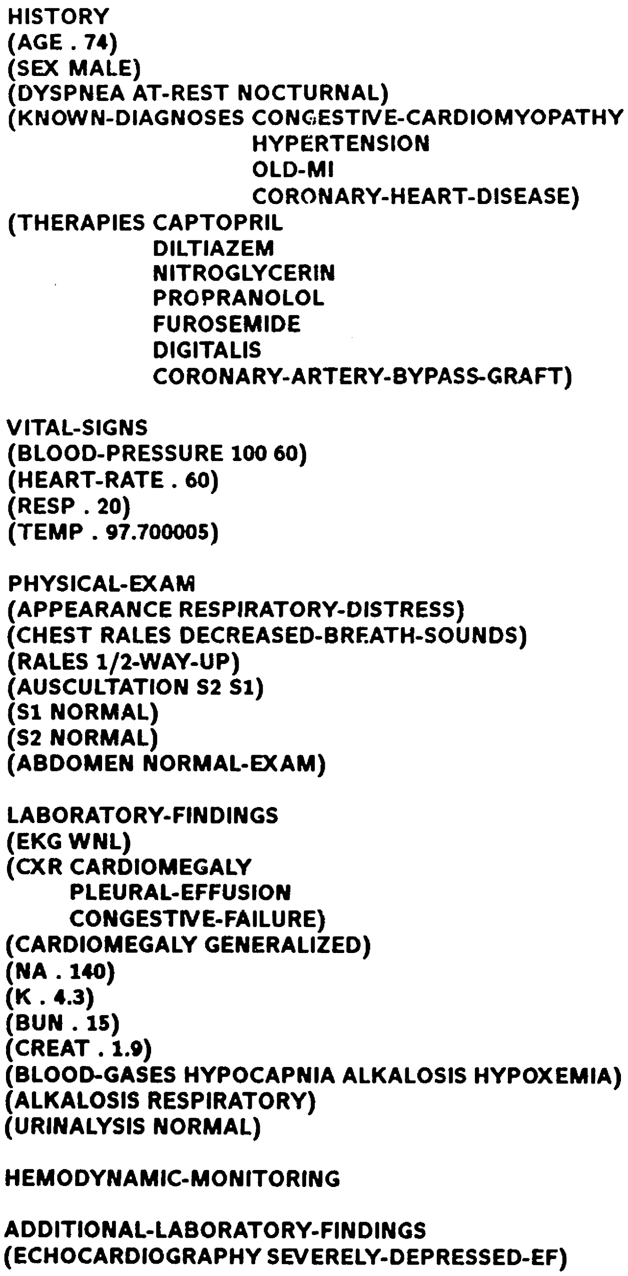
3.1 The Heart Failure Domain

The CASEY system, designed by Phyllis Koton, applies CBR to the medical domain of heart failure. Heart failure occurs in a patient when cardiac output, the induced rate of blood flow through the heart, is too low to serve the body's requirements. The condition may be triggered by many different cardiac diseases, and additionally, symptoms typical of heart failure may be produced by diseases of other organ systems. However, the goal of heart failure diagnosis is not merely to pinpoint underlying causes, because many of the implicated diseases are chronic and essentially incurable. Instead, the physician must infer the physiological mechanisms producing the symptoms, so that these causal processes can be intercepted through therapy. The task is complicated by the prevalence of multiple diseases and by the effects of previous intervention.

CASEY is not equipped to diagnose heart failure from first principles. It is grounded, however, on the model-based Heart Failure system [17], which automates this complex undertaking. Developed by William Long, Heart Failure reasons from a causal network of about 300 potential *pathophysiologic states.* Each state is represented by a node in the network and describes a condition such as AORTIC STENOSIS or HIGH LEFT ATRIAL PRESSURE. A causal *link* between two nodes specifies that one causes the other with a given conditional probability. Similar links embody the probabilistic causal relations between pathophysiologic states and symptoms (also called *findings*). Through these links, findings and intermediate states can be "explained" by one or more of their potential causes. A pathophysiologic state that needs no explanation is called a *primary cause.*

Heart Failure receives as input a list of symptoms, or patient findings, expressed as ⟨feature value⟩ pairs. Each of the several hundred potential features has a prespecified range of possible values. For example, the numerical feature "heart-rate" can be valued anywhere between 30 and 200 beats per minute. The categorical feature "cough" can be characterized as "present" or "absent." Normal values, when known, are included. Figure 3.1 exhibits a sample patient description.

Given the patient description, Heart Failure constructs a *causal explanation*, a diagram which links the patient's abnormal findings to underlying primary causes, through various intermediate states and causal links. Figure 3.2 displays a sample causal explanation; findings appear in lowercase, and primary causes are bold. Within such a hypothesis, every finding and state must be accounted for by another state unless it is primary. Thus, the causal explanation identifies a high probability pathophysiologic mechanism that produces



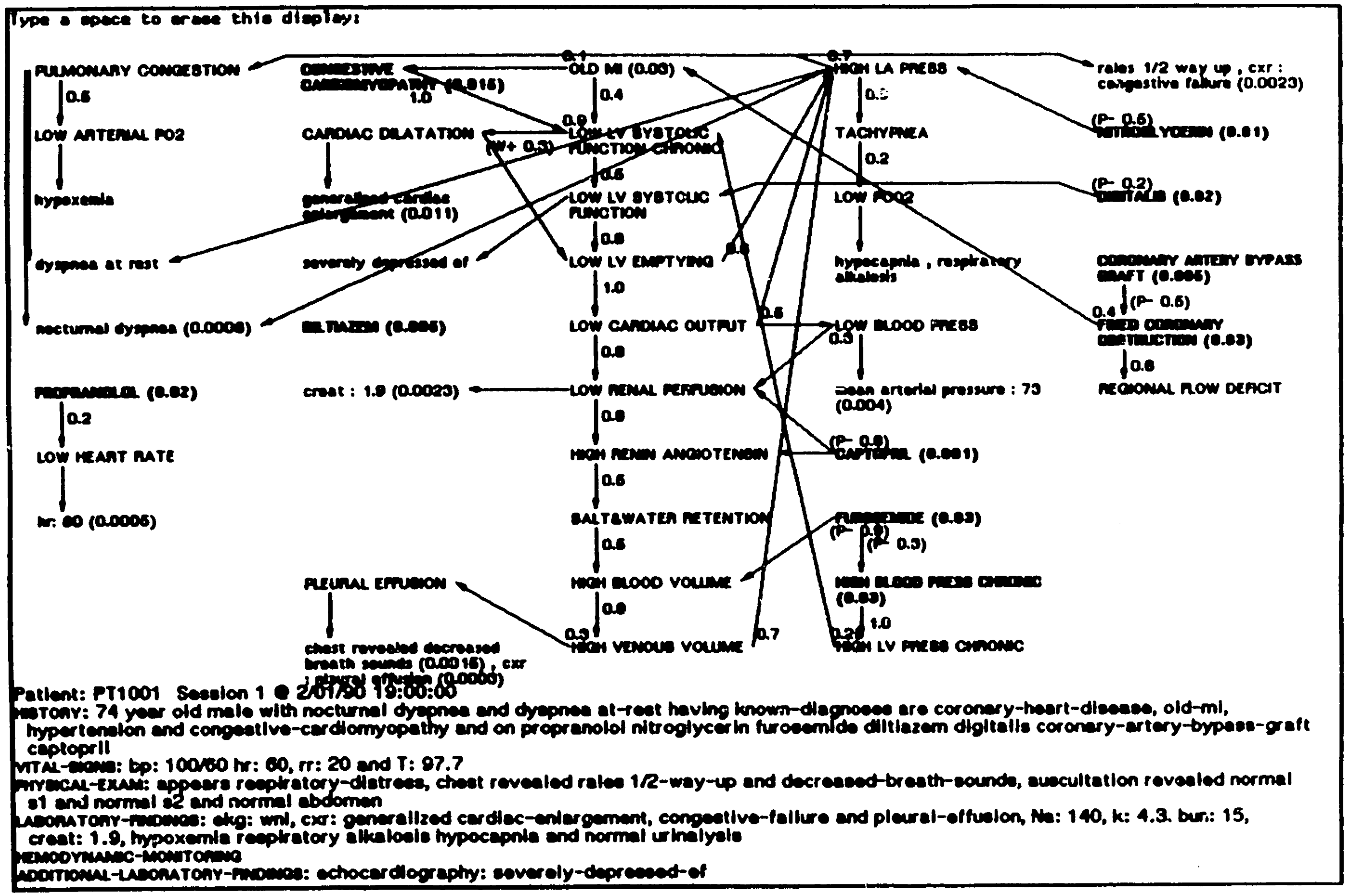
```
HISTORY
(AGE . 74)
(SEX MALE)
(DYSPNEA AT-REST NOCTURNAL)
(KNOWN-DIAGNOSES CONGESTIVE-CARDIOMYOPATHY
                 HYPERTENSION
                 OLD-MI
                 CORONARY-HEART-DISEASE)
(THERAPIES CAPTOPRIL
           DILTIAZEM
           NITROGLYCERIN
           PROPRANOLOL
           FUROSEMIDE
           DIGITALIS
           CORONARY-ARTERY-BYPASS-GRAFT)

VITAL-SIGNS
(BLOOD-PRESSURE 100 60)
(HEART-RATE . 60)
(RESP . 20)
(TEMP . 97.700005)

PHYSICAL-EXAM
(APPEARANCE RESPIRATORY-DISTRESS)
(CHEST RALES DECREASED-BREATH-SOUNDS)
(RALES 1/2-WAY-UP)
(AUSCULTATION S2 S1)
(S1 NORMAL)
(S2 NORMAL)
(ABDOMEN NORMAL-EXAM)

LABORATORY-FINDINGS
(EKG WNL)
(CXR CARDIOMEGALY
     PLEURAL-EFFUSION
     CONGESTIVE-FAILURE)
(CARDIOMEGALY GENERALIZED)
(NA . 140)
(K . 4.3)
(BUN . 15)
(CREAT . 1.9)
(BLOOD-GASES HYPOCAPNIA ALKALOSIS HYPOXEMIA)
(ALKALOSIS RESPIRATORY)
(URINALYSIS NORMAL)

HEMODYNAMIC-MONITORING

ADDITIONAL-LABORATORY-FINDINGS
(ECHOCARDIOGRAPHY SEVERELY-DEPRESSED-EF)
```

Figure 3.1: A patient description for case PT1001

the symptoms presented. In general, the connectivity of a causal explanation is determined uniquely by the states and findings it encompasses; once a particular set of pathophysiologic states has been hypothesized, the system assumes that all known influences among them will contribute to the disease mechanism.

Because each finding and intermediate state has several potential causes, the number of potential explanatory pathways grows exponentially. Since paths can span as many as twelve links, there are about 7,000 possible. Precomputing and pruning them relieves some of the complexity, but nonetheless, every path's probability must be calculated and ranked when the patient information is entered. Heart Failure searches backward through the paths from each finding for the primary nodes that account for it, identifying a minimal (or at least small) set of diseases that "cover" most of the findings. As findings are integrated into the complete explanation, each potentially relevant path is considered. In fact, several consistent combinations are constructed simultaneously. Finally, the hypotheses are evaluated and the most probable ones are selected.

Given the problem of constructing, from scratch, a high probability causal explanation for a set of findings, the system is surprisingly efficient. Its clever heuristics prune the unmanageable space of solutions so that diagnosis is fast enough to be pragmatic and yet robust enough to be powerful. Indeed, Heart Failure can diagnose a typical patient in under a minute, while still able to solve a broad range of common and unusual cases. However, although its heuristics reduce the diagnostic computation significantly, the system must *inevitably* consider, evaluate, and combine exponentially many potentially rel-



Type a space to erase this display:

Patient: PT1001 Session 1 @ 2/01/90 19:00:00
HISTORY: 74 year old male with nocturnal dyspnea and dyspnea at-rest having known-diagnoses are coronary-heart-disease, old-mi, hypertension and congestive-cardiomyopathy and on propranolol nitroglycerin furosemide diltiazem digitalis coronary-artery-bypass-graft captopril
VITAL-SIGNS: bp: 100/60 hr: 60, rr: 20 and T: 97.7
PHYSICAL-EXAM: appears respiratory-distress, chest revealed rales 1/2-way-up and decreased-breath-sounds, auscultation revealed normal s1 and normal s2 and normal abdomen
LABORATORY-FINDINGS: ekg: wnl, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 140, k: 4.3. bun: 15, creat: 1.9, hypoxemia respiratory alkalosis hypocapnia and normal urinalysis
HEMODYNAMIC-MONITORING
ADDITIONAL-LABORATORY-FINDINGS: echocardiography: severely-depressed-ef

Figure 3.2: A causal explanation for case PT1001

evant pathways. Since remembering its solutions might save the system this computation in the future, the domain appears ripe for case-based reasoning.

3.2 Implementation

CASEY [14] attempts to achieve efficiency through case-based reasoning without sacrificing the effectiveness and completeness of Heart Failure. For every patient it sees, it remembers the set of findings input and a complete causal explanation. Thus, when it recognizes a new problem as familiar, it can retrieve and transfer the appropriate solution from its case base. However, CASEY also realizes when a problem is not familiar, calling Heart Failure to construct a new diagnosis, which it adds to its repertoire. The system's functionality can be divided into three aspects, following the case-based reasoning paradigm: precedent retrieval, precedent evaluation, and solution transfer.

To understand CASEY's approach to precedent retrieval requires an understanding of its memory, whose implementation closely resembles Kolodner's construction [10]. Individual cases are stored beneath a discrimination network of *generalizations* (GENs), such that each GEN records resemblances between the cases and sub-generalizations it subsumes, while indexing these cases and sub-generalizations by the features that distinguish them. A GEN retains the patient findings that are common to at least 2/3 of the cases it encompasses and stores pathophysiologic states included by the causal explanations of all of its cases. Figure 3.3 displays a sample GEN.

Within a GEN, cases and sub-generalizations are discriminated on the basis of two levels of indexing: the first level specifies distinguishing features, and

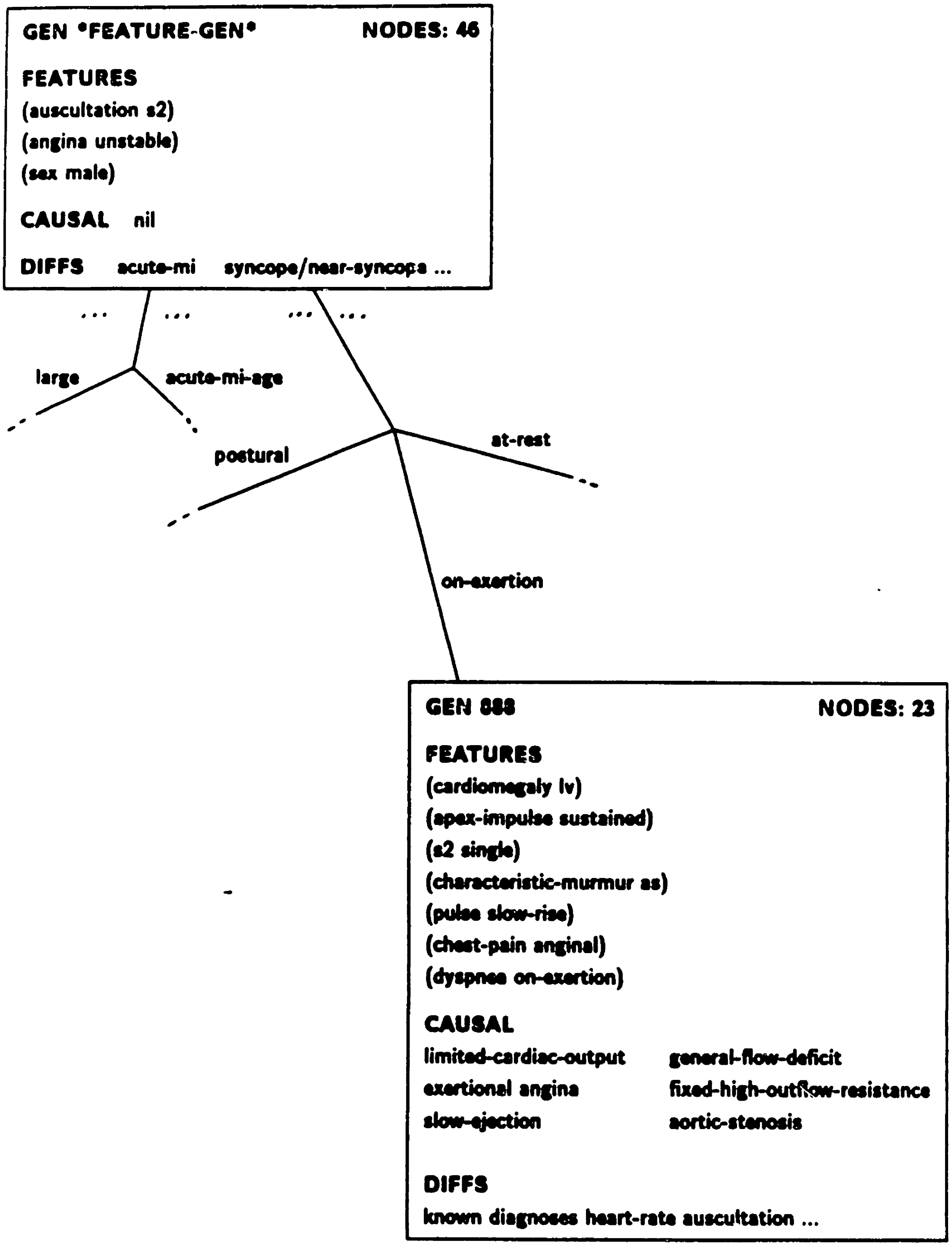


Figure 3.3: A fragment of CASEY's memory structure (taken from Koton's doctoral dissertation [14])

the second, distinguishing values for those features. For example, if a case is atypical by virtue of the finding ⟨syncope on-exertion⟩, it is indexed first under "syncope", and then under "on-exertion". The indexing scheme is redundant, because it catalogs a case according to *every* finding that differentiates it from the norm, at *every* level of the discrimination network. Thus, there exist several paths through memory to any one case, each generalizing the case on the basis of different attributes. Regrettably, the redundancy is necessary, as subsequent cases may match any subset of these attributes without matching the others. Indeed, different findings are important to different cases, and thus, restricting the set of indices might cause CASEY to miss the most suitable precedent for a new case.

CASEY is reminded of precedents in the process of incorporating a new case into its memory. When the new case encounters a GEN, CASEY compares them, remembering the GEN if the new case is representative in all aspects. Otherwise, CASEY searches every index-pair specifying a finding that distinguishes the new case from the GEN norm. If any of these index-pairs are "empty," the case is stored there, and the GEN becomes a reminding. If subgeneralizations are encountered, they are searched recursively. Finally, if an individual case is found, CASEY replaces it with a new GEN that embodies its similarity to the new case. The two cases are then stored beneath this GEN on the basis of their distinguishing features, and the precedent case becomes a reminding.

Once CASEY retrieves potential precedents, it must evaluate their applicability to the new case. In the heart failure domain, as in many others, different features can be manifestations of the same underlying state. For example, the findings ⟨ekg lv-strain⟩ and ⟨chest-xray lv-enlargement⟩ both provide evidence for the more general state LV-HYPERTROPHY.[1] Thus, the feature-based similarity that was used to retrieve the precedents is not necessarily the optimal basis for evaluation and comparison. Instead, cases are more likely to have similar causal explanations if they share *generalized causal features*, such as "evidence of LV-HYPERTROPHY," which subsumes both ⟨ekg lv-strain⟩ and ⟨chest-xray lv-enlargement⟩. In fact, CASEY maintains a second discrimination net that indexes cases exclusively through their generalized causal features; CASEY searches this memory for additional remindings.

CASEY's similarity metric for ranking potential precedents counts generalized causal features matched by new case evidence and subtracts from these the number unmatched by new case evidence. Thus, precedents are given priority if (1) they match a large number of new case findings and (2) the new case accounts for a large proportion of their pathophysiologic states. After the precedents are ranked, CASEY attempts to transfer the solution of its best candidate. If the attempt fails, it continues through the remaining precedents until a successful transfer occurs. When none of its precedents are appropriate, it requests a diagnosis from Heart Failure.

CASEY also makes extensive use of Heart Failure in justifying and adapting a solution for transfer. Solution adaptation proceeds in three stages. First, the system removes from the precedent hypothesis all disease states contradicted by new case findings. Secondly, CASEY incorporates those findings that can be directly explained by precedent pathophysiologic states. Symptoms remaining

[1] "lv" is an abbreviation for "left ventricular."

unexplained are accounted for, using the model, by augmenting the causal explanation with additional pathways. If any findings defy explanation, given the precedent diagnostic scheme, the transfer fails. Finally, CASEY prunes the resulting solution by eliminating all disease states bereft of evidence, failing here if no primary states remain. Otherwise, transfer is complete.

Although CASEY guarantees a valid and plausible explanation for each new case, its reasoning is heavily biased by the diagnostic hypothesis established in the precedent. Indeed, its very efficiency derives from the locality of its modifications and from the fact that these modifications are bounded by the number of differences between the precedent and the new case. Consequently, it is not able, like Heart Failure, to ensure a high probability explanation for the new case findings considered as a whole. Presumably, it will approach this ideal when the differences are small.

Chapter 4

Evaluating CASEY

4.1 Criteria For Success

Before we can even begin to evaluate CASEY, we must answer the question: "What are our criteria for success?" Certainly, a range of judgments may be made using measures as different as elegance, cognitive validity, and utility. For the present evaluation, utility seems the most objective and most appropriate standard; even so, there exist several possible ways to define and assess it.

Two primary and necessary contributors to utility are accuracy and efficiency. CASEY's accuracy ultimately refers to the quality of its diagnoses. Yet because all domain knowledge and precedent solutions are obtained exclusively from Heart Failure, CASEY is constrained by the correctness of this underlying system; it must not be penalized for reasoning from incorrect solutions that it receives on faith. Therefore, to judge the accuracy of the case-based reasoning involved, independent of Heart Failure, we establish the model-based solutions as ideal, measuring CASEY's analogically derived causal explanations against that standard.

One technique for assessing relative accuracy might quantify the discrepancy between the ideal solution and the case-based approximation, by enumerating the pathophysiologic states *missing* from CASEY's causal explanation along with the superfluous states included *spuriously* therein. The error fraction ϵ will be computed accordingly, as:

$$\frac{\begin{array}{l}\text{Number of \textit{missing} states}\\ \text{included in Heart Failure's}\\ \text{causal explanation but}\\ \text{omitted from CASEY's}\end{array} + \begin{array}{l}\text{Number of \textit{spurious} states}\\ \text{included in CASEY's causal}\\ \text{explanation but not in Heart}\\ \text{Failure's}\end{array}}{\text{Total number of states in Heart Failure's causal explanation[1]}}$$

The error measure ϵ is merely a syntactic, structural criterion, computed independently of the semantics and function of the diagnoses that it assesses. An alternative and perhaps more sensitive measure of accuracy would consider the likelihood that CASEY's causal explanation is the actual mechanism behind the findings presented.[2] The Heart Failure System can make this assessment using the prevalences and link probabilities specified in its causal model. Because I am interested in relative accuracy, however, I will compute the relative likelihood λ as the ratio of the likelihood of CASEY's solution to the likelihood

[1] This fraction can exceed the value 1 when the spurious states outnumber the accurate pathophysiologic states.

[2] While even richer criteria might take into account the cost of terminal illness or the utility of treatable disease, this study lacks a solid basis for such judgments and therefore will not go beyond the two more conventional measures.

of the ideal:

$$\frac{\text{Likelihood of CASEY's causal explanation}}{\text{Likelihood of Heart Failure's causal explanation}}$$

As its experience grows, CASEY's relative accuracy should increase asymptotically. The system is efficient, then, as long as it requires only moderately increasing amounts of memory and time to store and retrieve the growing number of precedents. CASEY's memory requirements can be measured by the number of cases stored in the case base and the number of GENs beneath which these are indexed. Its search time is bounded by the number of cases and GENs that it explores during precedent retrieval. The number of modifications the system must make to precedent solutions is an additional component of time complexity.

For CASEY to be worthwhile, it must supersede Heart Failure in some measure. Although CASEY can, in general, only emulate Heart Failure's accuracy,[3] it can surpass its efficiency, presumably by avoiding the computationally intensive model-based problem solving.

[3] Occasionally, CASEY may find a precedent whose solution is a more likely explanation of the new case findings than the Heart Failure ideal; however, it is *unlikely* that the model-based system would construct this more accurate hypothesis for the precedent but overlook it for the new case. Alternately, CASEY could overtake the accuracy of the original Heart Failure system by obtaining less fallible precedents either from physicians or from a less heuristic, more robust, more time consuming version of the model-based system.

4.2 Experimental Method

In 1867, Louis Pasteur introduced to scientific thought the notion of a controlled experiment. Bacteria flourished in an open flask he filled with unfermented liquids but failed to grow in a similar, sealed container. Before his breakthrough, scientists believed that bacteria were spontaneously generated from food and liquids under all circumstances.

In 1990, many Artificial Intelligence researchers continue to conduct evaluations without any semblance of experimental control. Studies that investigate the viability of a new method do not test different circumstances across which its performance may vary. However, in this examination of case-based reasoning, I use experience itself as an experimental variable: rather than merely evaluating reasoning from experience *per se*, I determine the effects of *increasing* experience on both accuracy and efficiency. My results are still inevitably limited, since they are derived from only a single domain. However, by uncovering and making explicit characteristics of heart failure that influence CASEY's performance, through statistical analysis of the test cases, I will be able to generalize my conclusions to encompass similarly characterized domains.

4.3 Experimental Design

This study measured various facets of CASEY's performance and their dependence upon experience, using a pool of 240 cases provided by Tufts New England Medical Center. To modulate the amount of experience, I composed varying size collections of precedents, including 25, 50, 75, 100, 125, 150, 175, and 200 patients respectively. Each case base contained 25 new precedents, in addition to *all* the precedents contained in the previous case base, so that experience was both uniformly incremental and cumulative.

Employing each case base in turn, CASEY diagnosed the 40 remaining patients, which were designated as "test cases." Each such diagnosis was subsequently evaluated according to the following criteria:

- ϵ, the diagnostic error of CASEY's causal explanation, with respect to the Heart Failure ideal (see section 4.1)

- ϵ_P, the raw precedent error, representing the diagnostic error of the precedent causal explanation *before* solution adaptation, with respect to the Heart Failure ideal

- λ, the relative likelihood of CASEY's causal explanation for the findings, with respect to the Heart Failure ideal (see section 4.1)[4]

- the number of memory nodes (GENs and precedents) searched during precedent retrieval

In addition, each case base was assessed for:

- the number of memory nodes (GENs and precedents) it contained

[4] Whenever CASEY found no appropriate precedents for a particular test case, ϵ was assigned the default value of 1, because CASEY's null solution was, by definition, missing every ideal state. Since the likelihood of a null solution is 0, λ was assigned this default value. The default value of ϵ_P, however, was calculated using the best matched precedent retrieved, even though CASEY could not transfer its solution to the test case.

For the purpose of computing the relative accuracy criteria, "ideal" diagnoses for all 40 test cases were provided by the Heart Failure system. All 200 precedents were also diagnosed by Heart Failure, and to maintain consistency, CASEY's causal explanations for test cases were not incorporated dynamically into the case bases, as they might have been otherwise. Thus, the capabilities of case-based reasoning were isolated from the quality of the precedent solutions.

The sample of 240 patients is reasonably and convincingly representative of the real-world distribution of heart failure cases: it includes all patients diagnosed at Tufts New England Medical Center, from all DRGs related to heart failure, during a period of about two years. Thus, the sample contains no bias, as did the 45 cases used to test CASEY initially; indeed, 20 of those 45 were hand selected from patients with either coronary artery disease or aortic stenosis [14]. Each case base incorporates, in order, the first N cases from a prespecified random permutation on the 200 potential precedents. The random permutation eliminated the skew intrinsic to the initial sequence, which organized the patients by DRG. Moreover, enforcing the same permutation over all case bases factored the issue of learning order out of the study.

The 40 test cases had also been selected at random from the pool, and to ascertain whether they represent a valid statistical sample, I constructed four additional test sets, each also containing 40 randomly chosen cases, in order to measure variances. With a case base of 25 precedents, CASEY diagnosed the 40 test cases from each of the five sets, and I computed, for each set, averages of the relative accuracy criteria and an average of the number of memory nodes searched.

Figure 4.1 presents the spread of the five averages along each criterion. The scale of each plot represents the natural range of the measure it quantifies (see figures 5.4, 5.5, 5.6, and 6.5), and because the variances extend over relatively small portions of these ranges, we can conclude that 40 test cases are a sufficient predictor of average performance. The boxes contain 50% of the spread, or three set averages, while the fences encompass the remaining averages, with the exception of extreme outlyers, denoted by asterisks.

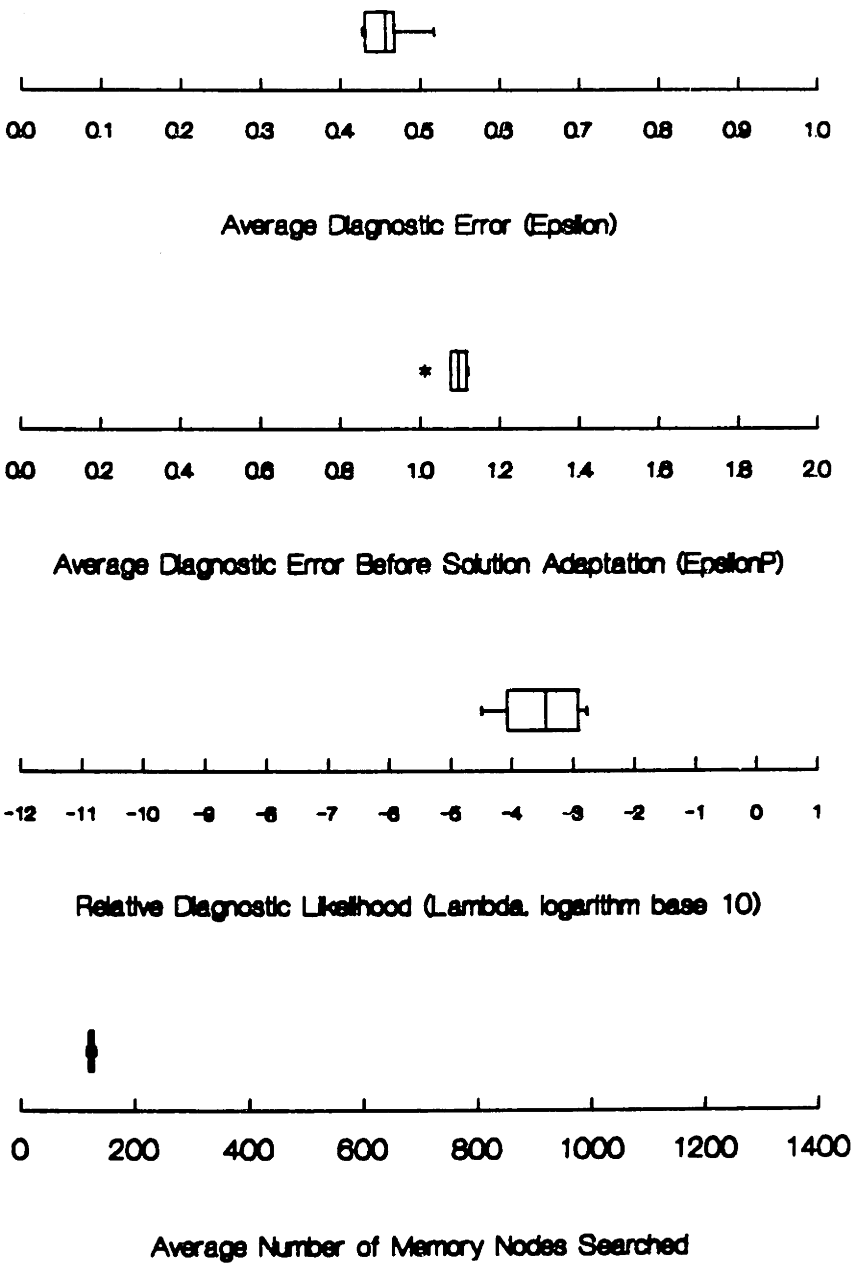


Figure 4.1: Variances for four performance criteria

Chapter 5

Accuracy

5.1 Predictions

Even now that we have a basis for evaluating CASEY, we must postpone asking "How well does CASEY work?" until we have answered "Why does CASEY work?", or more simply: "Does CASEY work?". Fundamentally, we should investigate whether the heart failure domain is a suitable application for case-based reasoning. For a number of reasons, I hypothesized that it is.

First, not every combination of findings represents a plausible case. Because these findings are merely the outward manifestations of particular disease states, they occur not randomly but causally, appearing often in the company of certain related symptoms and almost never in the company of others. Furthermore, the primary causes (and multiple disease sets, as well) have varying likelihoods *a priori*. Thus, many of even the plausible cases will be rare, while a few common syndromes will appear time after time. For these "routine" cases, which are likely to begin with, history will repeat, and similar versions will recur. Accordingly, case-based reasoning does well to remember their solutions, which are destined to be all but worn out with future use. The method also relies on the converse tautology: cases for which it has most likely not encountered precedents are—just that!—rare.

Winston perceived, as early as 1982, that these "Causal relations identify the regularities that can be exploited from past experience" [25]. Invariably, the same causal mechanisms that explain a particular case will entail similar explanations for slight variations on the case. Rarely does hypothesizing a disease depend upon the presence or absence of a single symptom. Relationships between states in the Heart Failure model are expressed locally, as links, and therefore small differences will necessitate only local modifications. However, as the differences grow, the entire backbone of a diagnosis may be eroded, no longer supported by the same constellation of symptoms. Accordingly, although the new set of findings may be explained plausibly by the precedent causal explanation, they may be more probably explained by a completely different pattern of pathophysiologic states. Indeed, with increased diagnostic confidence comes the subtle, increased danger of transforming an unsuitable precedent solution into a technically possible, but improbable causal explanation for the new case.

Because of the primacy of obtaining an appropriate precedent, the case memory will be unequivocally CASEY's primary "source of power." However, I expect that CASEY's extensive use of the Heart Failure model has significant effects on its performance. For example, indexing and evaluating precedents on the basis of generalized causal features is extremely clever, as it allows

CASEY to preview potential components of the new case solution and match them with components of precedent causal explanations. Clearly, the system has much to gain and little to lose from these local, but powerful operations. In terms of generalized causal features, similar problems will recur even more frequently and will harbor even fewer differences between their solutions.

CASEY's solution transfer process is probably the implementation's largest contribution to performance. Through local application of the Heart Failure model, CASEY can make a close solution perfect, a moderately appropriate solution close, and a distant solution plausible. The very facility of these modifications allows CASEY to search for broad equivalence classes of diagnostic hypotheses, rather than particular hypotheses themselves, thus enhancing the generality and applicability of its case-based reasoning.

Given that similar problems recur, CASEY can become proficient at "routine" diagnosis without an inordinate number of precedents. As the number of precedents increases, the best matched previous case will come closer and closer to the new case. Moreover, if similar problems require similar solutions, the error measures ϵ, ϵ_P, and λ should improve rapidly as the case base grows, particularly as the decreasing difference between the new case and its precedent increases the locality of potential modifications.

Unfortunately, these intuitions were fundamentally misleading.

5.2 Accuracy and Experience

In actuality, CASEY's accuracy does not improve with experience. Figures 5.1, 5.2, and 5.3 show three criteria of accuracy unchanging as experience increases from 25 to 200 precedents. Each point represents an average over the 40 test cases, diagnosed using a case base of the specified size. Given the variances illustrated in figure 4.1, the small changes that these graphs experience are insignificant.

In figure 5.1, the diagnostic error ϵ, which was expected to asymptote towards zero, hovers around 0.5. Thus, even with 200 precedents to reason from, CASEY constructs causal explanations in which, on average, half of the pathophysiologic states are either missing or spurious. According to this criterion, case-based reasoning is both ineffective and unimproving.

Figure 5.2 plots the relative diagnostic likelihood λ, which should asymptote towards 1 if experience were successful. λ is plotted on a logarithmic scale and averaged logarithmically as well, because it is calculated as a product, using conditional probabilities. These conditional probabilities, embodied in the causal links of a diagnosis, represent the chance that a particular finding or intermediate state is caused by the pathophysiologic state used to explain it. Because each link contributes multiplicatively to the likelihood, the measure is extremely sensitive to change. Intuitively though, an increasing number of precedents should improve CASEY's chances of obtaining a high probability solution, or at least, a solution closer to Heart Failure's. Instead, CASEY's causal explanations for test case findings remain approximately 1000 times less likely that the explanations constructed by Heart Failure for the same cases.

How crucial or incidental to CASEY's reasoning is the solution adaptation process? During this process, CASEY uses the Heart Failure model to reduce the difference between a precedent causal explanation and the ideal solution for the new case. To isolate the contribution of the underlying case-based reasoning, we can ablate solution adaptation, computing the raw precedent error ϵ_P as the difference between the pathophysiologic states from the most appropriate precedent, *before transfer*, and those of the ideal solution for the new case.

Figure 5.3 displays both the raw error ϵ_P before solution adaptation (dotted line) and the error ϵ after solution adaptation (solid line). Obviously, adaptation is significant, as it allows CASEY on average to reduce error by half. The difference $|\epsilon_P - \epsilon|$ approximates the number of modifications that CASEY makes during adaptation, in proportion to the size of the solution itself; since the average causal explanation contains roughly 23.5 pathophysiologic states, CASEY actually changes half of these, or 11.75 states. Even more significant, however, is the fact that the error ϵ_P before solution adaptation does not decrease with increasing case base size. CASEY simply does not find better precedents, even when its choices grow substantially.

Almost as surprising as the fact that accuracy does not improve are the occasional *increases* in relative error and the occasional *decreases* in relative likelihood. Undoubtedly, these retrogressions are miniscule, but the reader may wonder how they could occur *at all.* Once CASEY solves a test case using a particular precedent, achieving a certain level of accuracy, then adding more patients to the case base should not compromise this achievement, because the system always has the option of reverting to the original precedent. This argument is misleading, however, because it assumes that the most appropriate precedent will also bear the most resemblance to the test case. Since CASEY does not yet know the diagnosis of the new case, it cannot judge precedents according to the suitability of their causal explanations. Instead, it must rely on a more fallible similarity metric based on findings or generalized causal features. Consequently, the system may retrieve from the larger case base a *less* appropriate precedent which it deems *more* similar to the new case.

5.3 Accuracy Range

Although average accuracy remains constant with increasing case base size, accuracy results vary considerably over individual test cases. Figures 5.4, 5.5, and 5.6 display histograms for the three accuracy criteria over 320 tests, which comprise the diagnoses of all 40 test cases by each of the 8 case bases.

Figure 5.4 reveals that the diagnostic error ϵ resides primarily between 0 and 1. However, several outlying points testify that it can exceed 1 when the number of spurious states overwhelms the number of correct pathophysiologic states.

In figure 5.5, the raw diagnostic error ϵ_P before solution adaptation sprawls predominantly between 0.2 and 1.8. Because CASEY has not yet eliminated flagrantly spurious states from or incorporated blatantly missing states into the precedent causal explanation, the number of incorrect states overwhelms the number of accurate states as often as not. Five extreme outlying points at $\epsilon_P = 5.0$ and one at $\epsilon_P = 5.3$ did not even fit within the graph boundaries.

Figure 5.6 displays the relative diagnostic likelihood λ ranging from 10^{-12} to 10^1. Only twice in 320 tests did CASEY arrive at a causal explanation more likely than Heart Failure's. Otherwise, many solutions were plausible, while a few were abysmal, at nearly one trillion times less likely than Heart Failure's diagnosis.

5.4 Precedents and Accuracy

Figures 5.7, 5.8, and 5.9 plot the average number of precedents that were available when CASEY obtained solutions with the specified accuracy. The lines represent locally weighted scatterplot smoothings. Intuitively, improved accuracy values should imply that larger case bases were used to obtain them. However, because accuracy remains constant as case base size varies, this supposition does not hold. If anything, the number of precedents used to support increasing accuracy *decreases*, but the points are too scattered and uncorrelated to support conclusions of any sort.

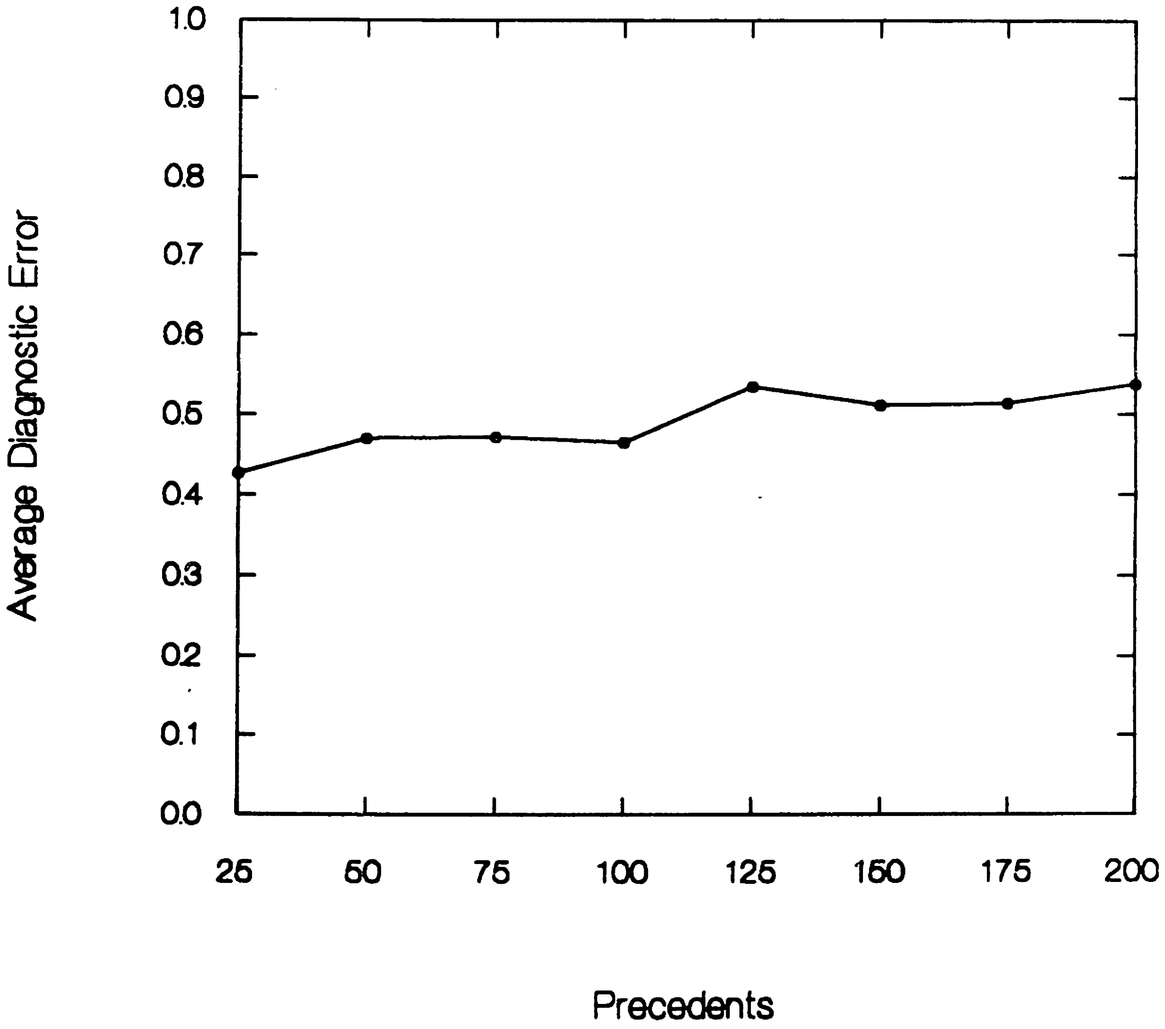


Figure 5.1: Effects of increasing experience on the average diagnostic error ϵ

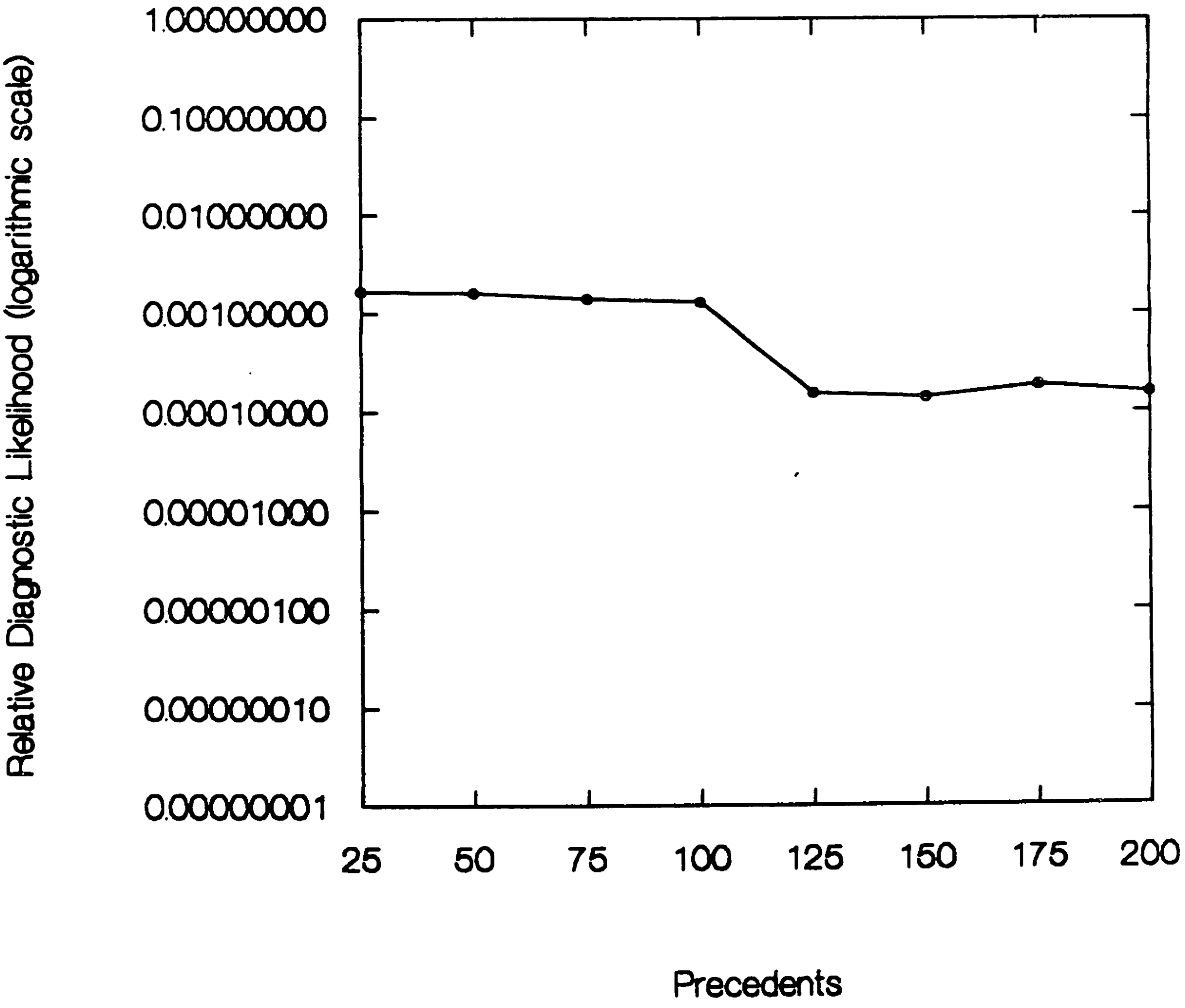


Figure 5.2: Effects of increasing experience on the relative diagnostic likelihood λ

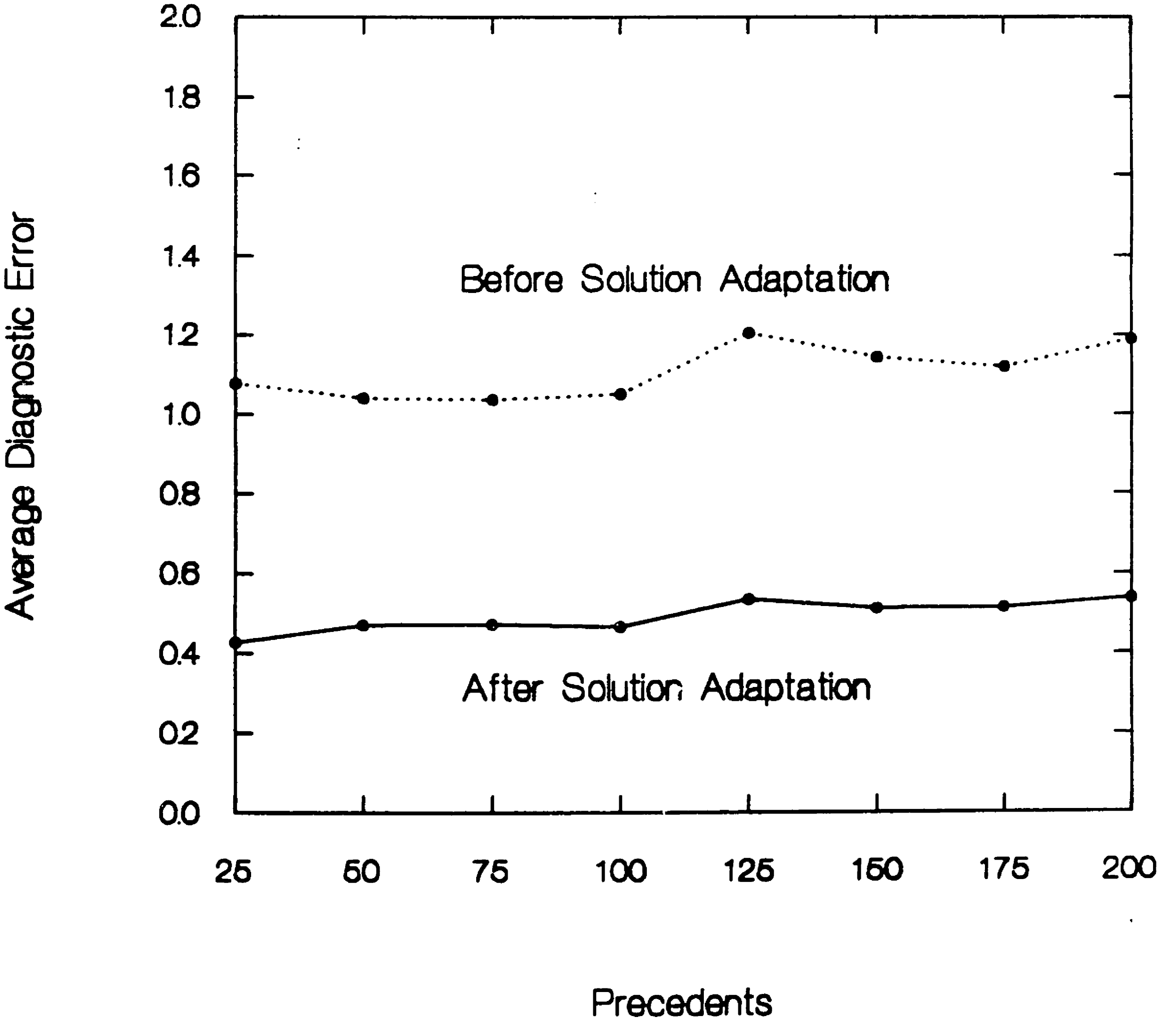


Figure 5.3: Effects of increasing experience on the average diagnostic error before (ϵ_P) and after (ϵ) solution adaptation

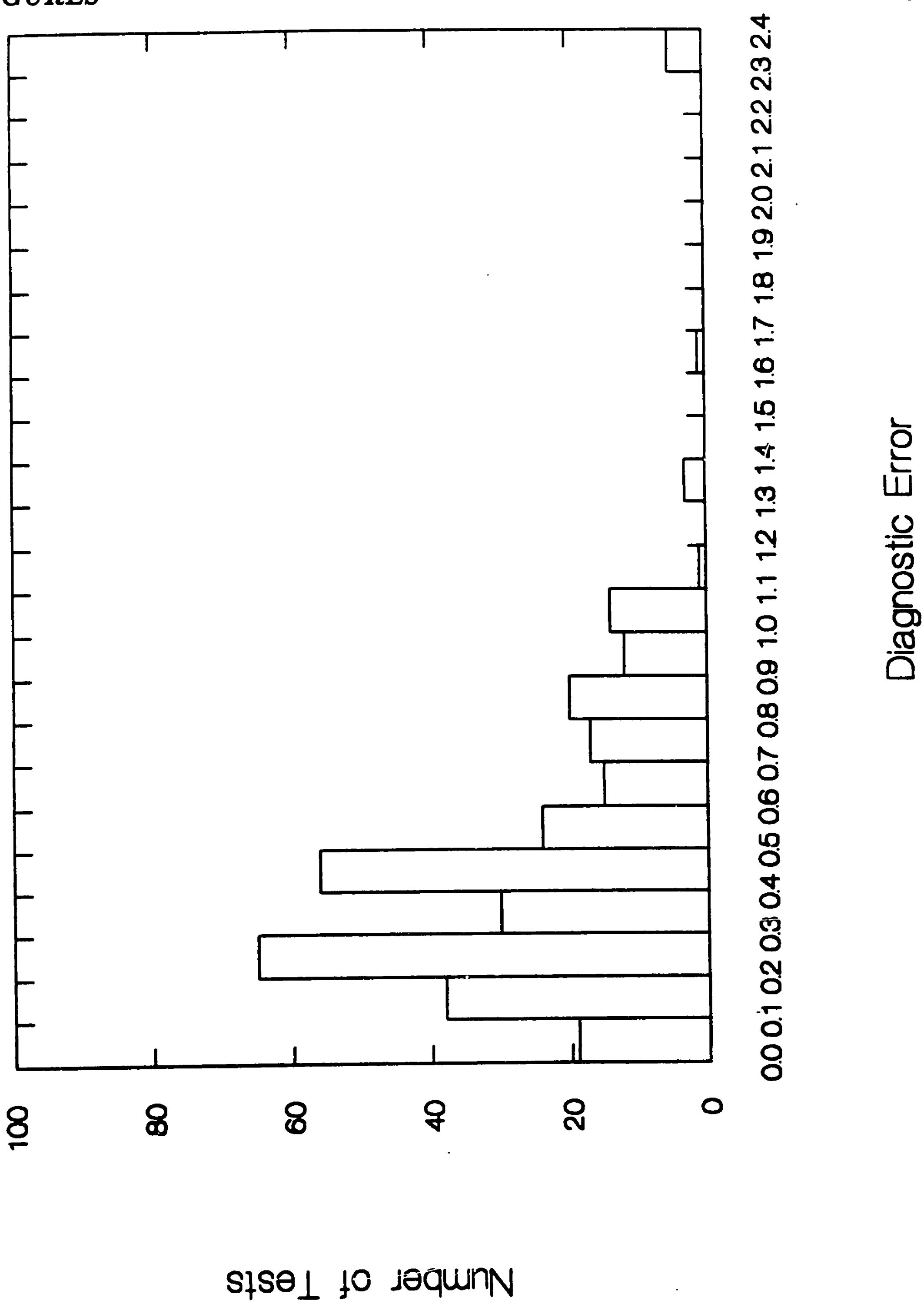


Figure 5.4: Range of the diagnostic error ϵ

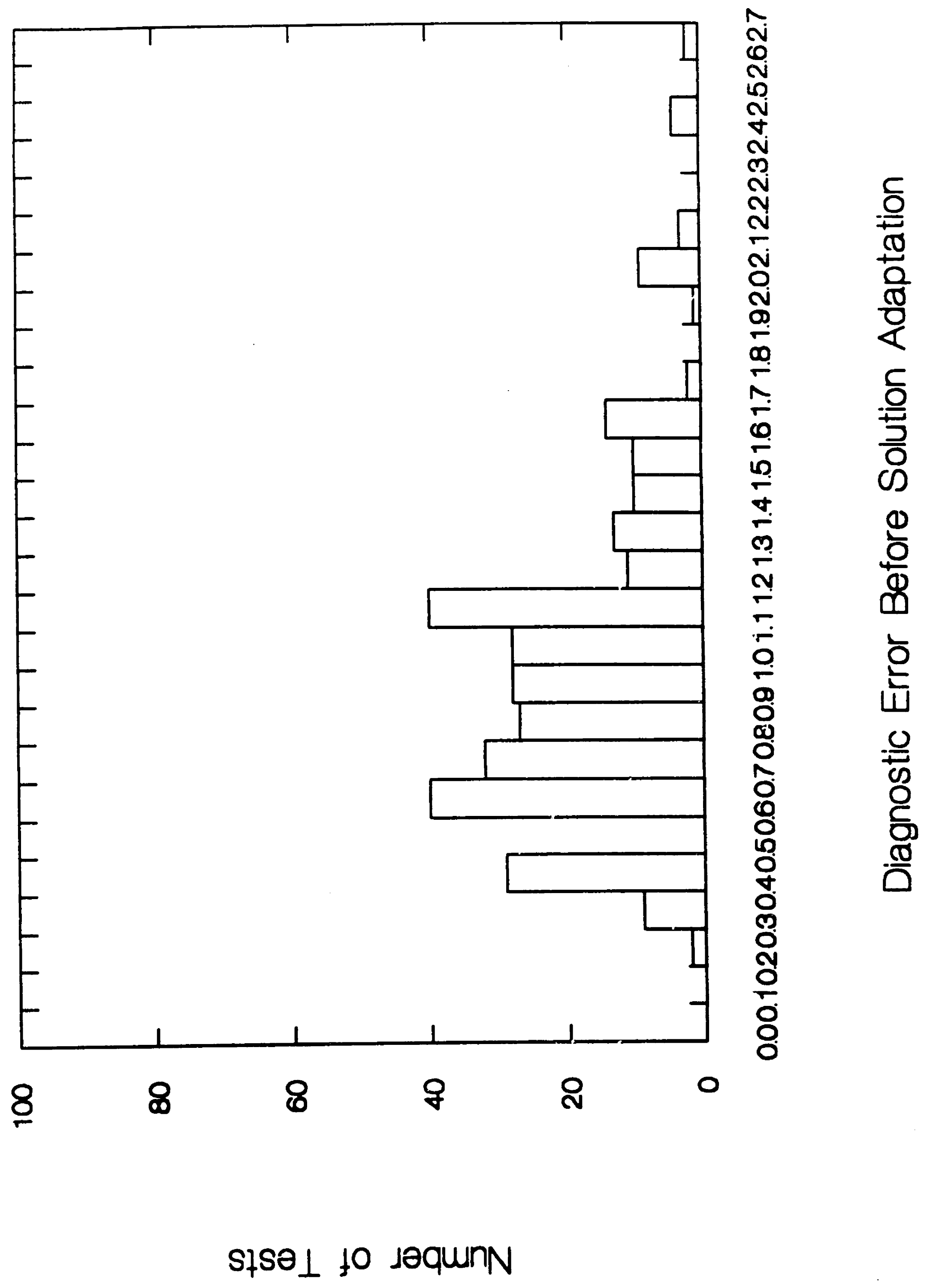


Figure 5.5: Range of the diagnostic error ϵ_P before solution adaptation

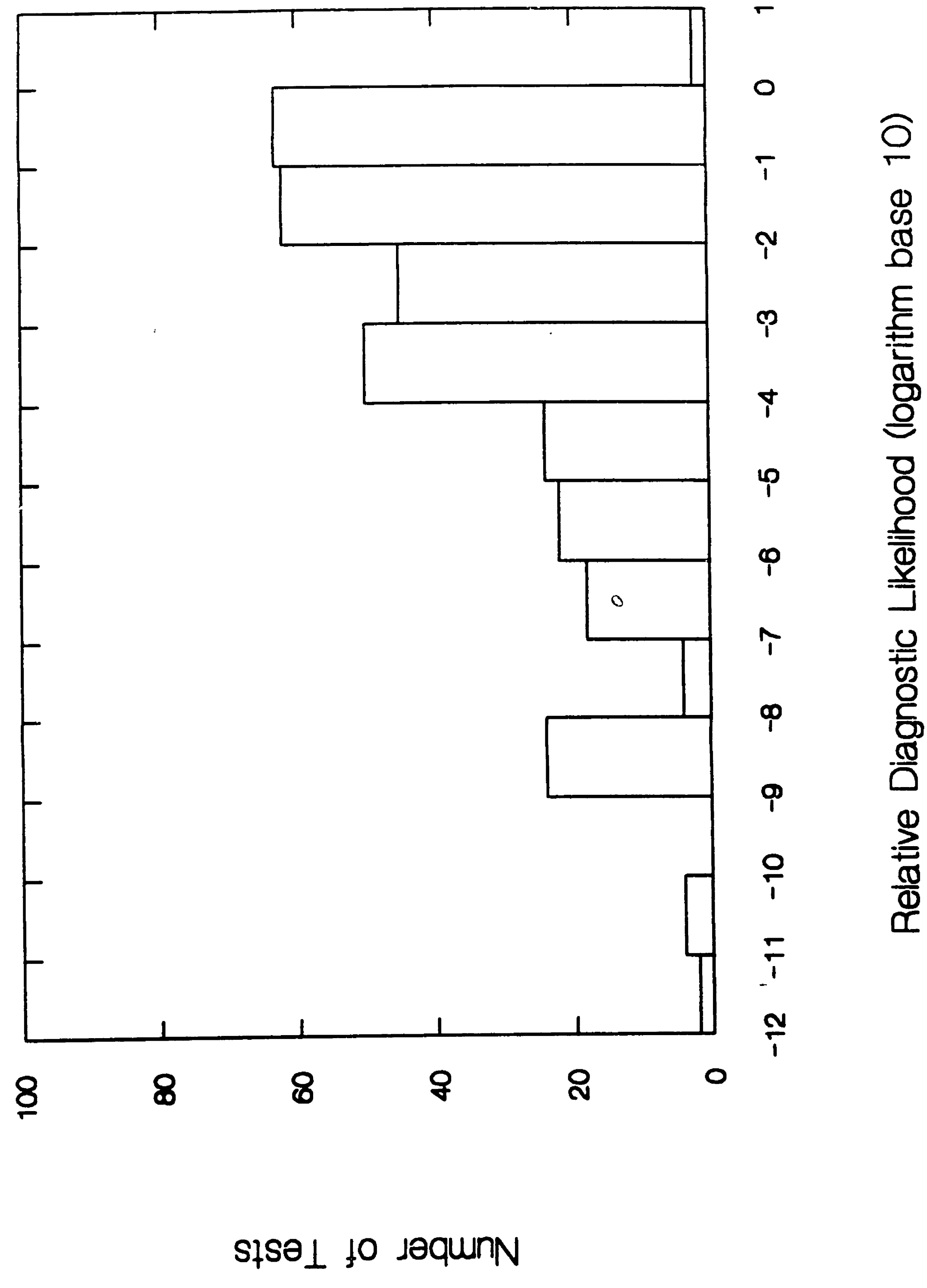


Figure 5.6: Range of the relative diagnostic likelihood λ

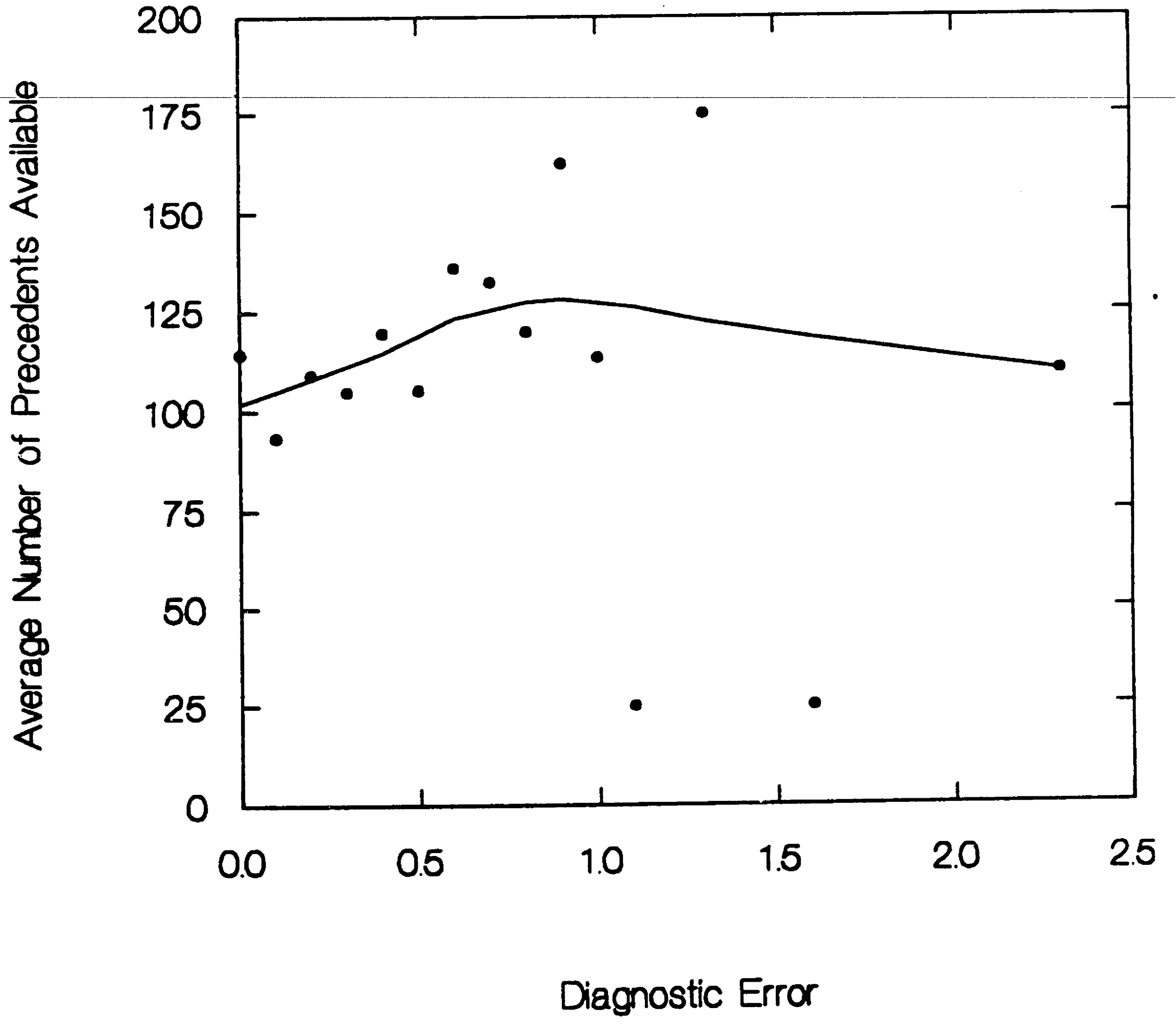


Figure 5.7: Correlation between the diagnostic error ϵ and the average number of precedents available

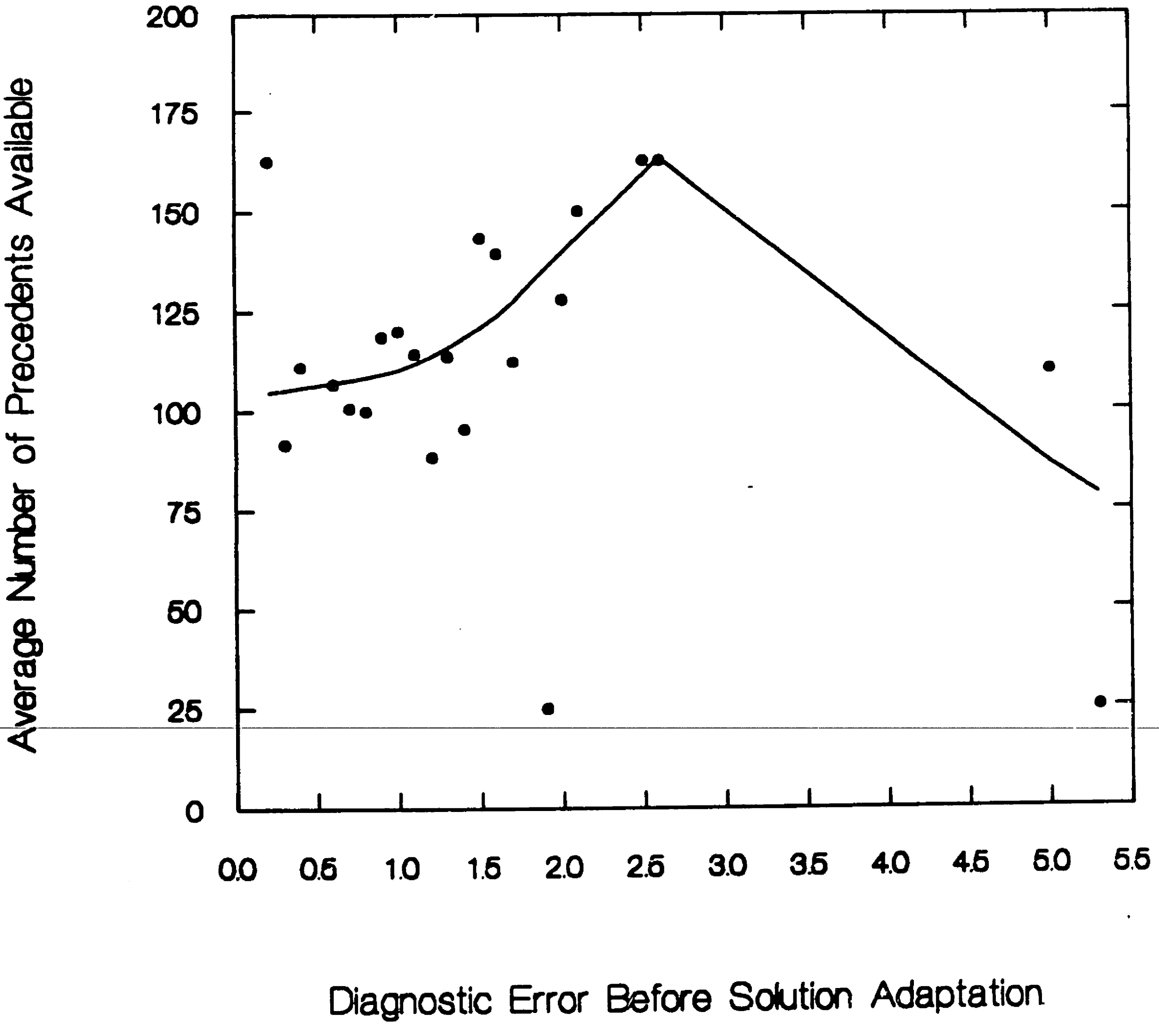


Figure 5.8: Correlation between the diagnostic error ϵ_P before solution adaptation and the average number of precedents available

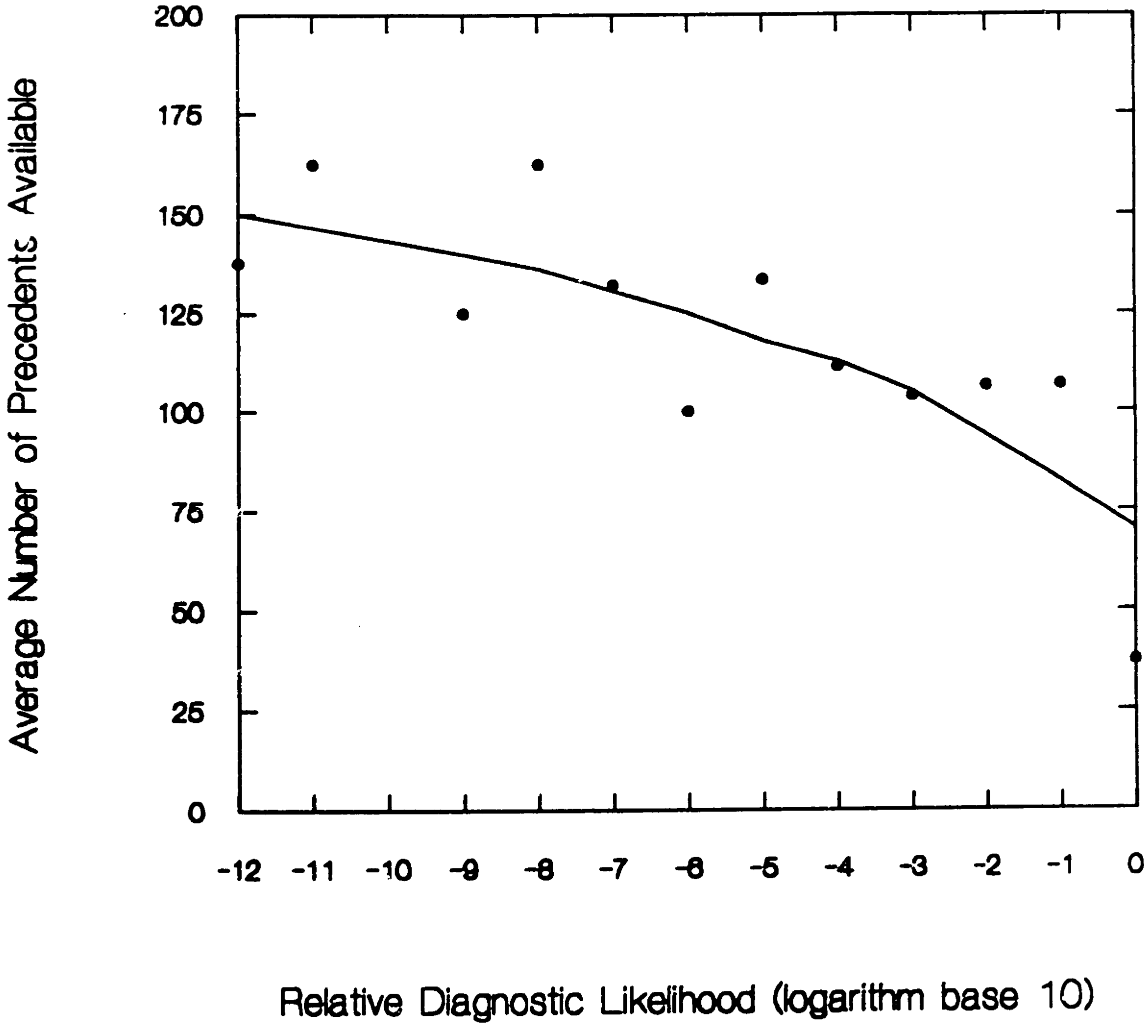


Figure 5.9: Correlation between the relative diagnostic likelihood λ and the average number of precedents available

Chapter 6

Efficiency

6.1 Predictions

Although CASEY's accuracy is primarily domain determined, its efficiency depends heavily upon implementation. In particular, CASEY's memory requirements may be significantly larger than the number of precedents might suggest. Kolodner's construction (the basis for CASEY's memory) is both complex and redundant. The number of GENs it creates may equal and even overwhelm the number of actual cases. Recall that in CASEY a case is indexed, at each stage of the discrimination network, according to *every* finding that differentiates it from the norm.[1]

Because indexing is redundant, the size of the case base has the potential to grow exponentially with the number of cases. Hypothetically, there could be $O(2^N)$ GENs , each representing a different subset of the N precedents and

[1] Recall also that this redundancy is essential, as the most appropriate precedent may match the new case according to any of the combinatorially many subsets of possible findings.

defined by the unique group of findings common to that subset. Granted, cases that are prototypical for a GEN are not stored, and features included in the GEN norms are eliminated from further indexing. However, the intuition that the case memory grows *more* slowly after it has seen increasing numbers of cases [14] is not substantiated empirically or formally.

Retrieval time may also slow, despite the fact that CASEY never searches all paths through the discrimination net. At each stage of the network, the search must branch according to the number of differentiating features, which may decrease, but not dramatically. Although retrieval time seems to be constant for up to 44 cases [14], this relation may change as the case base becomes truly large.

To promote compactness, CASEY merges a GEN with its parent when it accumulates 2/3 of its parent's cases. Thus, every GEN's "case count" is at most 2/3 times the "case count" of its parent, and consequently, the memory structure is no deeper than $O(\log_{3/2} N) = O(\log N)$. As long as the precedent search branches at each stage of the discrimination net, it will visit $O(b^{\log N}) = O(N)$ nodes (where b is the branching factor). This does *not* imply that the search visits a constant fraction of the cases, however, as a case indexed into any parent GEN may be indexed, recursively, into several of its children, and may thus be visited several times in the course of a search.

The preceding estimates of exponential space and linear time are the results of a worst-case analysis. Undoubtedly, there are domains that elicit this notorious performance, and in such cases, even an $O(N)$ time brute force search through the simple $O(N)$ size precedent list would be preferable. However, the regularity intrinsic to the heart failure domain should simplify the memory's complexity in practice.

Because similar cases recur, they will cluster compactly into larger and fewer GENs. Moreover, clusters of co-occurring findings will be roughly disjoint, causing the rapidly diverging branching factor to converge on relatively few sub-generalizations. In the best case, the discrimination net will resemble a tree, expending only $O(N)$ space; if only a small fraction of the paths through this structure are actually searched, retrieval time may approach $O(\log N)$. Although CASEY does fall short of this ideal, I expected its memory to grow only as a small polynomial in the number of precedents and its time complexity to be sublinear.

6.2 Memory

An exceptionally smooth curve in Figure 6.1 outlines the number of memory nodes, including both GENs and cases, required to index precedent populations of varying size. Figure 6.2 plots the same points on a double logarithmic scale, and its perfect linearity demonstrates that CASEY's memory grows polynomially as predicted, and not exponentially.[2] A subsequent linear regression analysis using the simplex algorithm revealed the degree of the curve to be 1.668. However, although its order of growth is small, CASEY's memory grows rapidly enough to become infeasible even when the number of precedents is moderate. Almost 9000 memory nodes are required to index only 200 precedents.

As explained in the previous section, CASEY collapses a GEN into its

[2] If $\log Y = a \log X + b$, then $10^{\log Y} = 10^{a \log X} 10^{b}$, and thus $Y = 10^{b} X^{a}$, a polynomial.

parent when it subindexes 2/3 of its parent's cases. Figure 6.3 illustrates the consequences of failing to collapse such GENs. The solid line in the figure plots the number of memory nodes required when collapsing is performed; the dotted line represents memory growth without collapsing. Certainly, the process is significant, as it reduces the number of nodes required by approximately half. However, it does not diminish the memory's order of growth.

Section 3.2 discussed the fact that CASEY maintains two discrimination networks, one that indexes precedents based on their findings, and another that indexes them based on generalized causal features. Figure 6.4 depicts the breakdown of memory growth between the two networks. Since generalized causal features abstract findings to the level of the pathophysiologic states for which they provide evidence, they provide a superior representation for the regularity inherent in the cases. Consequently, the dotted line representing the generalized causal memory rises more slowly than does the solid line, which represents the memory based on findings *per se*. Clearly, the added abstraction facilitates more parsimonious, more compact generalization.

6.3 Retrieval Time

Figure 6.5 shows retrieval time fulfilling the worst case scenario of linear growth. Each point represents the number of memory nodes searched during precedent retrieval, averaged over the 40 test cases, for each case base. Considering the slope of the plot renders CASEY's retrieval process not only disappointing but acutely impractical as well. For instance, the system searched over 1200 memory nodes in a case base of only 200 precedents.

In figure 6.6, retrieval time is broken down between the findings memory (solid line) and the generalized causal memory (dotted line). Because a new case is, by definition, undiagnosed, its findings are not yet explained by pathophysiologic states. Consequently, CASEY must include among its generalized causal features *all* possible causes for all of its findings. Thus, a search through the generalized causal memory will branch according to a more extensive sequence of indices and will therefore visit a larger number of nodes.

How does retrieval time grow as accuracy increases? Intuitively, the case bases used to achieve higher accuracy values should be larger, and thus should require more time for retrieval. Also, cases for which good precedents exist are presumably well represented in the case base and, thus, must search through a relatively large segment of it. Conversely, rare cases will find few sub-generalizations with similar distinguishing features, and thus can quickly locate their best matches, which are actually quite poor.

Figures 6.7, 6.8, and 6.9 are scatterplots attempting to correlate retrieval time with the three accuracy criteria. Each point corresponds to a single test case, while each test case appears at eight points, representing different retrieval times and different diagnostic accuracies obtained from each of the eight case bases. Although curves were obtained using locally weighted scatterplot smoothing, the plots are sufficiently diffuse to warrant forbearing any unsubstantiated conclusions we might draw from them. Only in two of the graphs does retrieval time seem to increase with decreasing diagnostic error, but it does so in sections of the graphs that are almost entirely unpopulated. As the average number of precedents available did not increase with accuracy, we cannot expect retrieval time, a significantly more indirect measure of relevant experience, to increase.

6.4 Generalization Granularity

Because GEN norms are compiled from features common to at least 2/3 of the subsumed cases, CASEY merges a sub-generalization with its parent when it indexes the same fraction of its parent's precedents. How is CASEY's performance affected by changing the granularity of generalization through this fraction's value? To answer this question, I conducted an additional series of experiments, holding the number of precedents constant at 100, but allowing the granularity parameter to vary among the values 1/2, 2/3, 3/4, and 9/10.

Intuitively, decreasing this parameter yields larger, vaguer GENs. Thus, although the memory will be more compact and the retrieval time will decrease, I predicted that the quality of CASEY's solutions, as measured by ϵ, ϵ_P and λ, would diminish. Conversely, as the granularity parameter increases, GENs will become smaller and more specific. Consequently, the memory will be more discriminating, and unfortunately, more expansive, suggesting that retrieval time will increase and that solution quality will improve.

Figures 6.10 through 6.13 confirm increases in memory size and retrieval time with larger granularity values. Memory size experiences the slight increase shown in figure 6.10 because larger granularity causes GEN collapsing to occur less frequently. Figure 6.11 demonstrates that the findings memory (solid line) and the generalized causal memory (dotted line) are equally affected. Retrieval time increases, as shown in figure 6.12, because stricter requirements for including features in GEN norms leave more features available for branching; in addition, there are more GENs to search. Figure 6.13 portrays the more pronounced effects on the generalized causal memory's retrieval time (dotted line), encouraged by that memory's greater tendency to branch during search.

Figures 6.14, 6.15, and 6.16 plot the three accuracy criteria against generalization granularity.[3] Not surprisingly, since accuracy does not vary with case base size, it is "marble-constant" across all granularity values.

[3] A newer version of the Heart Failure system computed the likelihood of CASEY's solutions for this experiment. Its stricter "probability of hypothesis" function tended to produce lower λ values, relative to the precomputed ideal likelihoods. However, the increased magnitude of the disparity does not change the criterion's constancy over varying generalisation granularity.

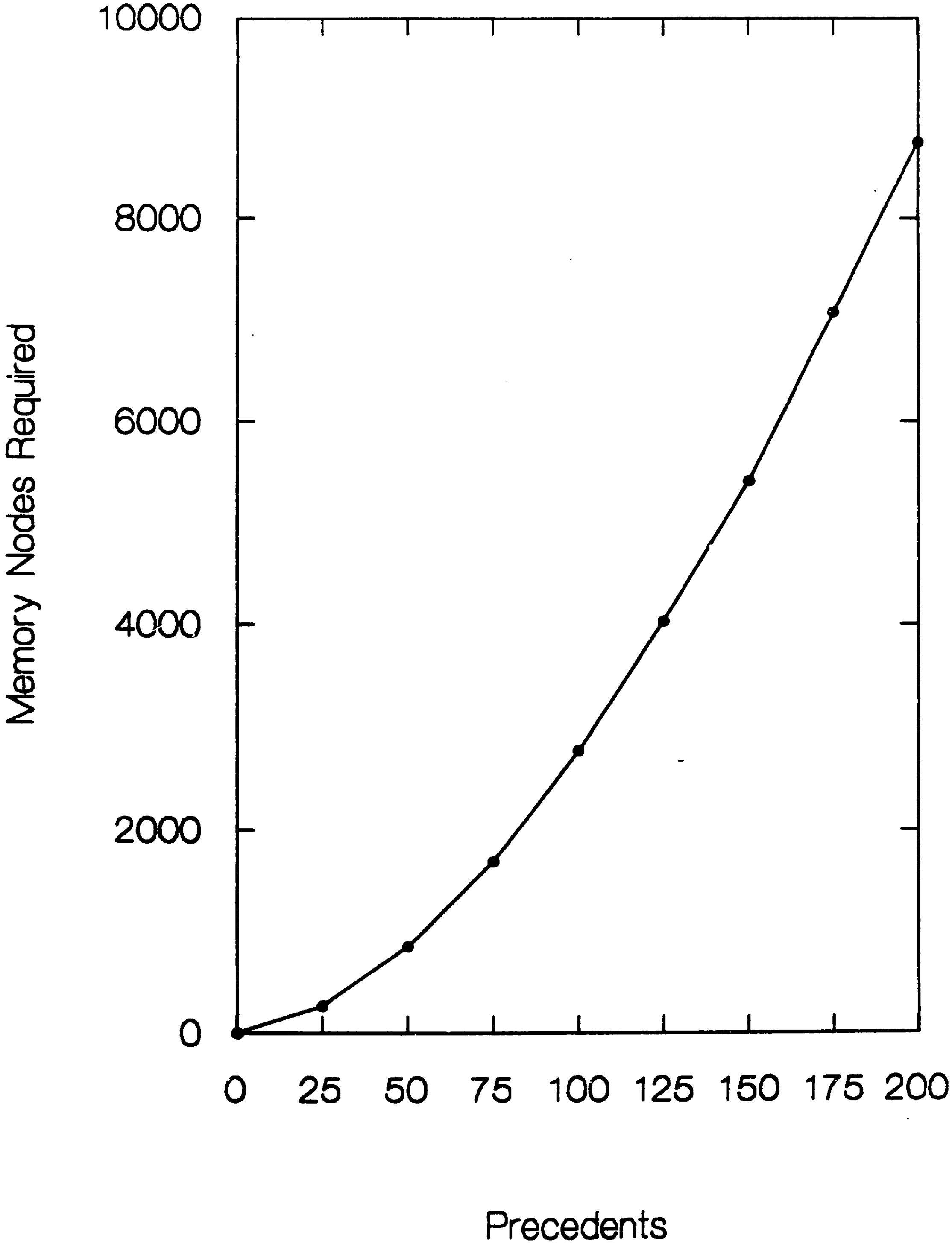


Figure 6.1: Growth of CASEY's memory

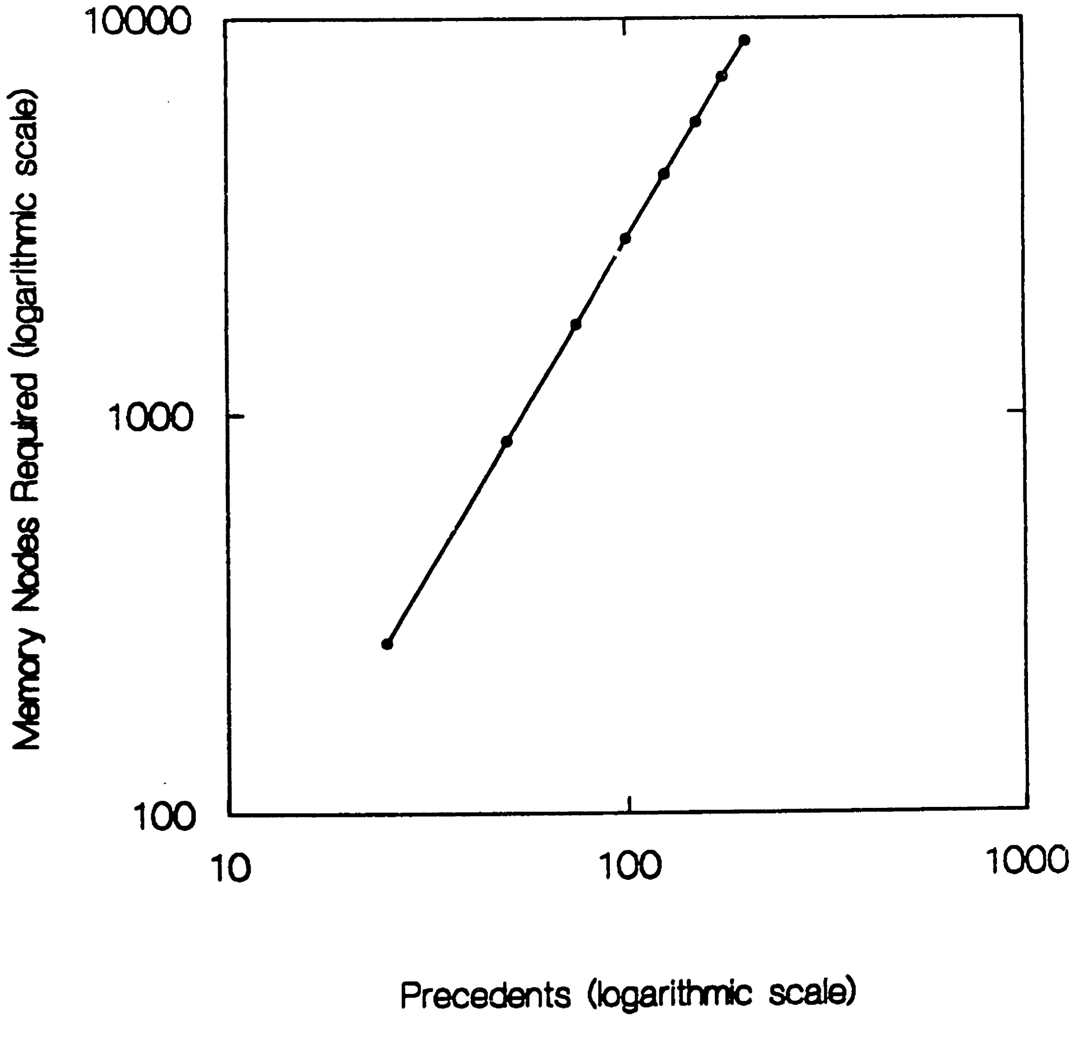


Figure 6.2: Log × log plot of memory growth

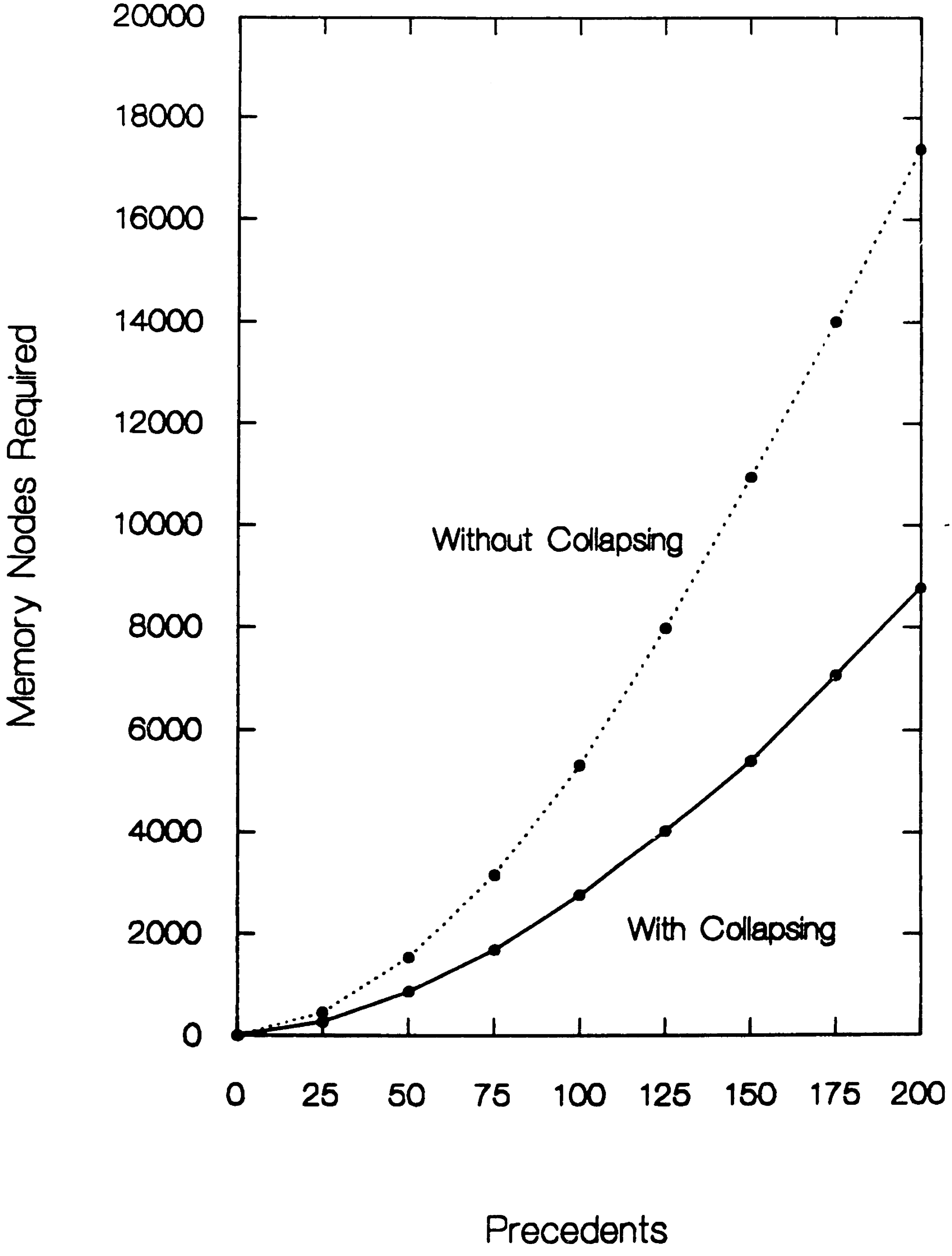


Figure 6.3: Memory growth, with and without GEN collapsing

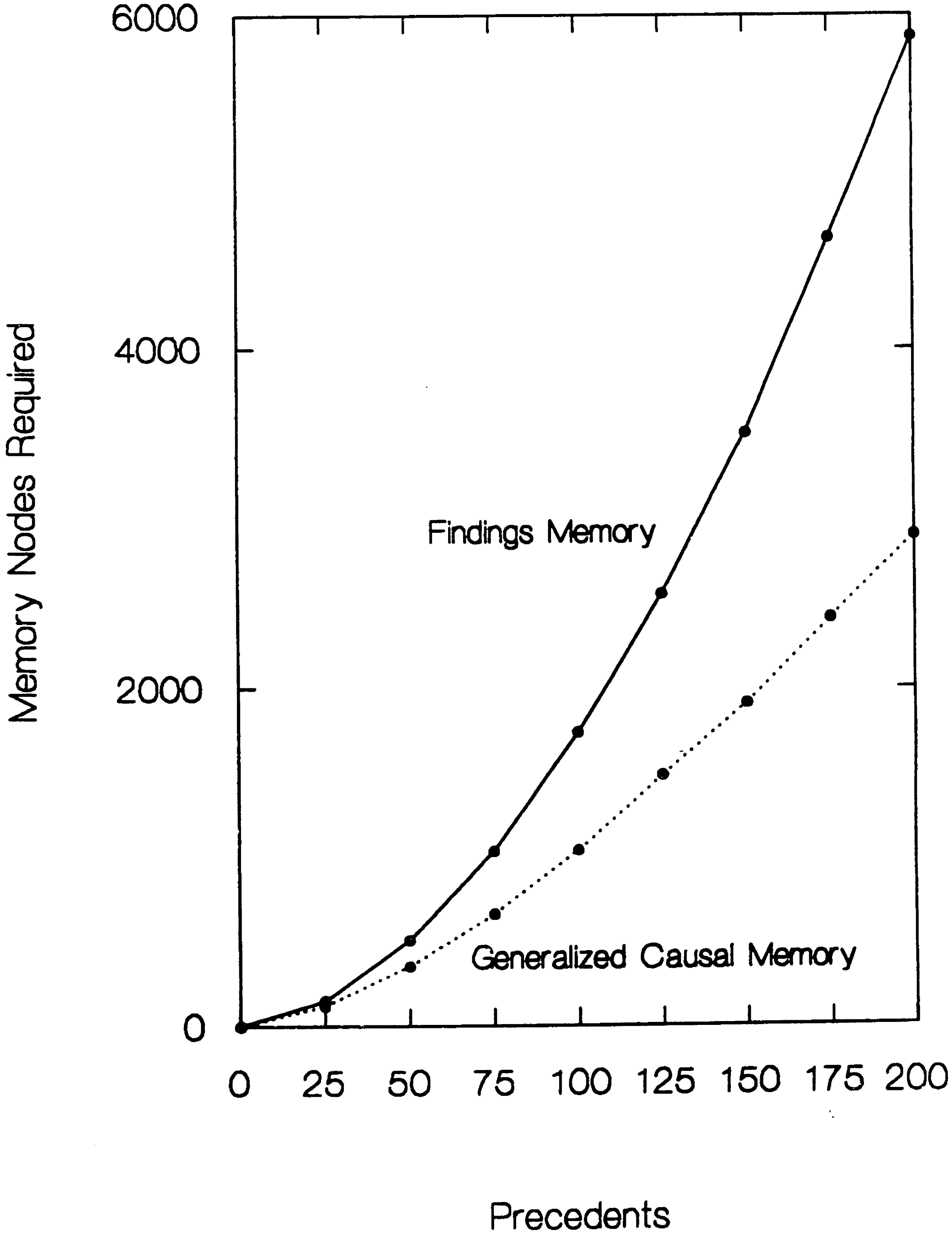


Figure 6.4: Memory growth, broken down between the findings memory and the generalized causal memory

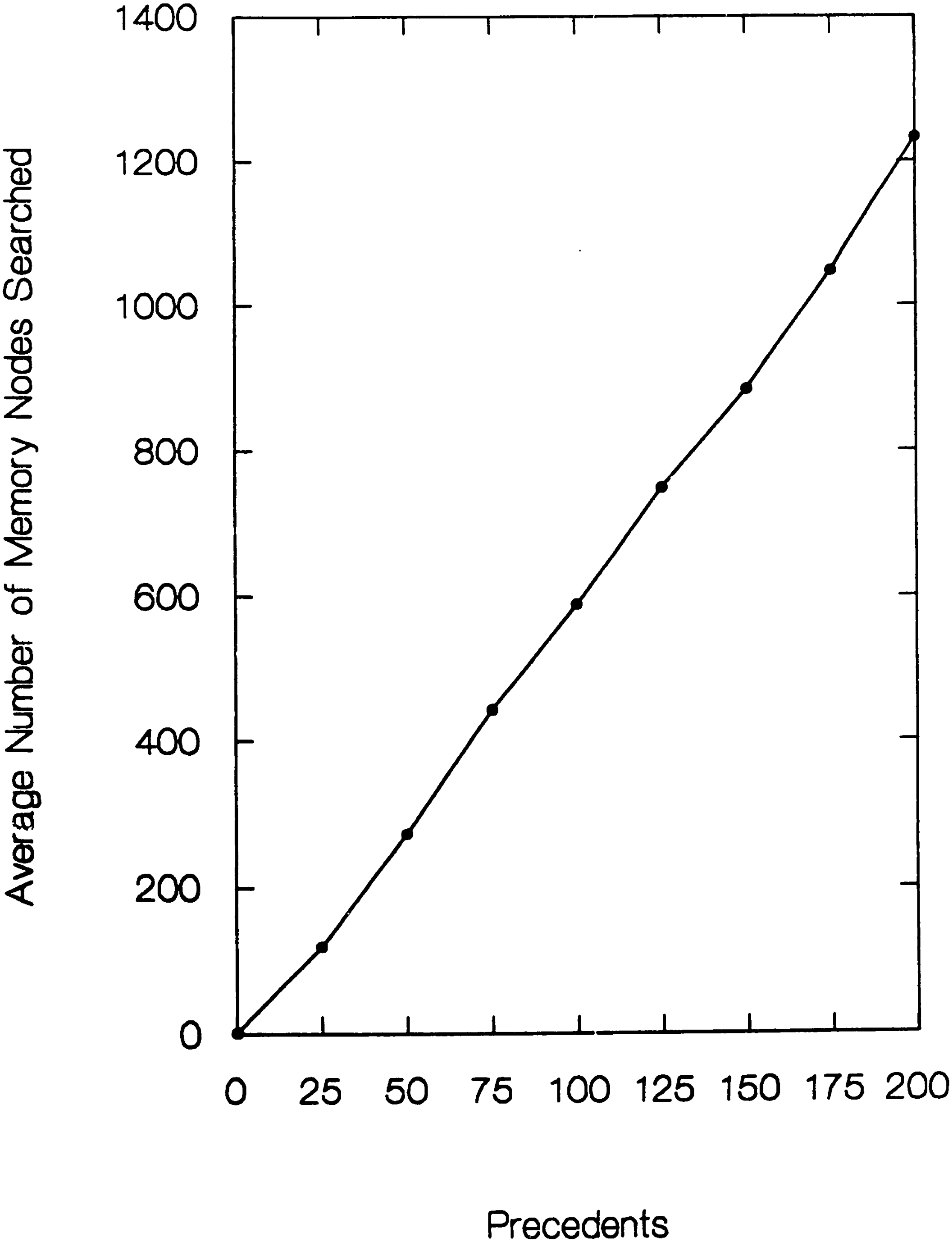


Figure 6.5: Growth of CASEY's retrieval time

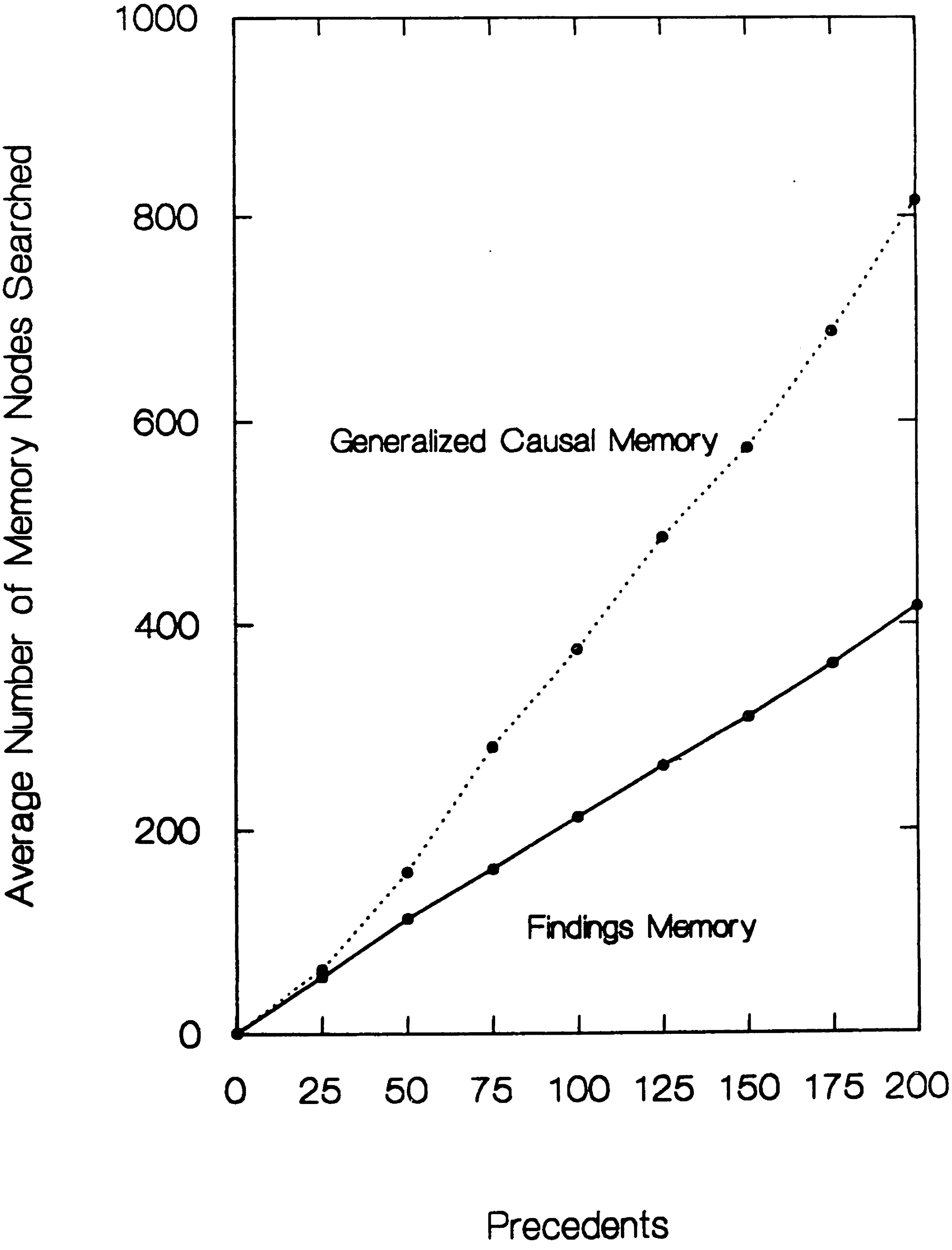


Figure 6.6: Retrieval time growth, broken down between the findings memory and the generalized causal memory

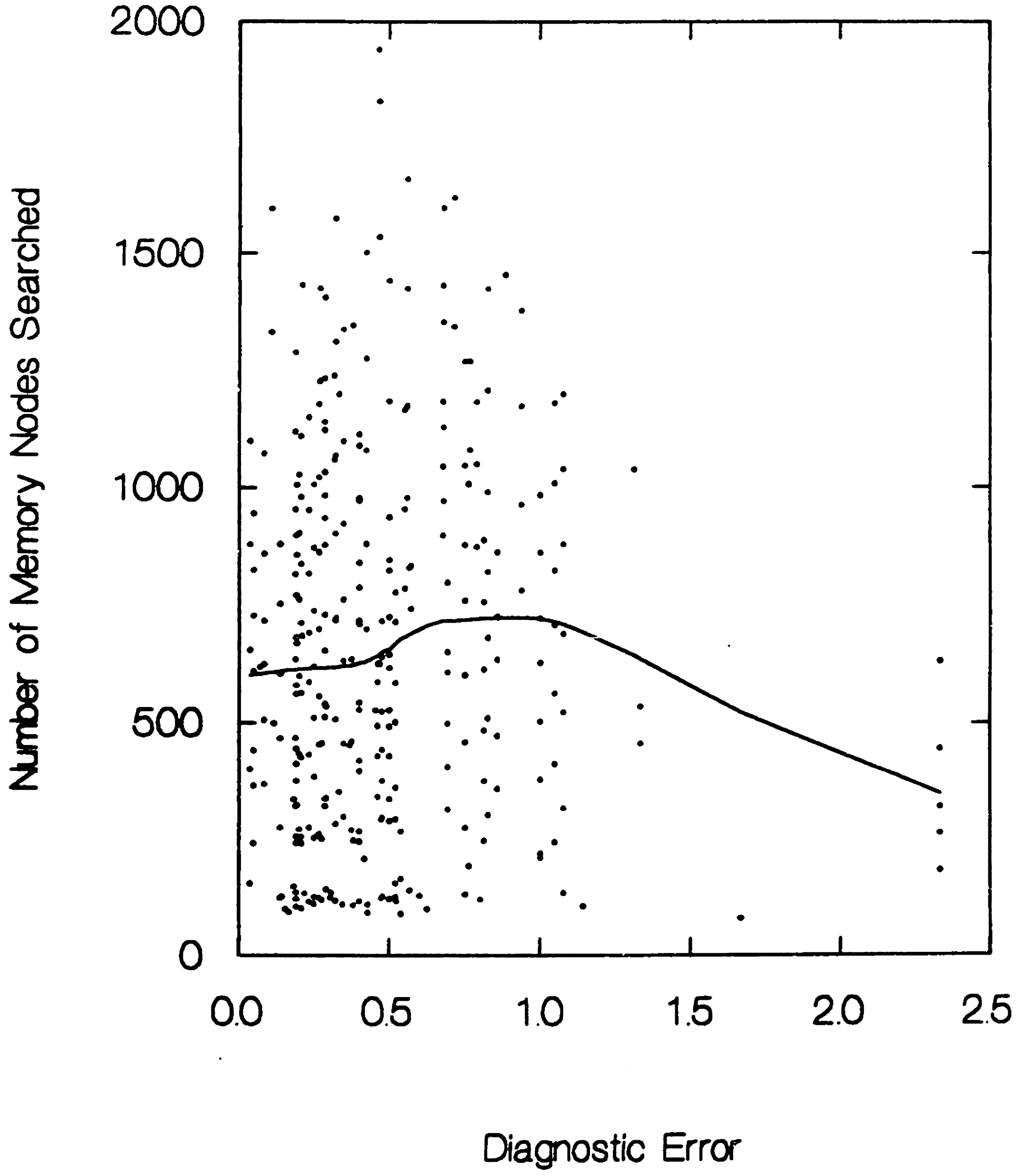


Figure 6.7: Correlation between the diagnostic error ϵ and retrieval time

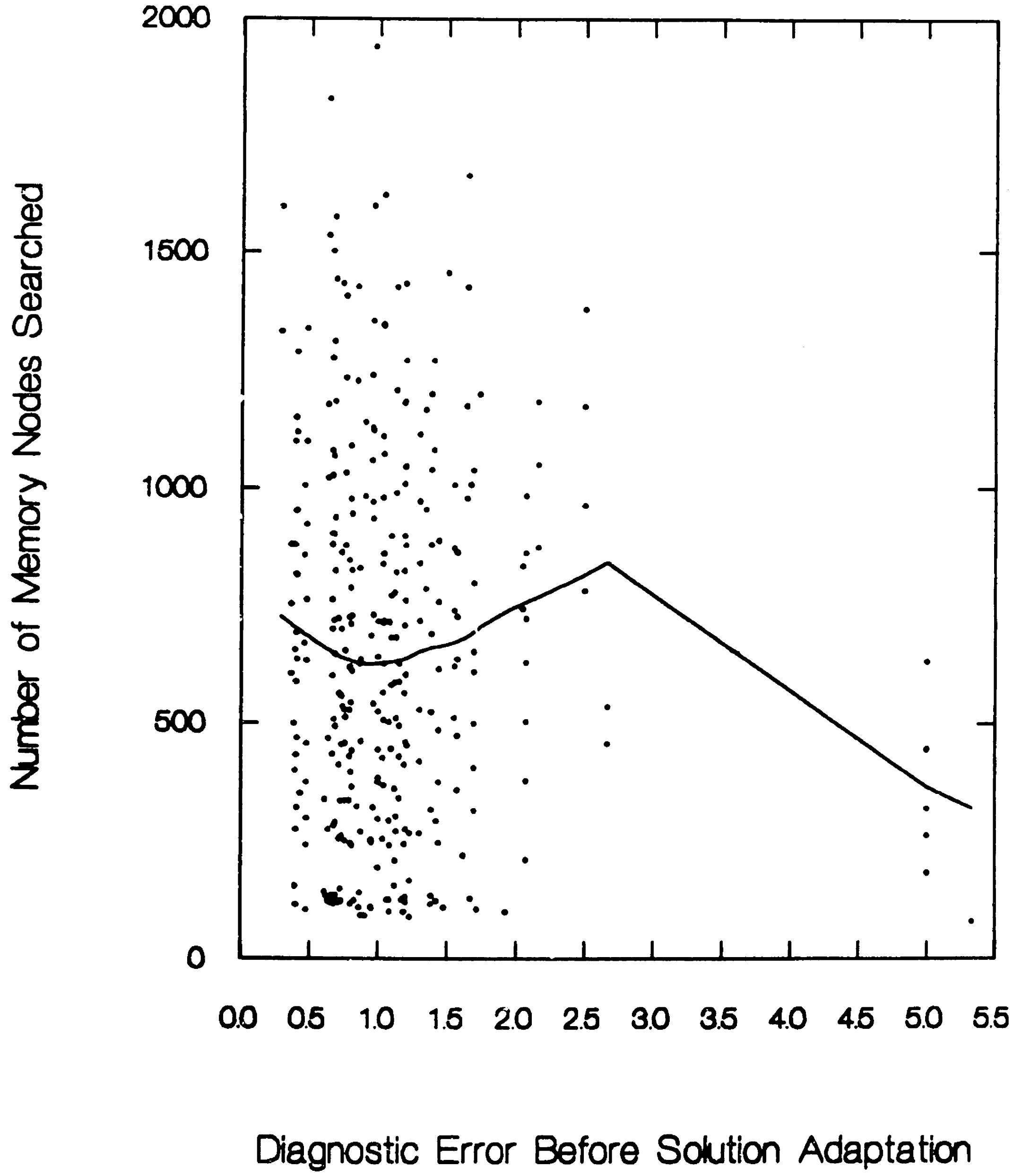


Figure 6.8: Correlation between the diagnostic error ϵ_P before solution adaptation and retrieval time

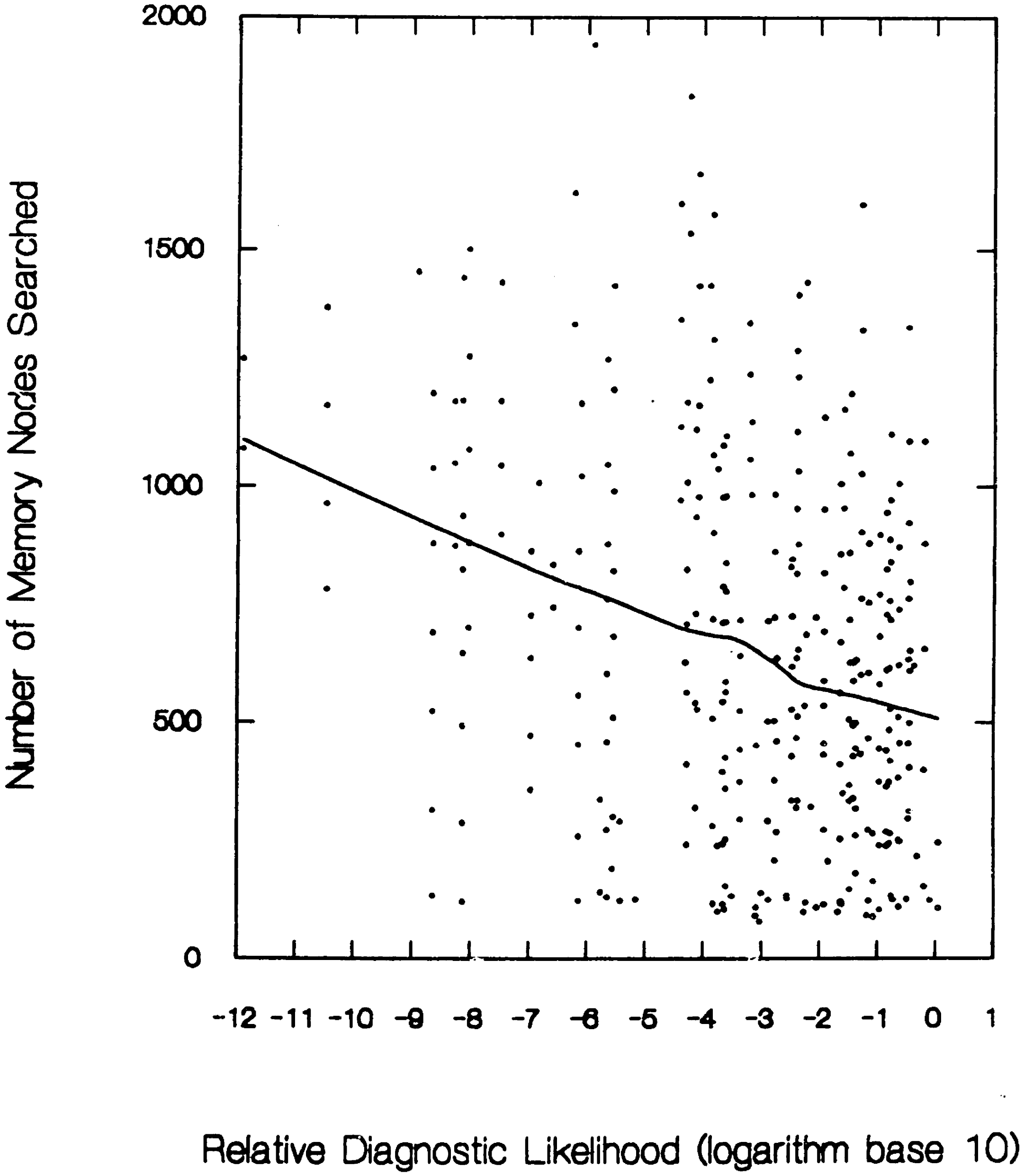


Figure 6.9: Correlation between the relative diagnostic likelihood λ and retrieval time

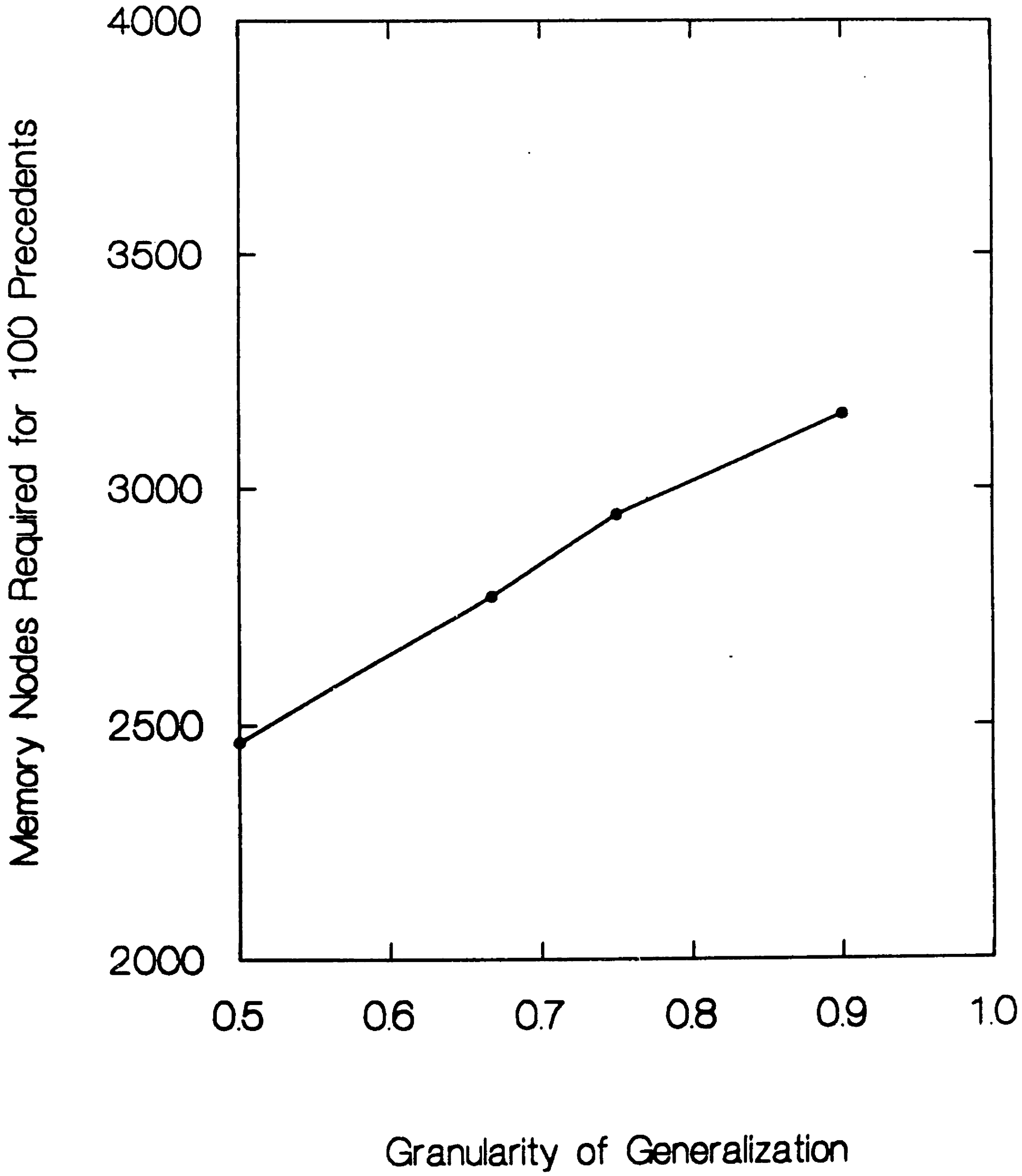


Figure 6.10: Effects of generalization granularity on memory growth

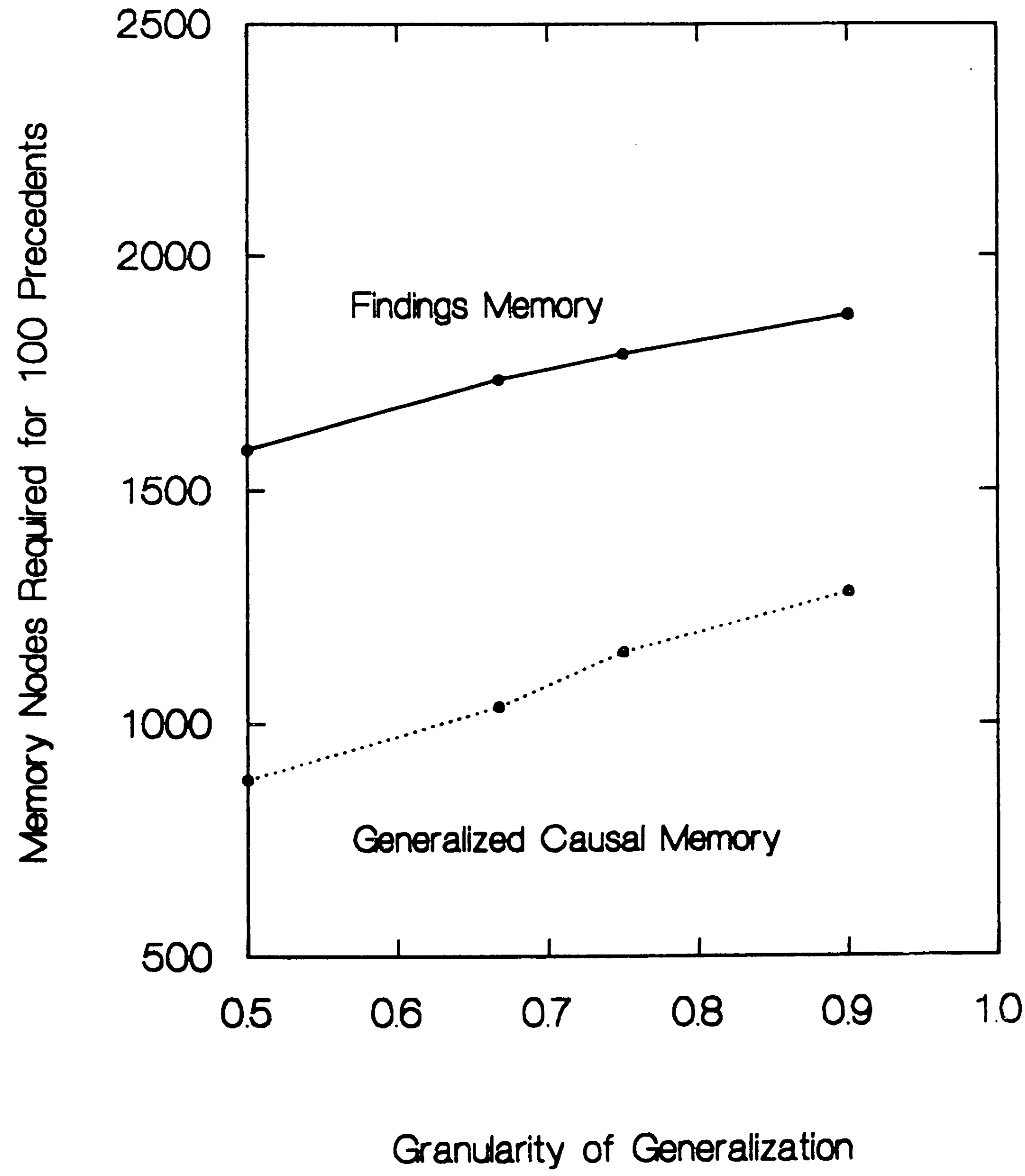


Figure 6.11: Effects of generalization granularity on memory growth, broken down between the findings memory and the generalized causal memory

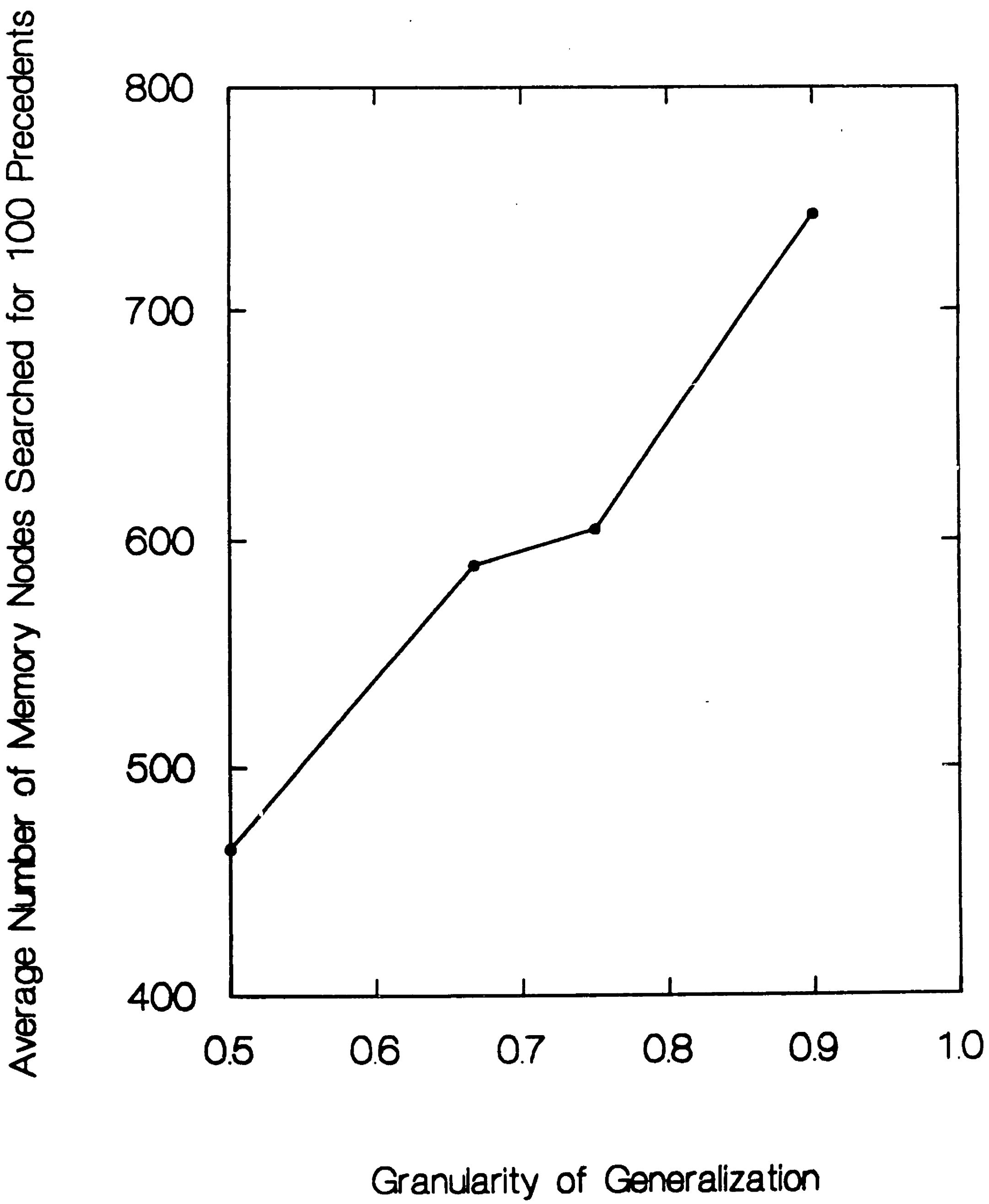


Figure 6.12: Effects of generalization grarularity on retrieval time

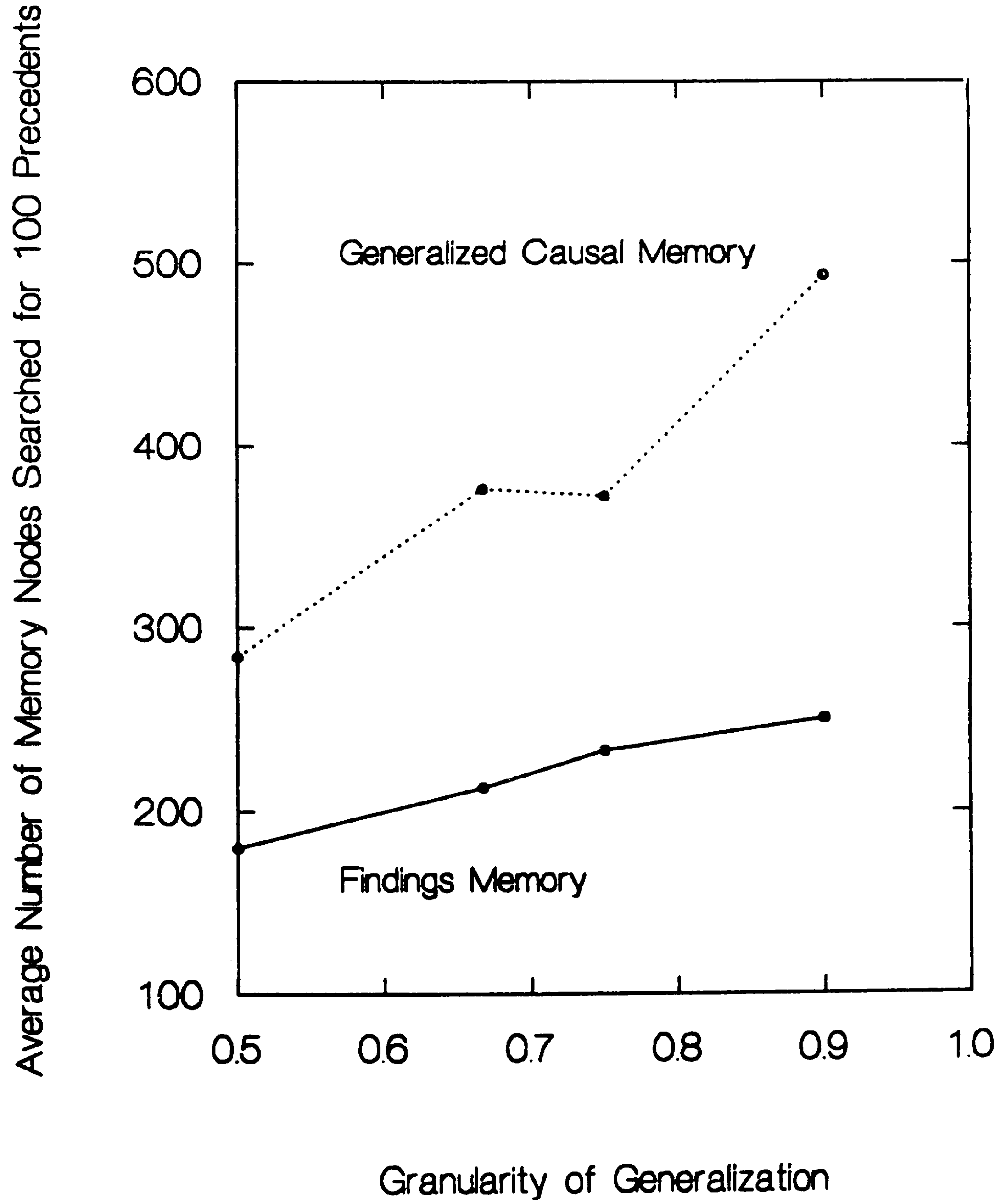


Figure 6.13: Effects of generalization granularity on retrieval time, broken down between the findings memory and the generalized causal memory

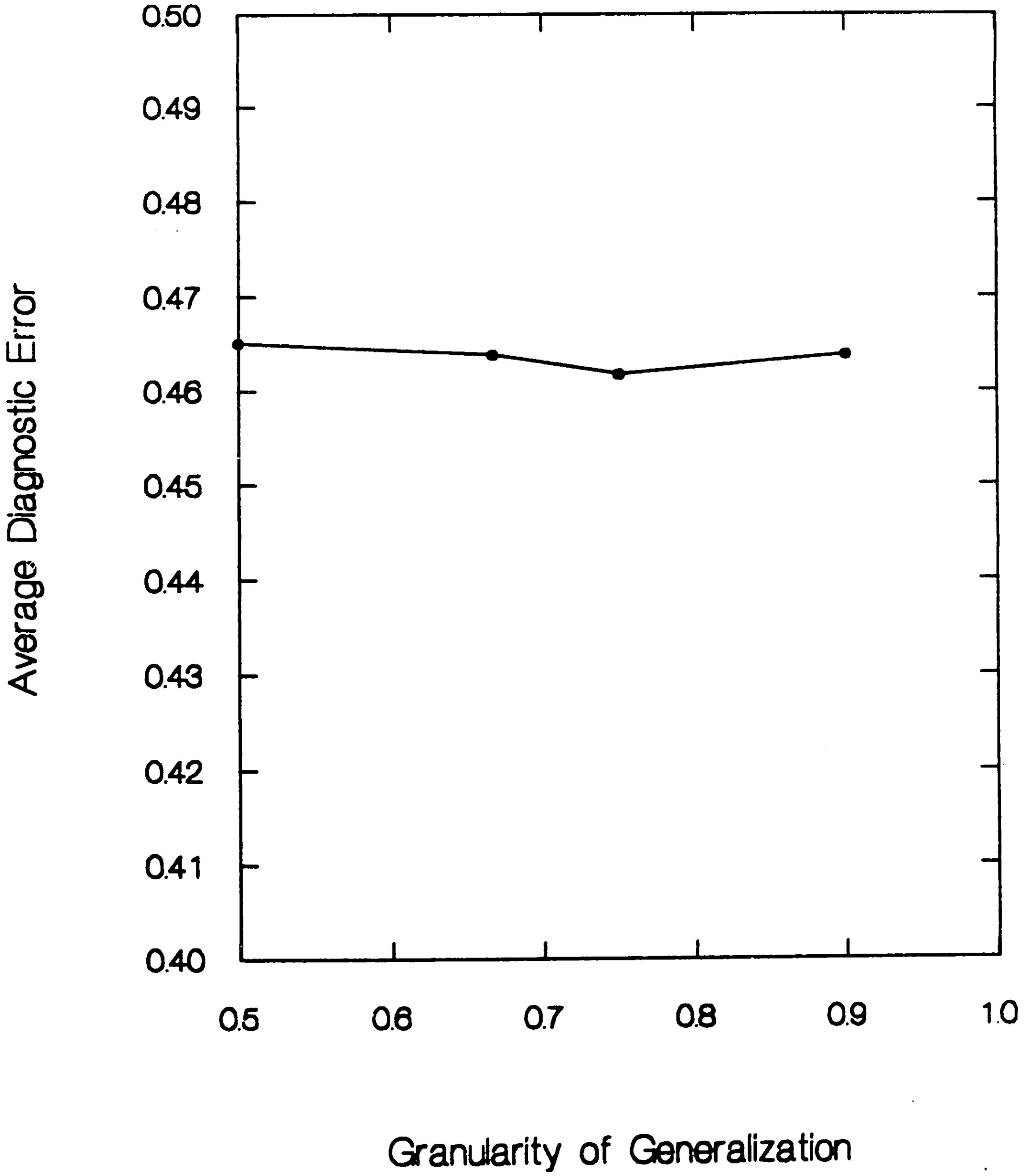


Figure 6.14: Effects of generalization granularity on the average diagnostic error ϵ

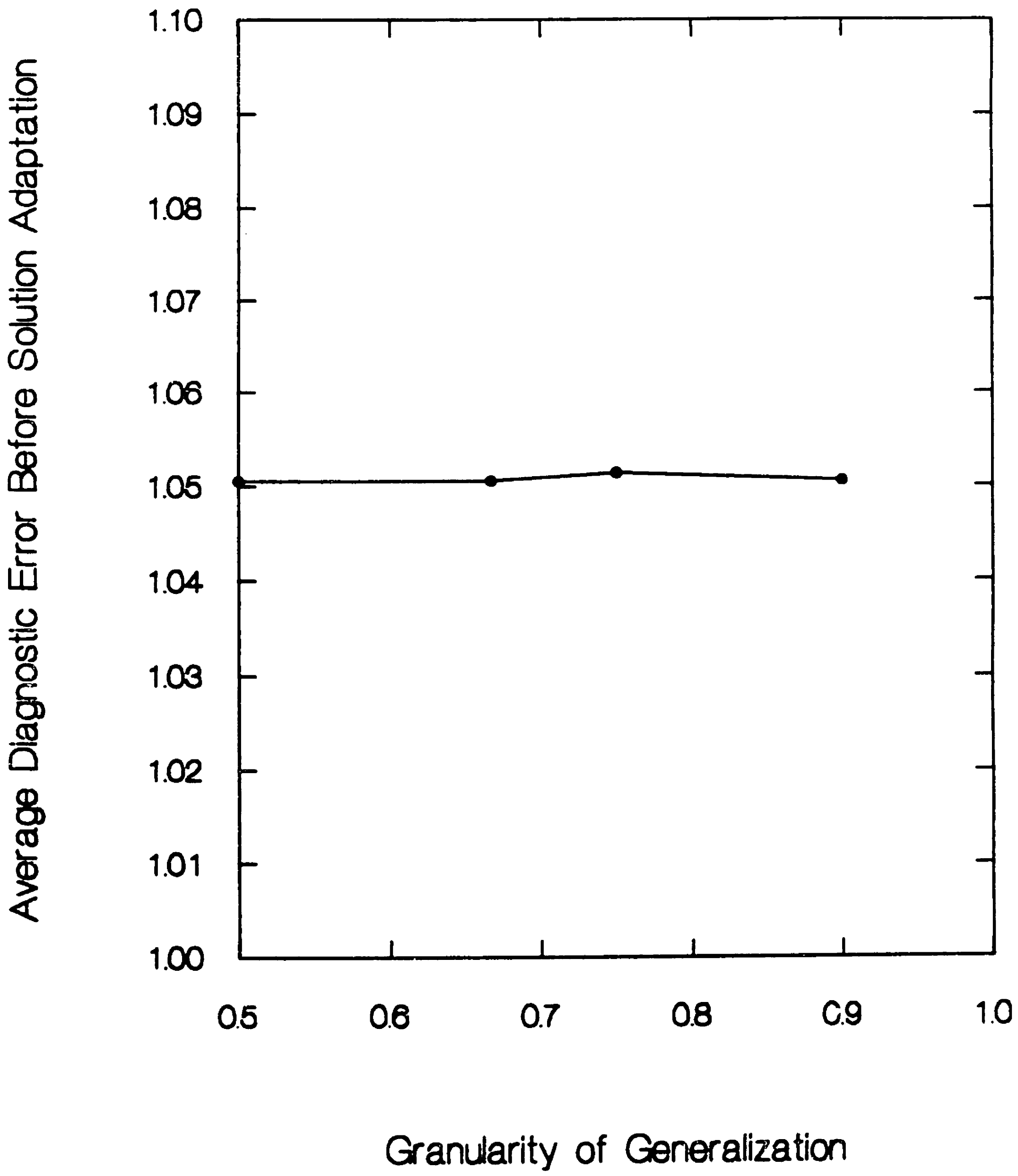


Figure 6.15: Effects of generalization granularity on the average diagnostic error ϵ_P before solution adaptation

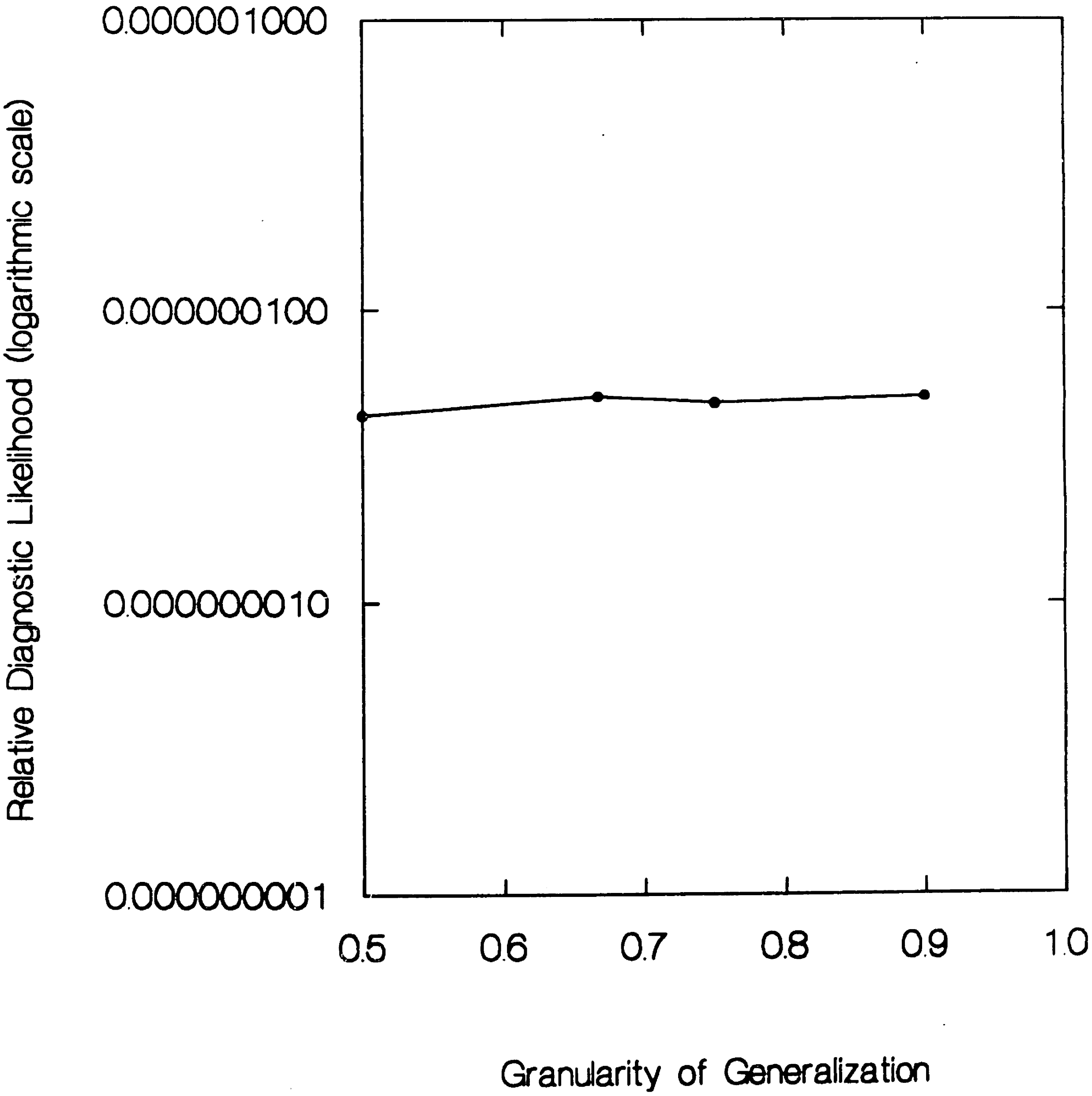


Figure 6.16: Effects of generalization granularity on the relative diagnostic likelihood λ

Chapter 7

Domain Regularity

7.1 Similar Cases

Cardiologists perceive a great deal of regularity in the heart failure domain, claiming that a preponderance of their diagnostic tasks are routine. The reason: many of the cases that they see are similar. However, if heart failure patients resemble each other significantly, then case-based reasoning should succeed, first in finding similar cases among potential precedents, and secondly, in transferring the diagnoses of these homologous precedents to new patients. As this study demonstrates, case-based reasoning confounds our expectations, unable to construct accurate diagnoses even with 200 potential precedents. Two aspects of regularity are consequently at issue.

Conceivably, *similar cases may not recur.* The resemblances that cardiologists suggest could be more diffuse than what CASEY expects, perhaps requiring thousands of cases for adequate representation. Human beings noιoriously make broad generalizations on an abstract, encompassing level while implicitly cognizant of important individual differences. For example, we can state the rule "look before you leap" with conviction and still be able to finesse circumstances under which "he who hesitates is lost."

Alternately, and perhaps in addition, *similar cases may not require similar solutions.* Symptomatic or syntactic resemblance does not necessarily entail the diagnostic, semantic homology upon which case-based reasoning depends. Heart failure is a complex, context sensitive domain, in which the *collective* evidence provided by clusters of findings reveals more than the *collected* contributions of findings considered in isolation. For example, a systolic ejection murmur is often evidence for high cardiac output. However, when other symptoms characterize the cardiac output as normal or low, the same murmur is more probably explained by aortic stenosis. Mitral regurgitation, although only weakly suggested by a particular murmur alone, is almost definitely the murmur's cause if it is supported by more specific evidence. Since the presence or absence of a particular finding can determine the explanation of others, two almost equivalent cases may diverge diagnostically. Thus, CASEY may retrieve "similar" precedents whose causal explanations do not lead to the best explanation for the new patient.

Uncovering the heart failure domain characteristics responsible for CASEY's diagnostic inefficacy will allow us to delineate a general class of domains for which case-based reasoning fails. Accordingly, this chapter presents a series of experiments designed to investigate the domain, with respect to whether:

- Similar cases recur

- Similar cases require similar solutions

However, before we undertake these inquiries, we must define what we mean by "similar cases."

Intuitively, two cases are similar if they have a large number of findings in common. However, because disease states can express themselves through any of a variety of findings, cases with equivalent disease states may differ superficially. Generalized causal features (see section 3.2), which represent the pathophysiologic states for which findings provide evidence, seem to be a more powerful predictor of diagnostic similarity. However, because new findings have not yet been assigned to particular pathophysiologic states, *all* of their *possible* causes must be included among the generalized causal features, thus diffusing the power of the representation.

CASEY's own similarity metric enumerates precedent pathophysiologic states matched by the new case's generalized causal features, but then penalizes the precedent for its unmatched states. In this way, the system avoids precedents that contain overwhelming amounts of inappropriate and misleading information. As precedents can be judged both by common findings and by common "gen causals," they can also be penalized according to both representations. Table 7.1 diagrams the four similarity metrics that result.

The last conceivable similarity metric is the perfect, or retrospective, metric, which judges two cases to be similar if their diagnoses agree. CASEY can not use this metric until the new case has been diagnosed, thus defeating its purpose, but I can exploit it to measure the absolute limits of regularity in the heart failure domain.

M1 Common Findings	M3 Common Gen Causals
M2 Common Findings - Unmatched Findings	M4 (CASEY's Metric) Common Gen Causals - Unmatched Gen Causals

Table 7.1: Four similarity metrics

7.2 Do Similar Cases Recur?

If similar cases recur, then as case base size increases, the best matched precedent should approximate the new case more and more closely, approaching perfect correspondence in the limit. To test this expectation, I compared each of the 40 test cases to the 200 precedents in order, measuring similarity based on the four metrics described in the previous section. For each metric, I plotted 200 points, each corresponding to the highest similarity achieved by any of the first N precedents. Thus, the graphs in figures 7.1 through 7.4 portray increases in the similarity of the best matched precedent, as case base size grows from 1 to 200.

Values along each metric rise sharply and then plateau abruptly before even 25 precedents have been considered. The plots thus illustrate simply and conclusively the reason for accuracy's unresponsiveness to experience. Figure 7.1 shows the number of common findings halting well below 40, the total number of findings in a typical case. When the number of unmatched findings is subtracted from this measure (figure 7.2), the resulting net similarity does not even reach zero, signifying that, on average, even the best matched precedent from a case base of 200 has more findings unmatched than matched. Encompassing all possible causes of the new findings, generalized causal features tend to be a catch-all net, matching large numbers of precedent pathophysiologic states, and matching larger numbers of states from larger precedents (notice the jump at precedent 86 in figure 7.3). Consequently, the number of common "gen causals" is not a particularly discriminating metric. CASEY's own metric, which subtracts the number of unmatched "gen causals," is more sensitive; accordingly, it reaches a net similarity of only 7 (figure 7.4).

Apparently, although somewhat similar cases populate even a small pool of 25 precedents, truly similar cases are rare in even a moderately large pool of 200. Given the slight and, moreover, the diminishing slope of the four curves, CASEY might need vast numbers of cases before any of the metrics (with the exception of the third) approach maximum value.

7.3 Do Similar Cases Require Similar Solutions?

For the purpose of analyzing the correlations between case similarity and solution similarity, I derived diagnoses for each of the 40 test cases using each of the 200 precedents and measured the following:

- Case similarity, according to the four metrics

- The difference between the precedent causal explanation and the ideal causal explanation for the test case, computed as the number of different pathophysiologic states

- The raw error ϵ_P of the precedent diagnosis before solutio adaptation, with respect to the ideal
- The error ϵ of CASEY's solution after adaptation, with respect to the ideal
- The relative likelihood λ of CASEY's solution, with respect to the ideal

Figure 7.5 displays a scatterplot matrix of diagnostic difference, ϵ_P, and ϵ graphed against the four similarity metrics. The points of each plot represent a random selection of approximately 250 of the original 8000 test case and precedent pairs;[1] the curves were computed using a locally weighted smoothing algorithm. On average, solution disparity decreases with increasing case similarity, and therefore, on average, similar cases *tend* to require similar solutions. However, the plots are diffuse and the confidence intervals are wide, so that, despite the trends, a similar precedent does not guarantee an appropriate diagnosis.

Although all correlations are weak, CASEY's similarity metric (M4) proves to be the strongest predictor of precedent suitability. Figures 7.6 through 7.11 enlarge the plots involving this metric and the baseline metric (M1), which measures the number of common findings. The larger plots incorporate 500 randomly selected points.

What appear to be horizontal lines across the ϵ plots in figures 7.5, 7.10, and 7.11 are in fact concentrations of points at the default value 1, assigned to ϵ when CASEY deems solution transfer inappropriate. Recall that solu-

[1] Restricting the point set was unavoidable given the capacity of the graphing package.

tion adaptation fails either when a new finding cannot be explained within the precedent diagnostic scheme, or when pruning unsupported pathophysiologic states leaves the causal explanation bereft of a diagnosis (see section 3.2). Because similar cases *tend* to require similar solutions, the proportion of unsuitable precedents should decrease with increasing similarity. The graphs in figure 7.12, however, refute this supposition. For each of the four metrics, I organized the 8000 test case and precedent pairs according to their similarity, computing the fraction of inappropriate matches at each similarity value. Although a locally weighted smoothing algorithm attempted to characterize curves for the plots, they were distinctly uncorrelated. Thus, despite the overall similarity between two cases, a few crucial, distinguishing findings can dichotomize their causal explanations to the point of incompatibility.

Figure 7.13 is a matrix of scatterplots involving the four similarity metrics and the relative likelihood λ, plotted on a logarithmic scale. 500 test pairs were selected for the plot, but because inappropriate solution transfer results in a default relative likelihood of 0, which is infinitely negative on a logarithmic scale, unsuitable matches were not plotted, leaving 324 points. Lines were computed using locally weighted smoothing, and large scale versions of the graphs for metrics M1 and M4 appear in figures 7.14 and 7.15.

Surprisingly, the relative likelihood criterion remains independent of precedent similarity. Unlike the Heart Failure system, CASEY does not take full account of the probability of the diagnostic hypotheses that it generates. Rather, it accepts a precedent scheme unquestioningly, linking new case findings to it whether these links are probable or not. Given the mediocre precedents that populate a 200 patient case base, finding a match that is not only qualitatively close, but quantitatively *likely* as well, appears to require different mechanisms. Perhaps precedent matches must be almost perfect before they can even begin to constrain the likelihood of solution transfer; alternatively, perhaps the notion of precedent matching is not the appropriate constraint.

In summary, *do* similar cases require similar solutions? The best answer that the heart failure domain can muster is "possibly." Case similarity only weakly influences solution difference and error of transfer; indeed, each similarity value admits a wide range of close and distant solution matches. The domain is even more noncommittal with respect to associating similar precedents with likely, or even appropriate, solutions for the new case. Interactions among findings that determine diagnostic success are too context sensitive and specific to be accounted for by an overall metric.

7.4 Retrospective Similarity

In order to probe the true limits of regularity in heart failure diagnosis, I will leave our four practical, but fallible, similarity metrics behind. Instead, I will judge two cases as close if their *diagnoses* are close, calling my metric retrospective, because it measures similarity by hindsight, after diagnosis has taken place. Using the new, perfect metric, we can combine the two questions "Do similar cases recur?" and "Do similar cases require similar solutions?", asking the more profound question:

> Do similar solutions recur?

To pursue the new line of inquiry, I used the entire pool of 240 cases, constructing the $\begin{pmatrix} 240 \\ 2 \end{pmatrix} = 26860$ possible case pairs and comparing the ideal diagnoses of the two cases within each pair. Figure 7.16 is a histogram of the number of case pairs differing in the specified number of pathophysiologic states. Unequivocally, similar solutions *do not* recur. All 240 solutions are mutually distinct, and the *closest* pair of solutions differ in *five* disease states. Moreover, the median pair of solutions differ in almost *thirty* states, which is larger than the size of a typical causal explanation (23.5 states), and therefore larger than half the size of the typical pair. Whatever regularity heart failure diagnosis contains must be exceedingly diffuse.

Figure 7.17 plots the fraction of the 26860 case pairs whose diagnostic difference was at most the specified number of states. Almost none of the case pairs differed in less than 10 states, while 90% of them differed in at least 18 states. Astonishingly, 80% of the case pairs differed in 21 states, almost half the size of a typical diagnostic pair.

The fraction plotted in figure 7.17 represents the proportion of pairs of cases whose solutions fell within a specific distance of each other. It can also be interpreted as an empirical estimate of the probability that the solutions of a random case matching will achieve the specified proximity. Therefore, the reciprocal of this probability bounds the expected number of precedents required to assure us that a new case will find a close match. Figure 7.18 graphs this empirical expectation on a logarithmic scale. The number of precedents required rises with dizzying rapidity as we demand better matches for the new case, and the double logarithmic plot in figure 7.19 assures us, through its upward concavity, that the ascent is exponential. CASEY would need a case base of tens or even hundreds of thousands of patients to provide its new cases with decent precedents.

According to this empirical curve, if CASEY has 240 patients in its case base, then its best matched precedents should, on average, have solutions different from the new case diagnosis in approximately 10 or 11 pathophysiologic states. To corroborate the curve's prediction, I located, for each of the 240 cases, the closest causal explanation among the remaining 239 and computed the diagnostic difference between them. Figure 7.20 displays the resulting histogram. Averaged over the 240 cases, the closest diagnosis differed from the current case's solution in 12 states. Thus, the previous curve's prediction is approximately confirmed.

Figure 7.21 plots the fraction of cases for which the closest matching diagnosis fell within the specified distance. Less than 7% of the cases could find matching diagnoses with fewer than 7 different states, while approximately 80% of the cases could not find diagnoses with fewer than 9 different states. Bearing in mind that the average causal explanation contains only 23.5 states, we realize that among 240 cases, appropriate precedents simply do not exist; even a perfect, retrospective metric would not help us find them.

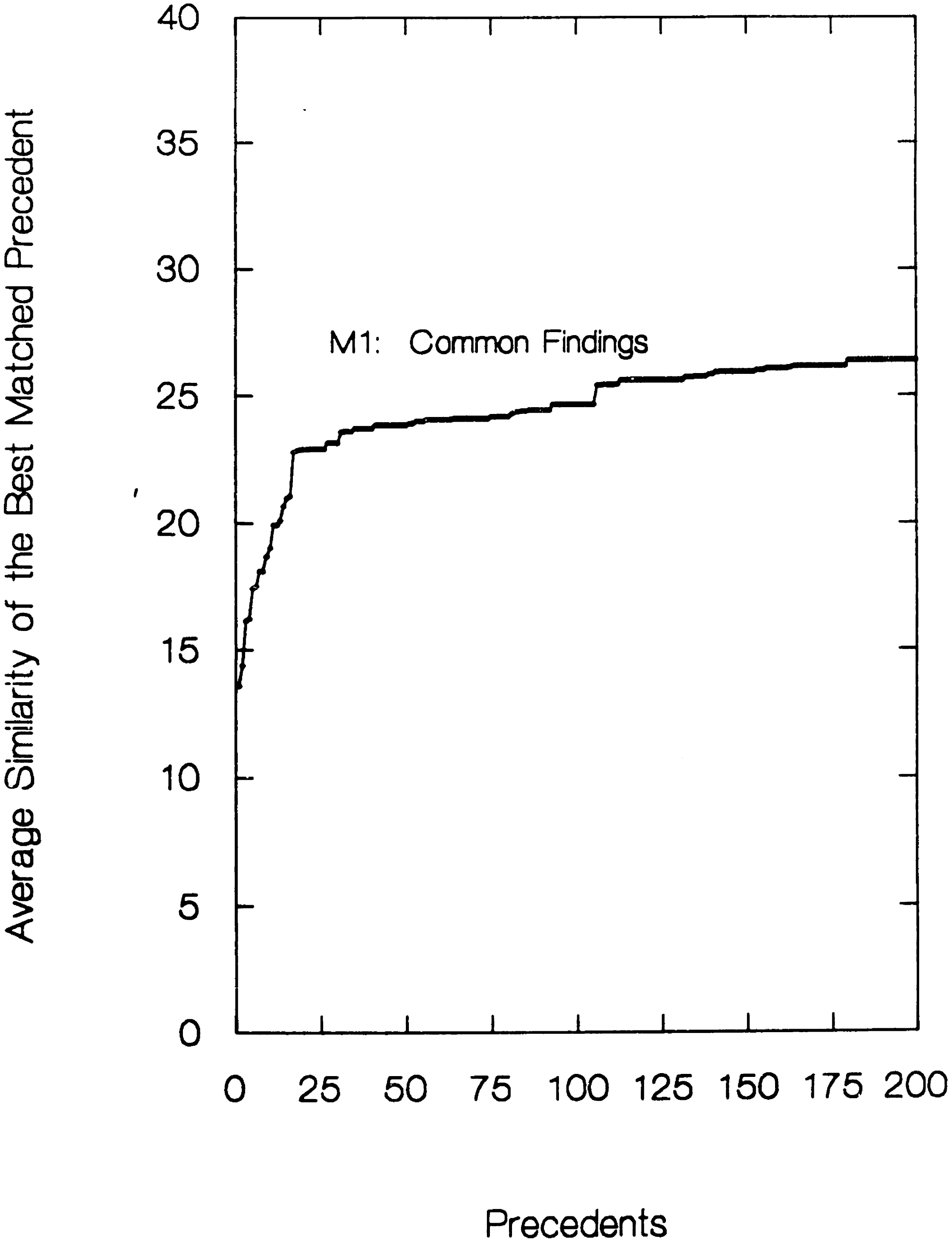


Figure 7.1: Effects of increasing experience on the similarity of the best matched precedent, judged by the number of common findings

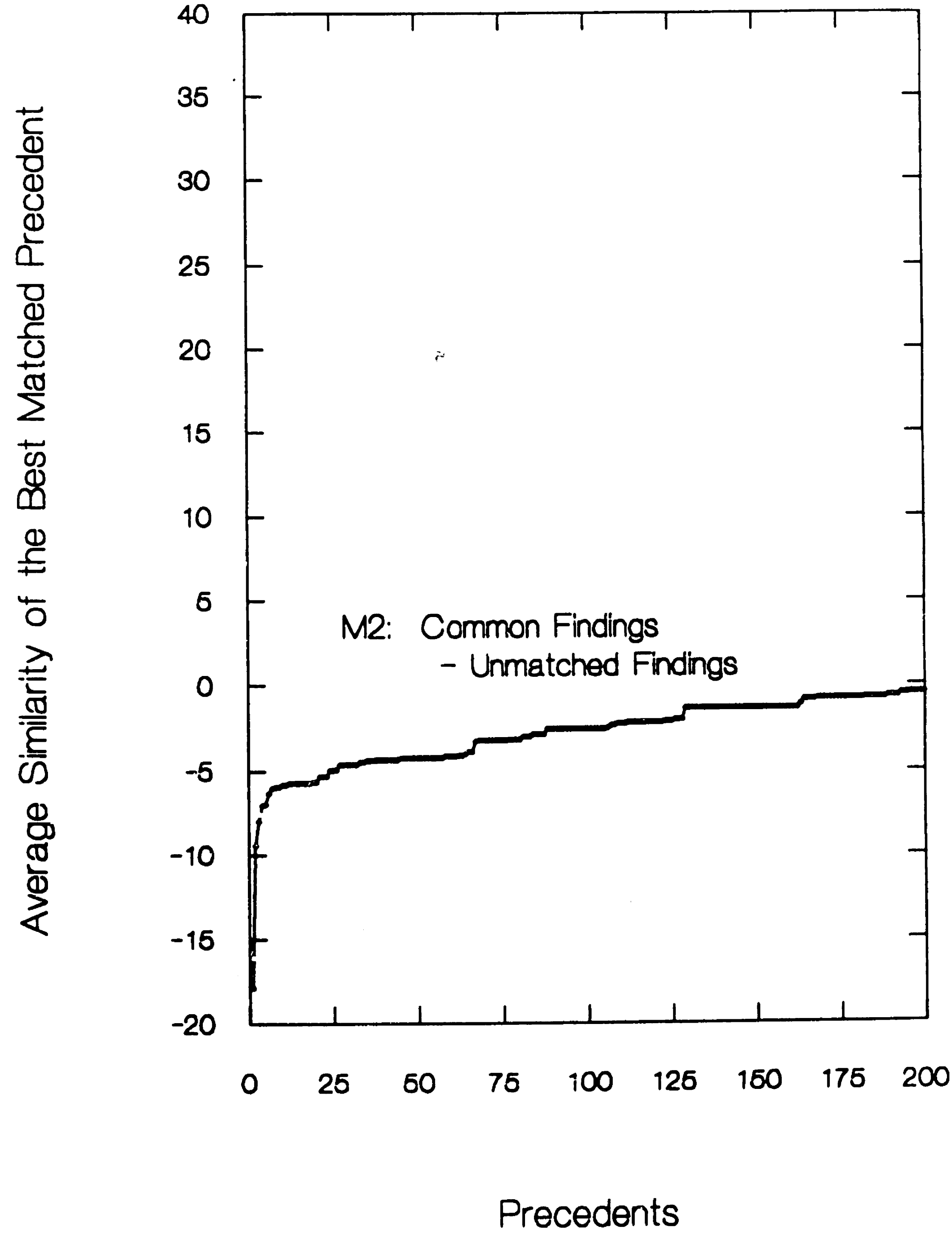


Figure 7.2: Effects of increasing experience on the similarity of the best matched precedent, judged by the number of common findings minus the number of unmatched findings

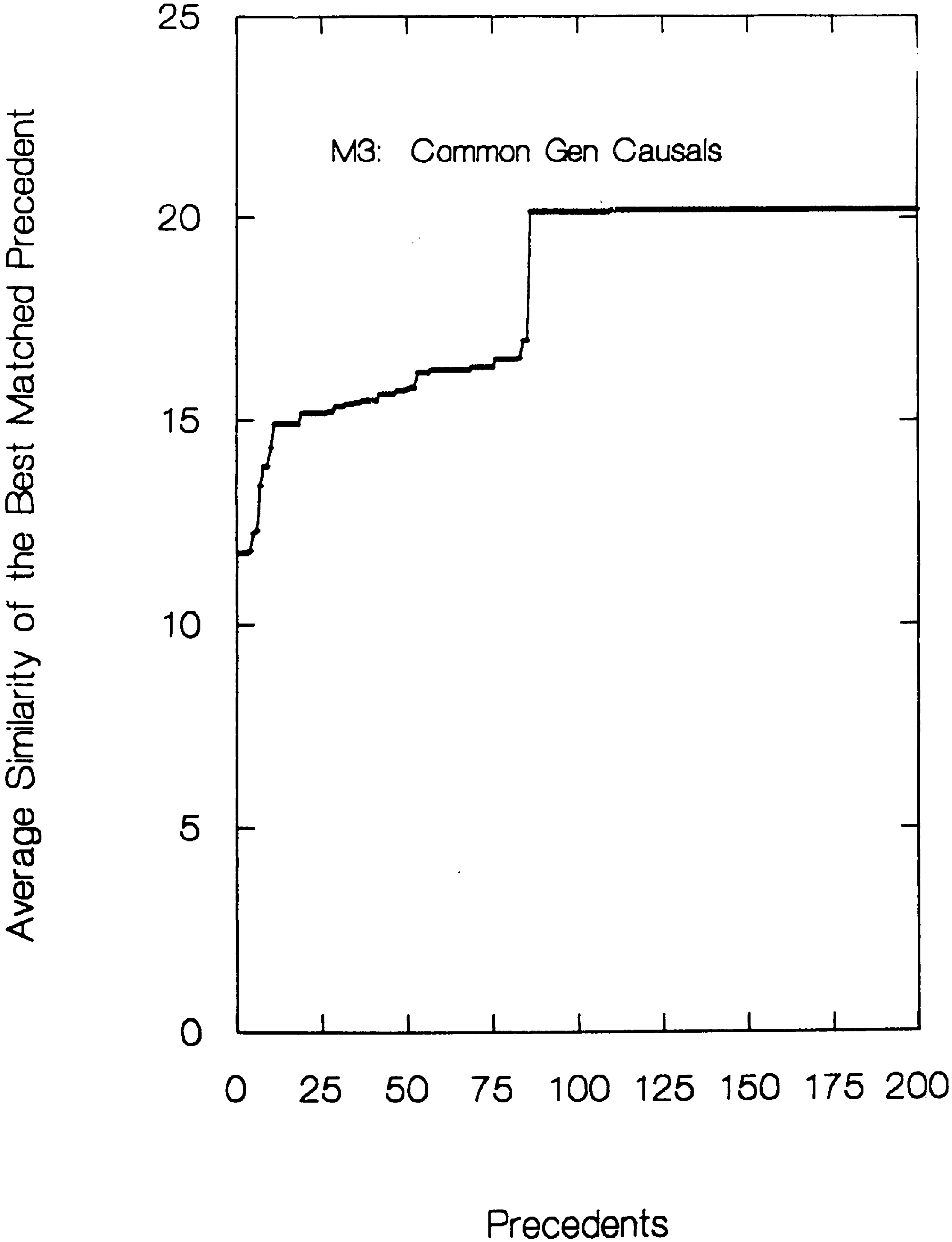


Figure 7.3: Effects of increasing experience on the similarity of the best matched precedent, judged by the number of common generalized causal features

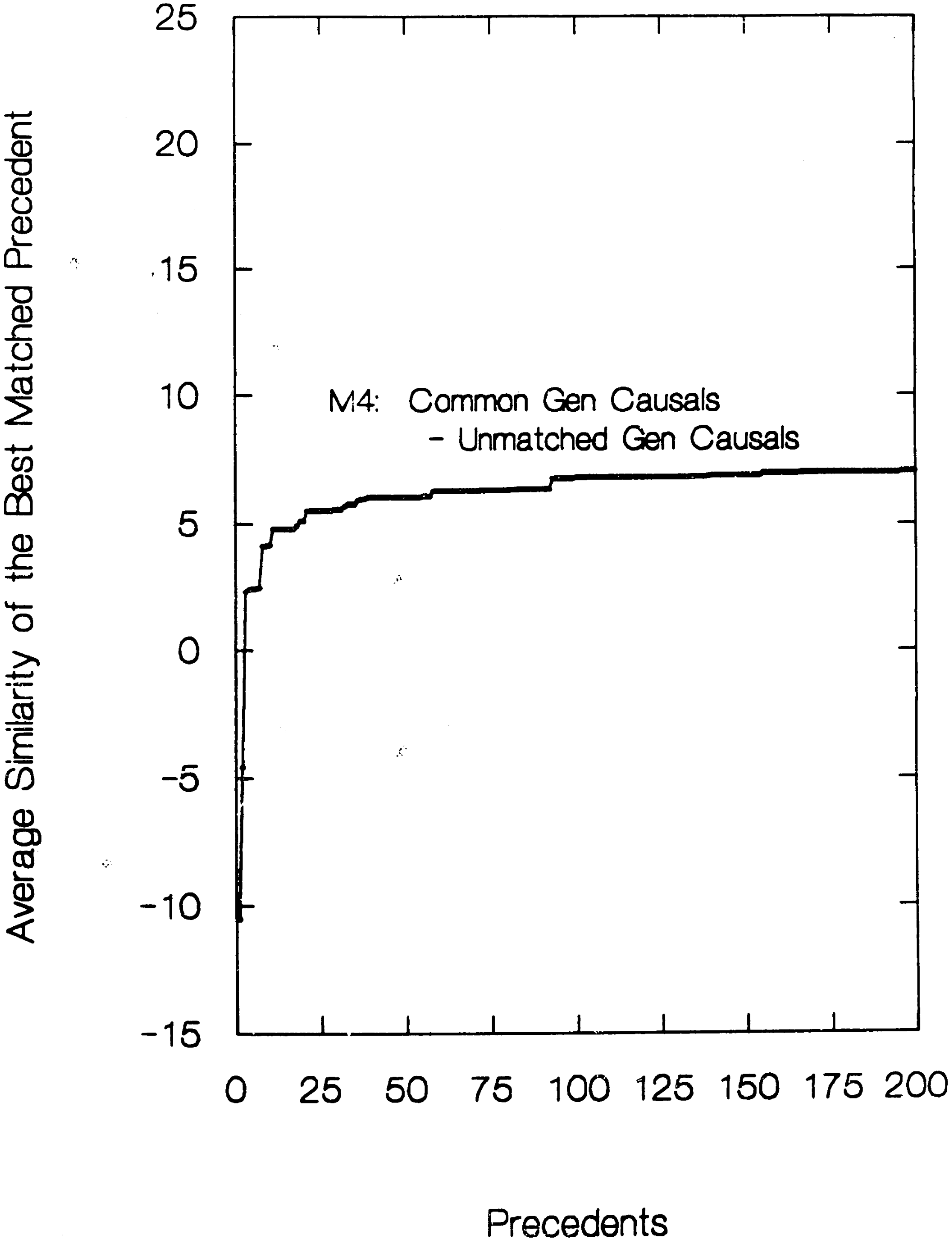


Figure 7.4: Effects of increasing experience on the similarity of the best matched precedent, judged by the number of common generalized causal features minus the number of unmatched generalized causal features

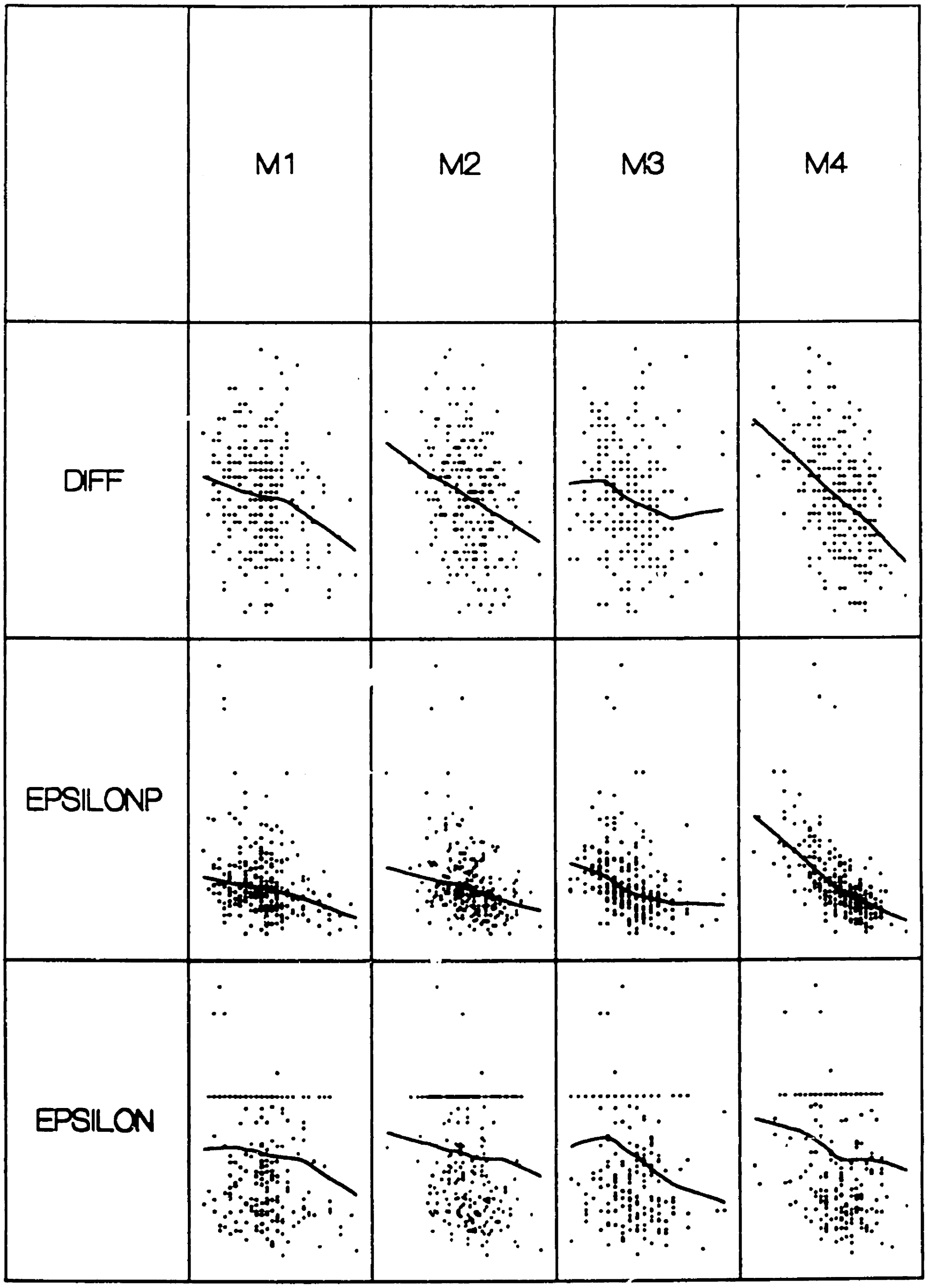


Figure 7.5: Correlations between case similarity and diagnostic disparity, measured by: 1. the diagnostic difference (DIFF); 2. the diagnostic error ϵ_P before solution adaptation (EPSILONP); and 3. the diagnostic error ϵ (EPSILON)

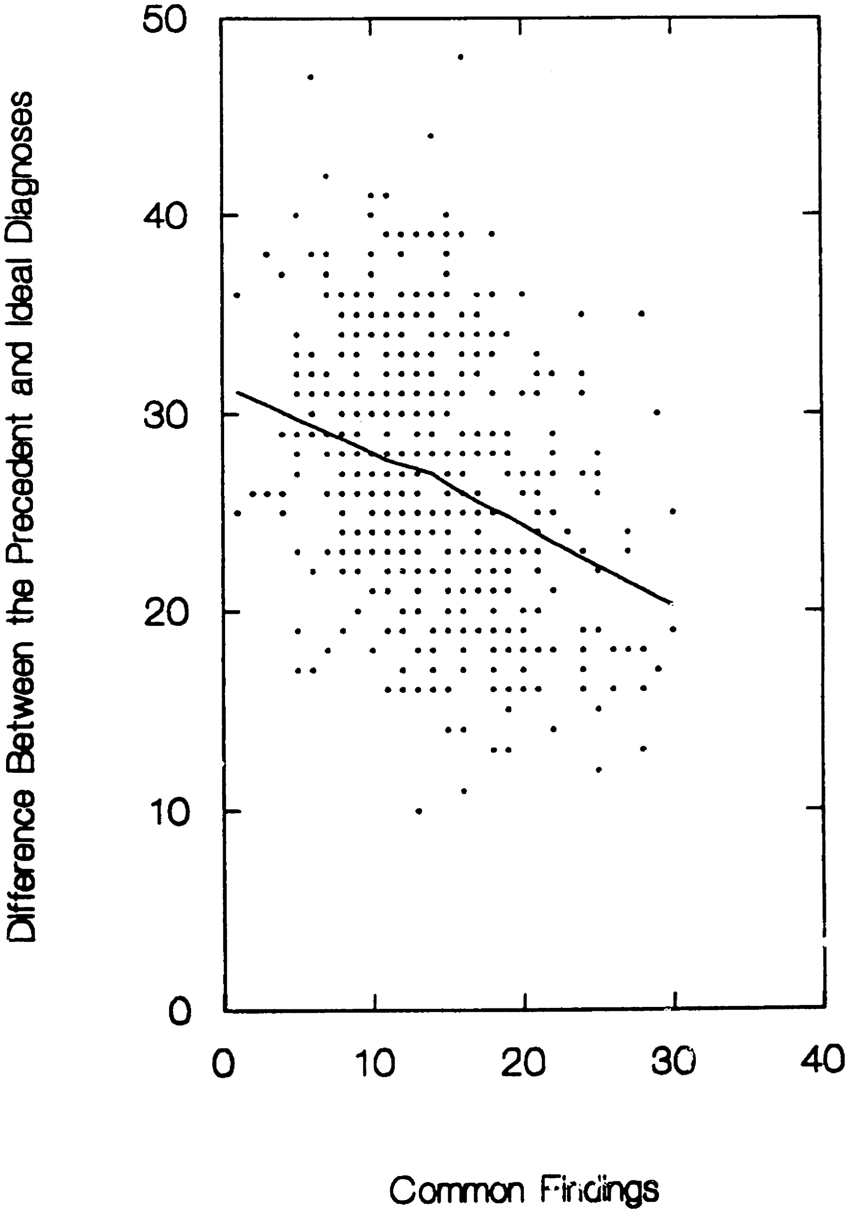


Figure 7.6: Correlation between the number of common findings (M1) and diagnostic difference

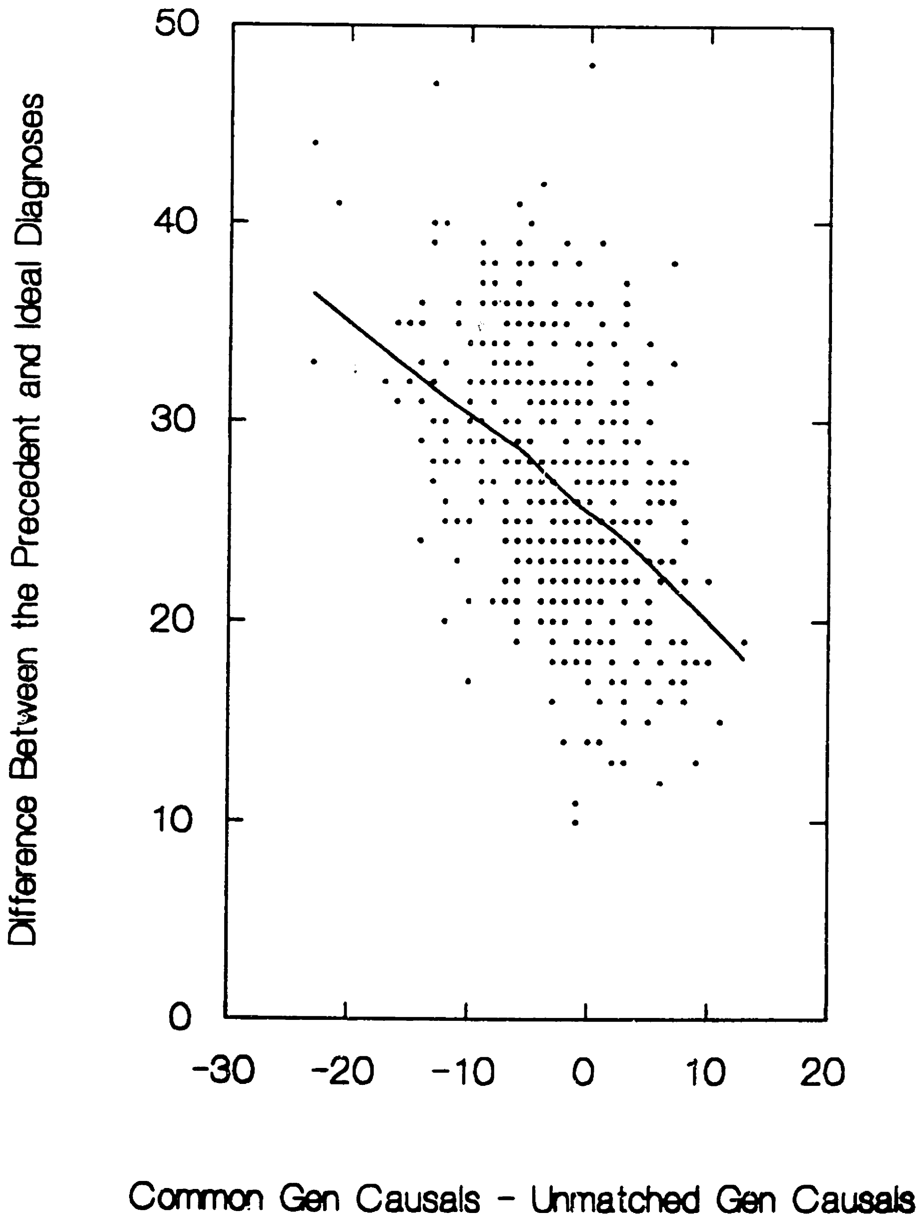


Figure 7.7: Correlation between CASEY's similarity metric (M4) and diagnostic difference

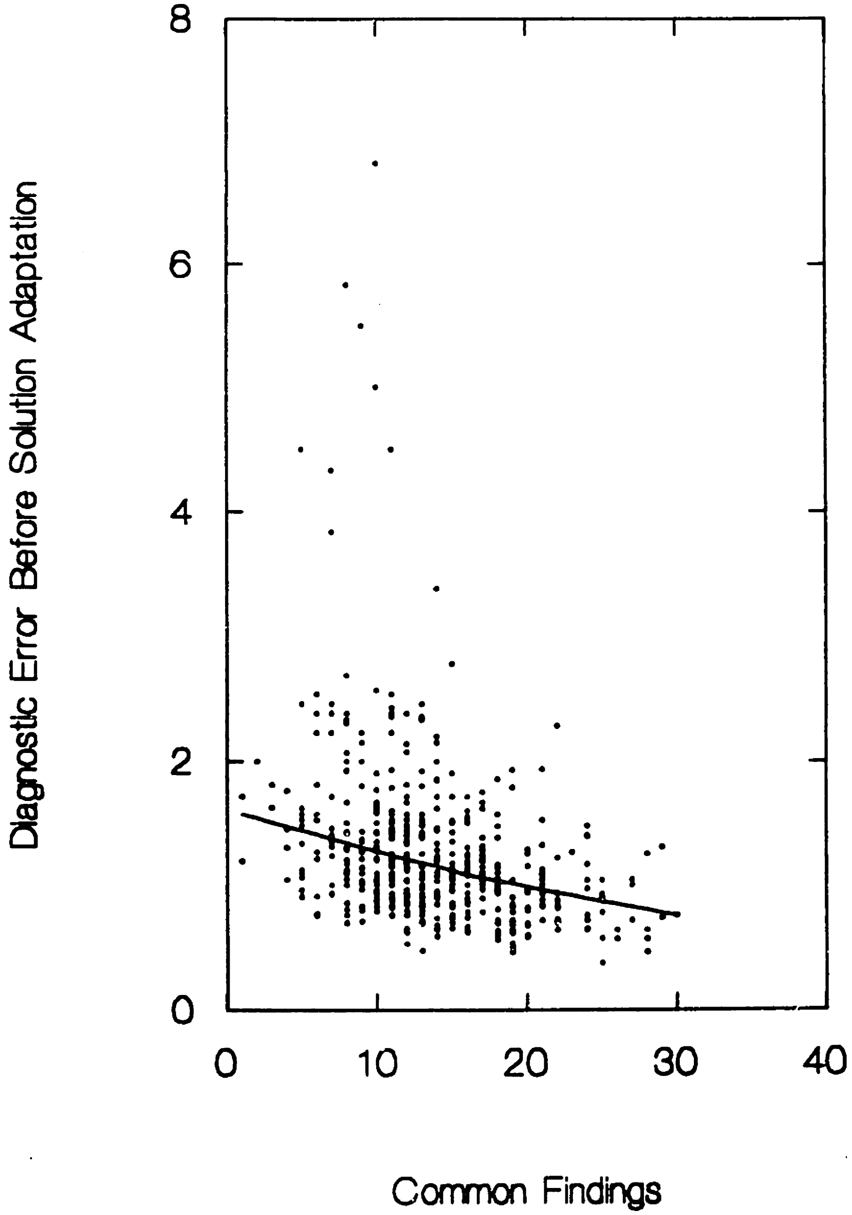


Figure 7.8: Correlation between the number of common findings (M1) and the diagnostic error ϵ_P before solution adaptation

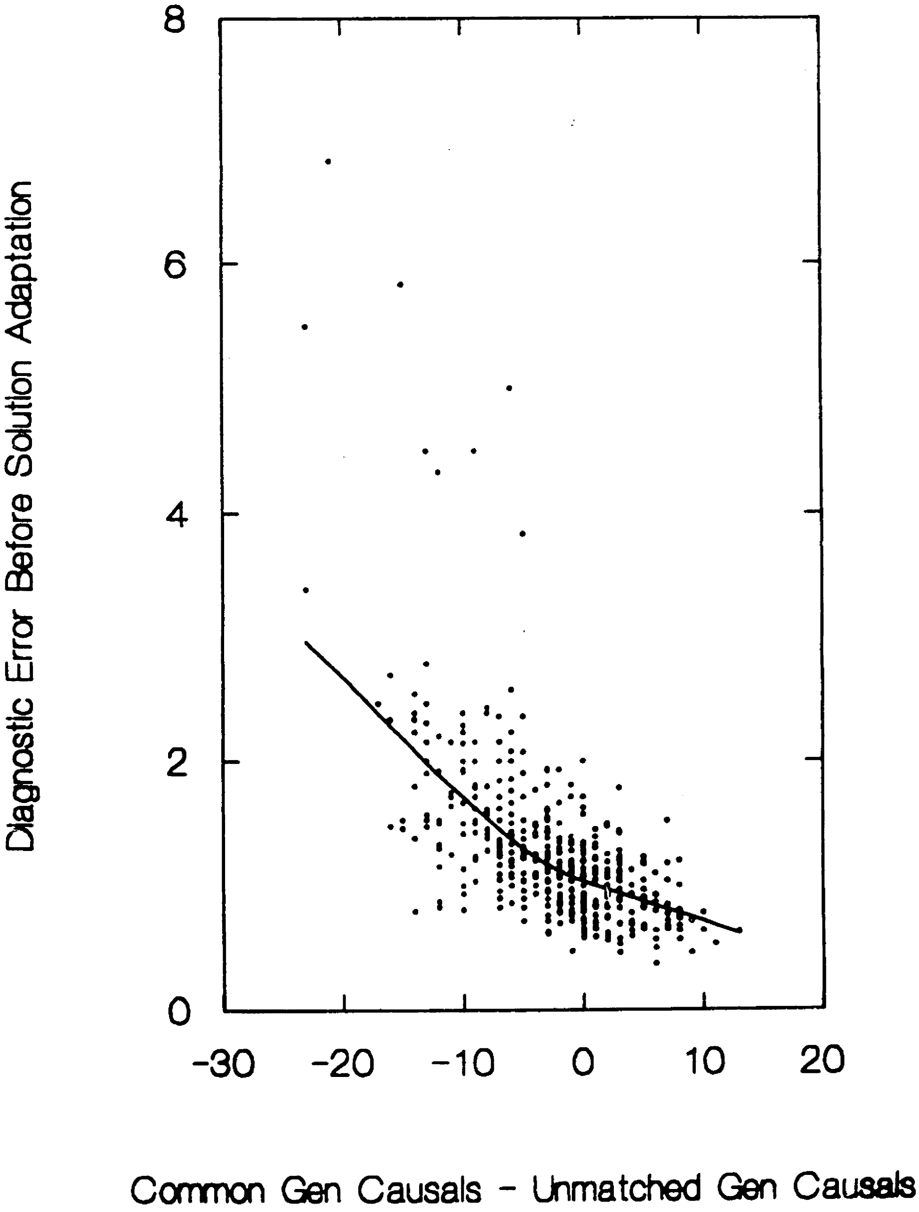


Figure 7.9: Correlation between CASEY's similarity metric (M4) and the diagnostic error ϵ_P before solution adaptation

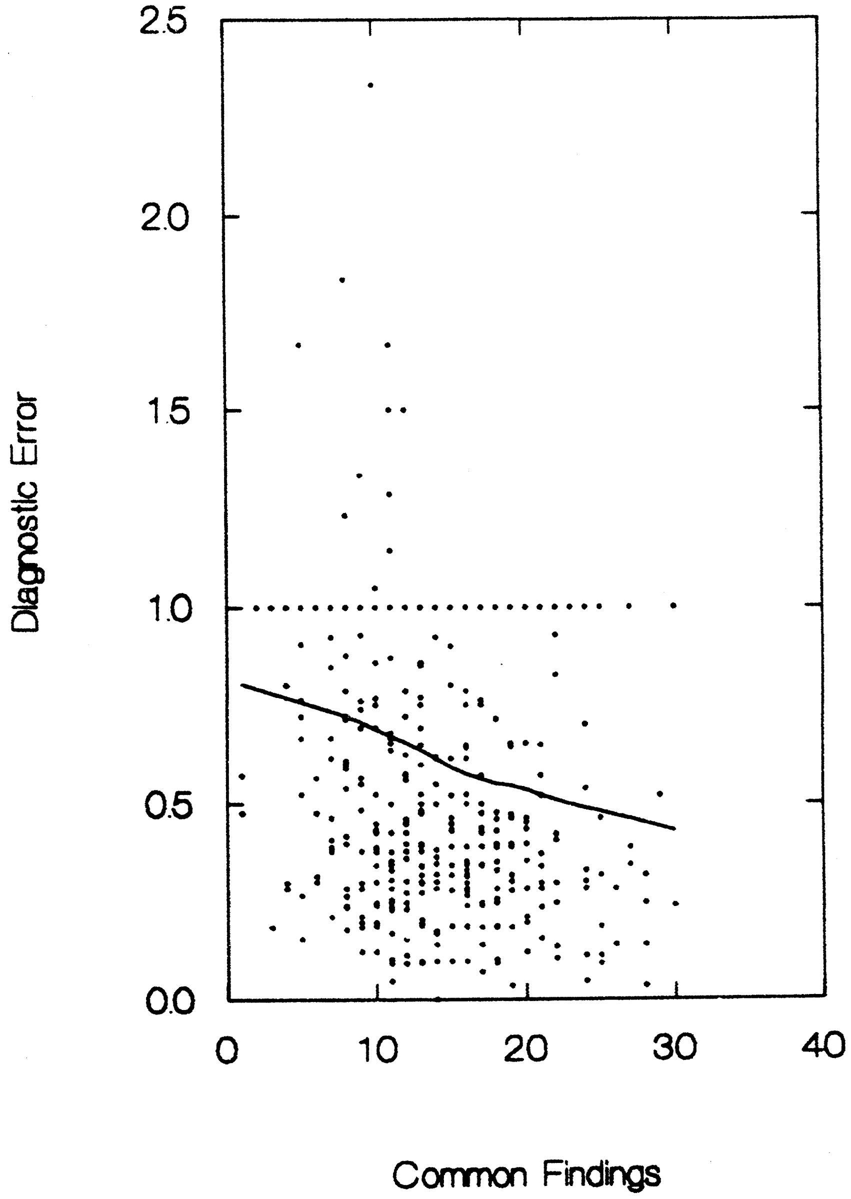


Figure 7.10: Correlation between the number of common findings (M1) and the diagnostic error ϵ

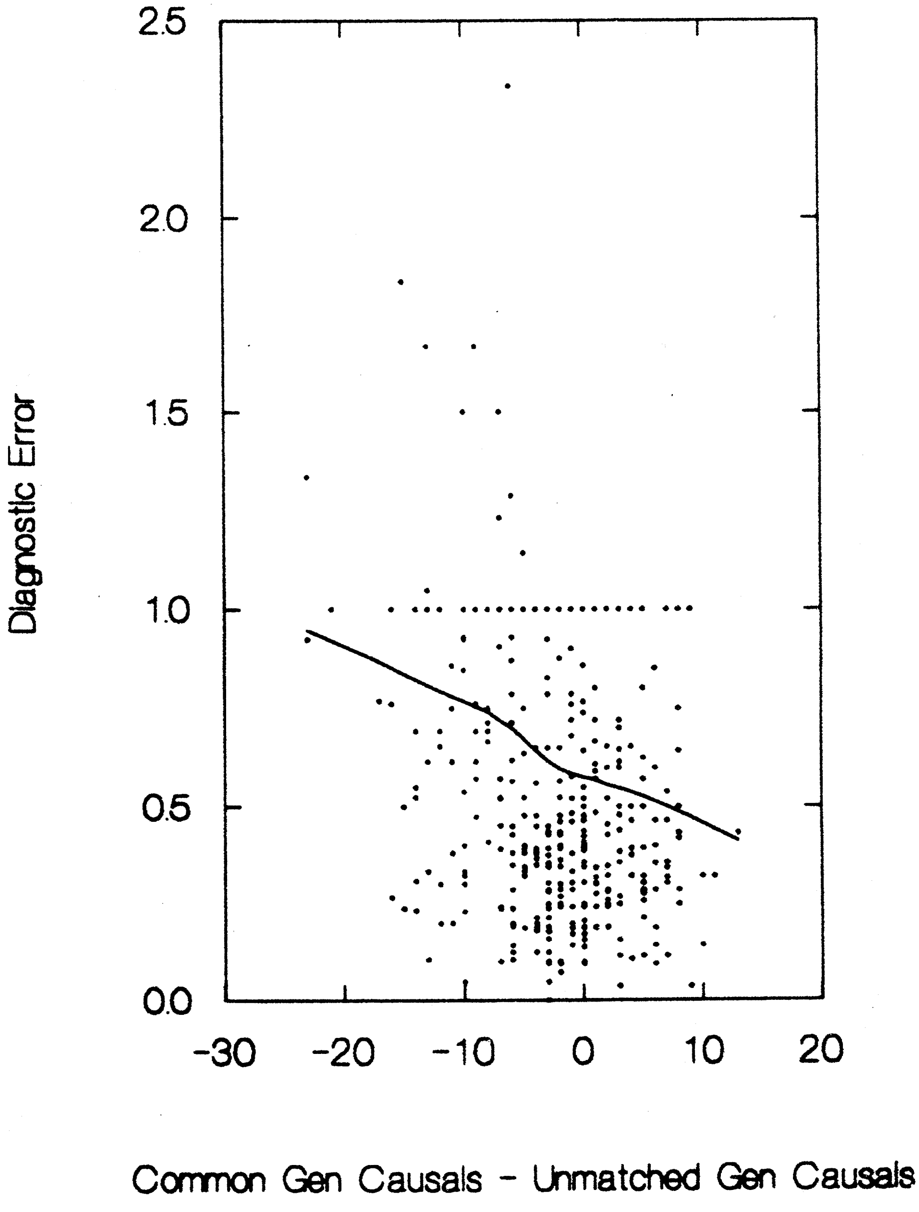


Figure 7.11: Correlation between CASEY's similarity metric (M4) and the diagnostic error ϵ

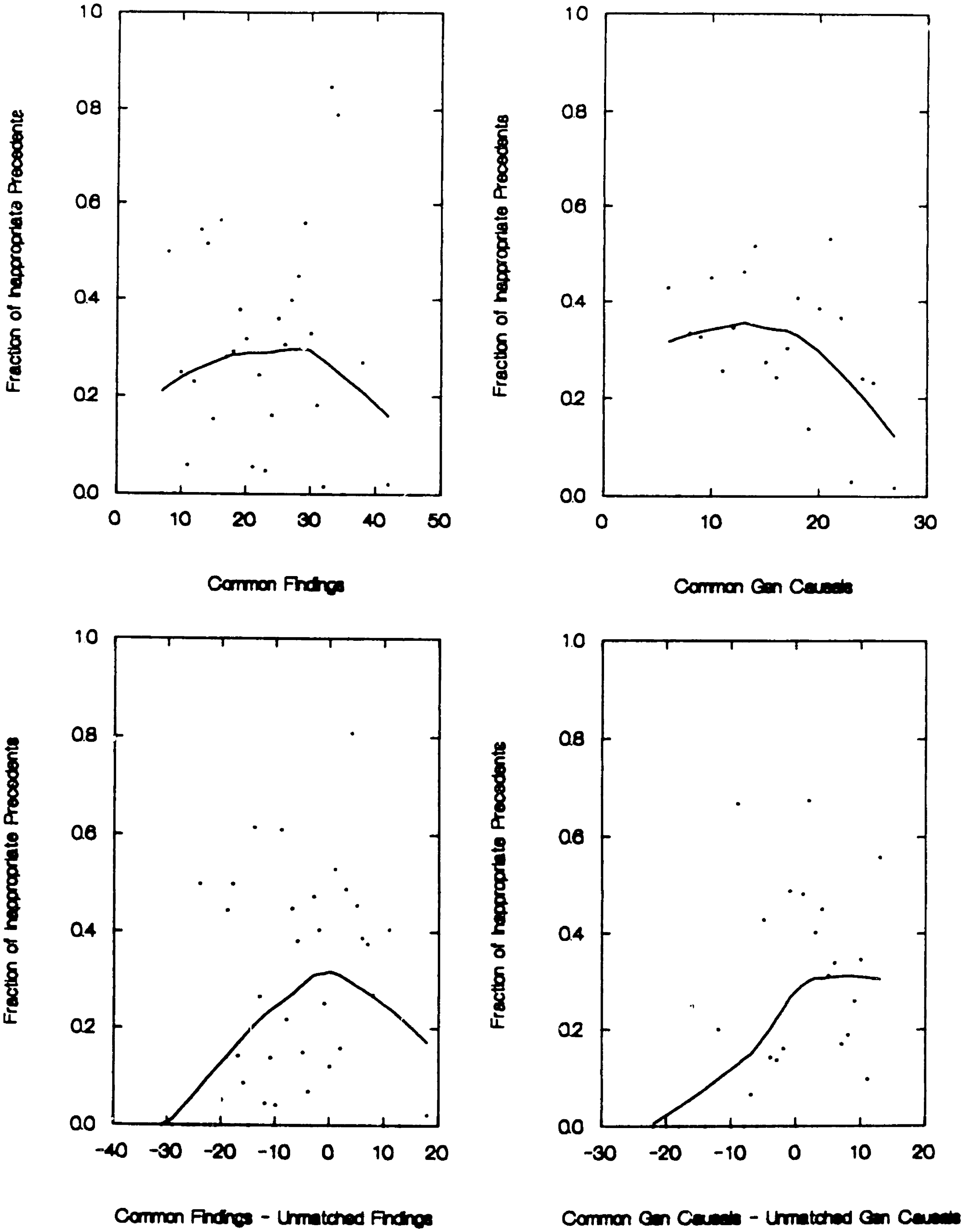


Figure 7.12: Correlations between case similarity and precedent suitability

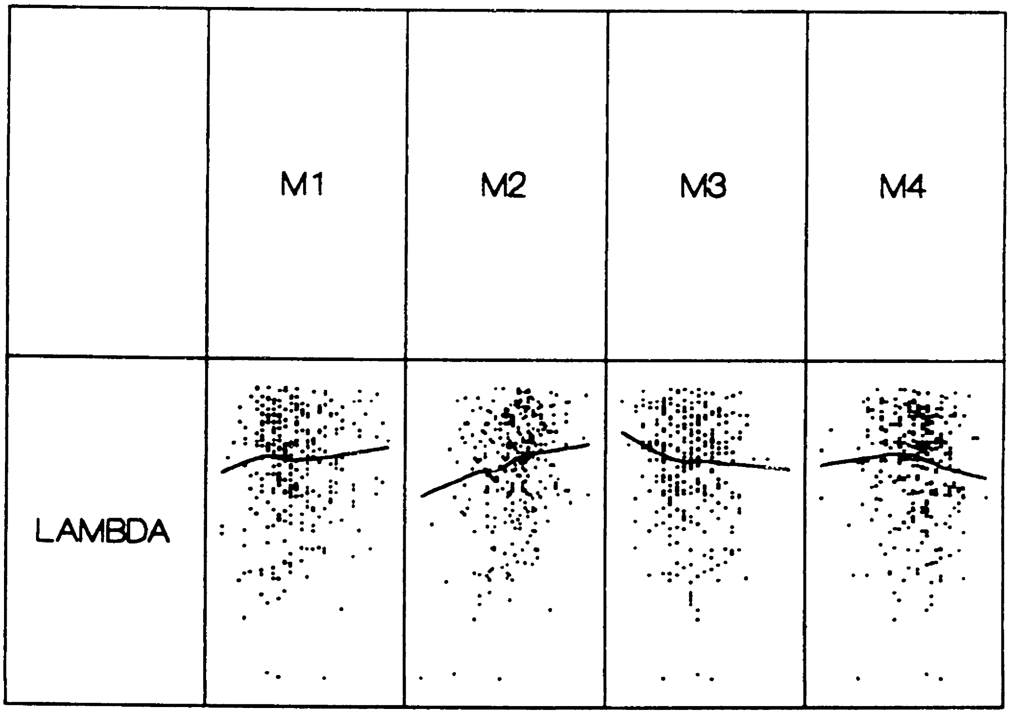


Figure 7.13: Correlations between case similarity and the relative diagnostic likelihood λ

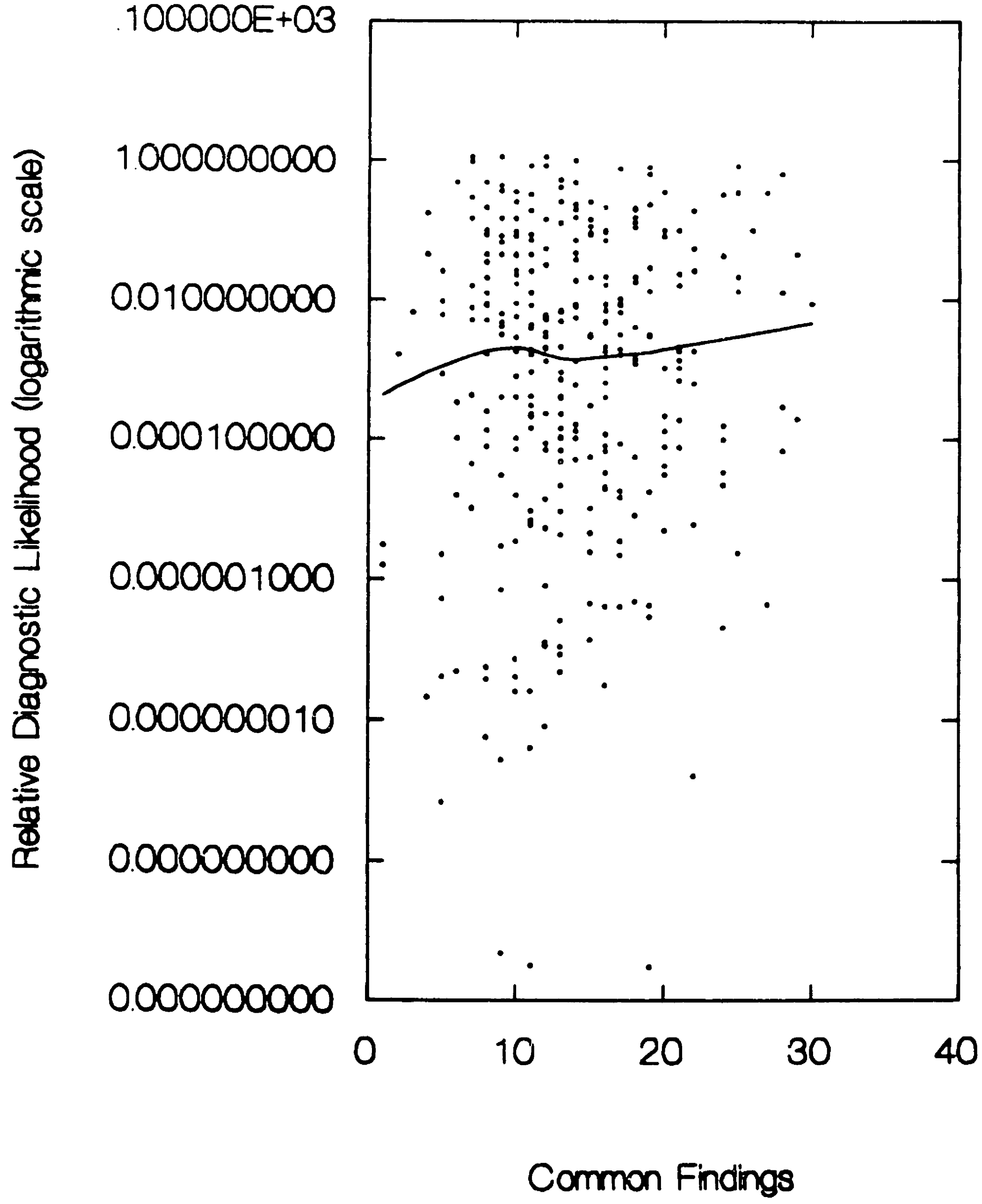


Figure 7.14: Correlation between the number of common findings (M1) and the relative diagnostic likelihood λ

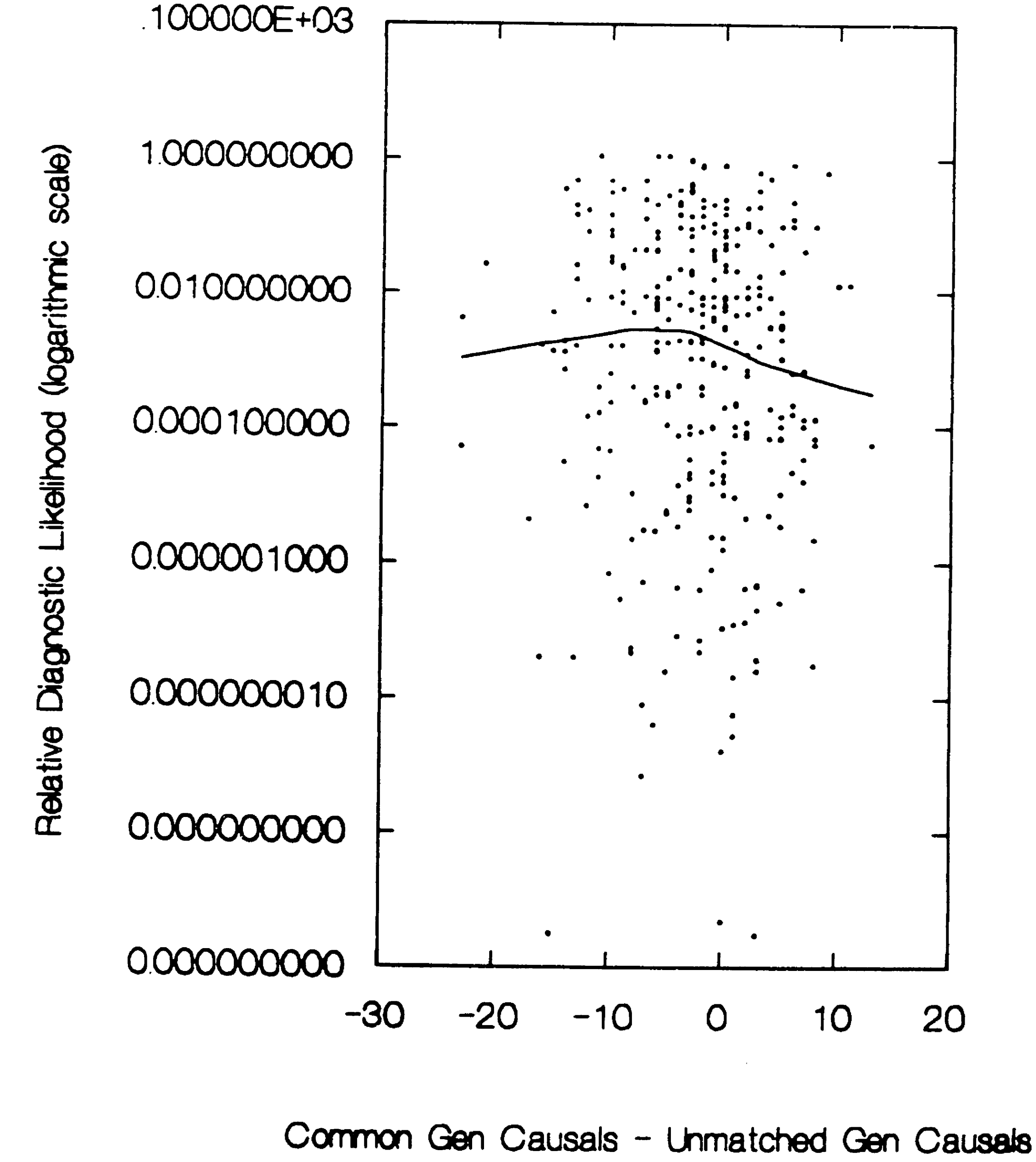


Figure 7.15: Correlation between CASEY's similarity metric (M4) and the relative diagnostic likelihood λ

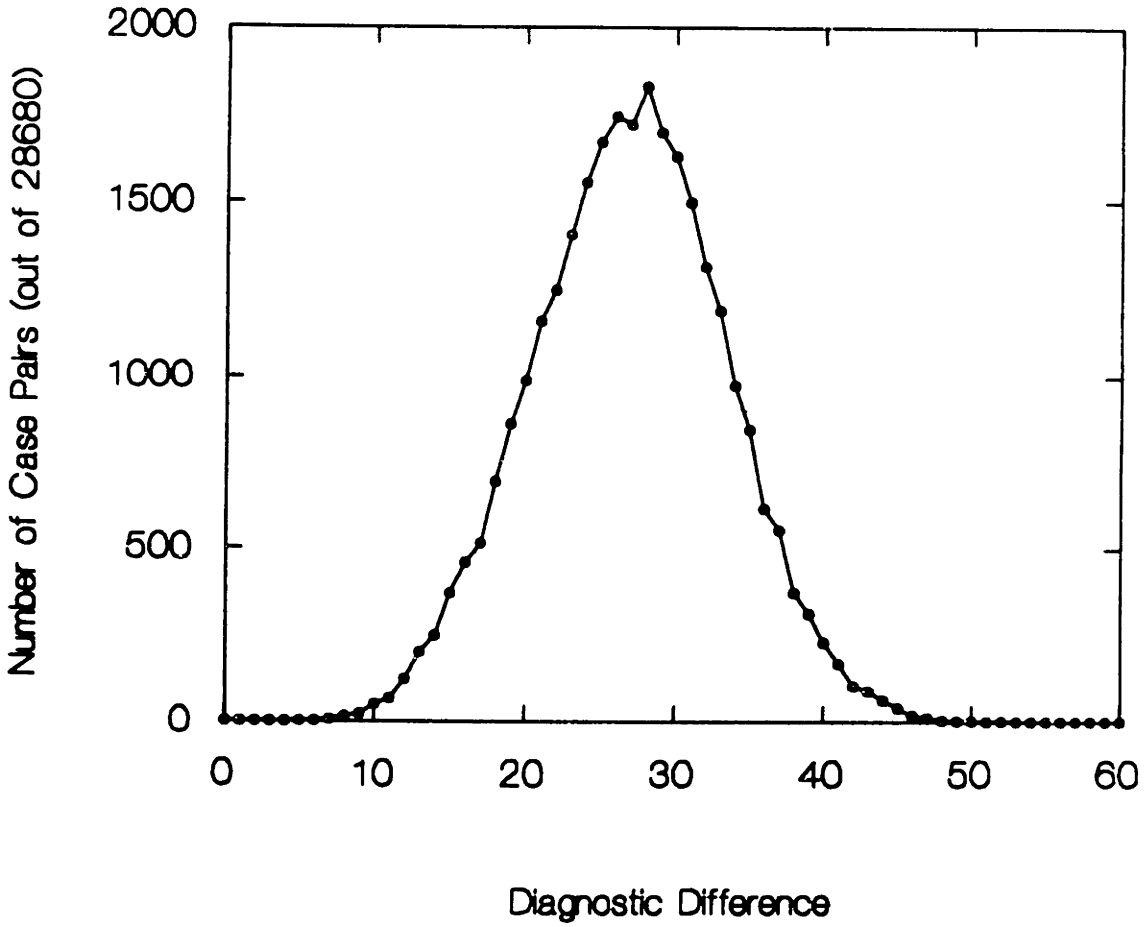


Figure 7.16: Histogram of diagnostic difference, over 26860 case pairs

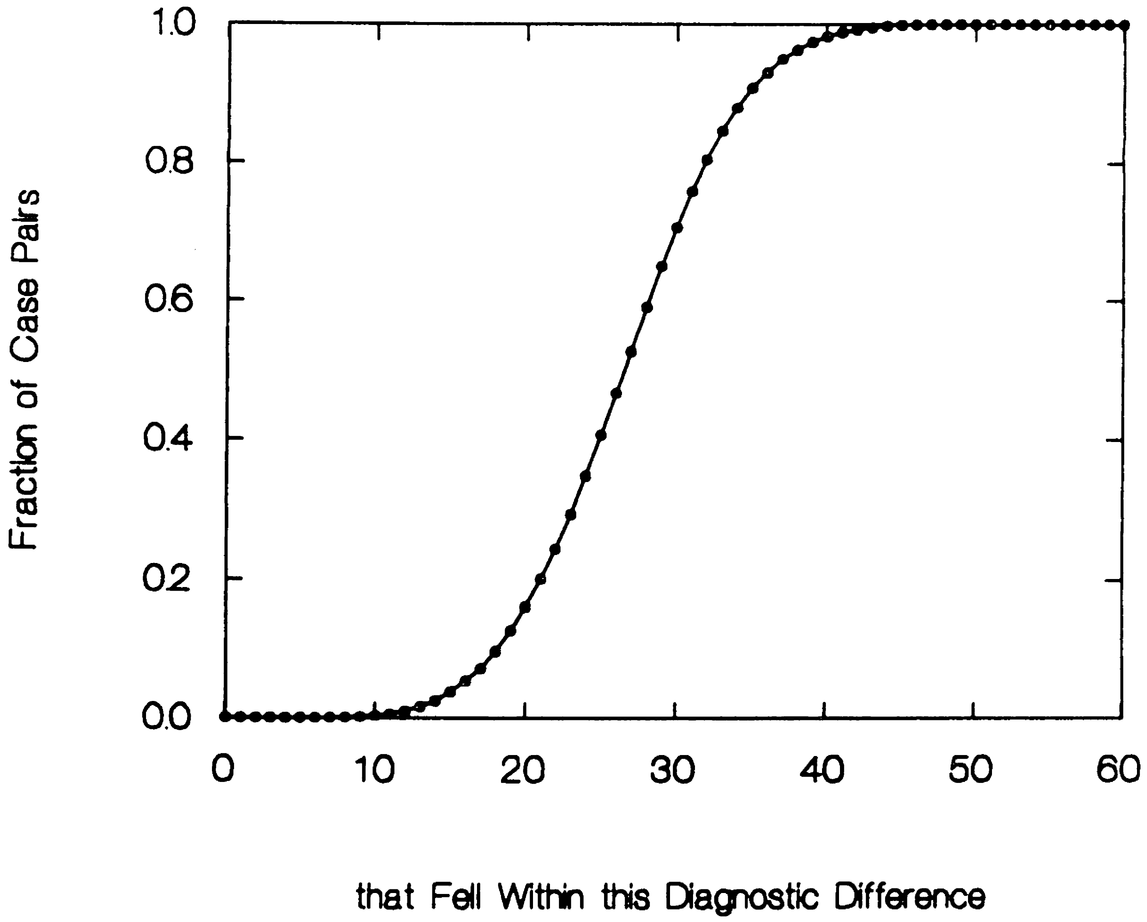


Figure 7.17: Fraction of case pairs achieving the specified diagnostic proximity

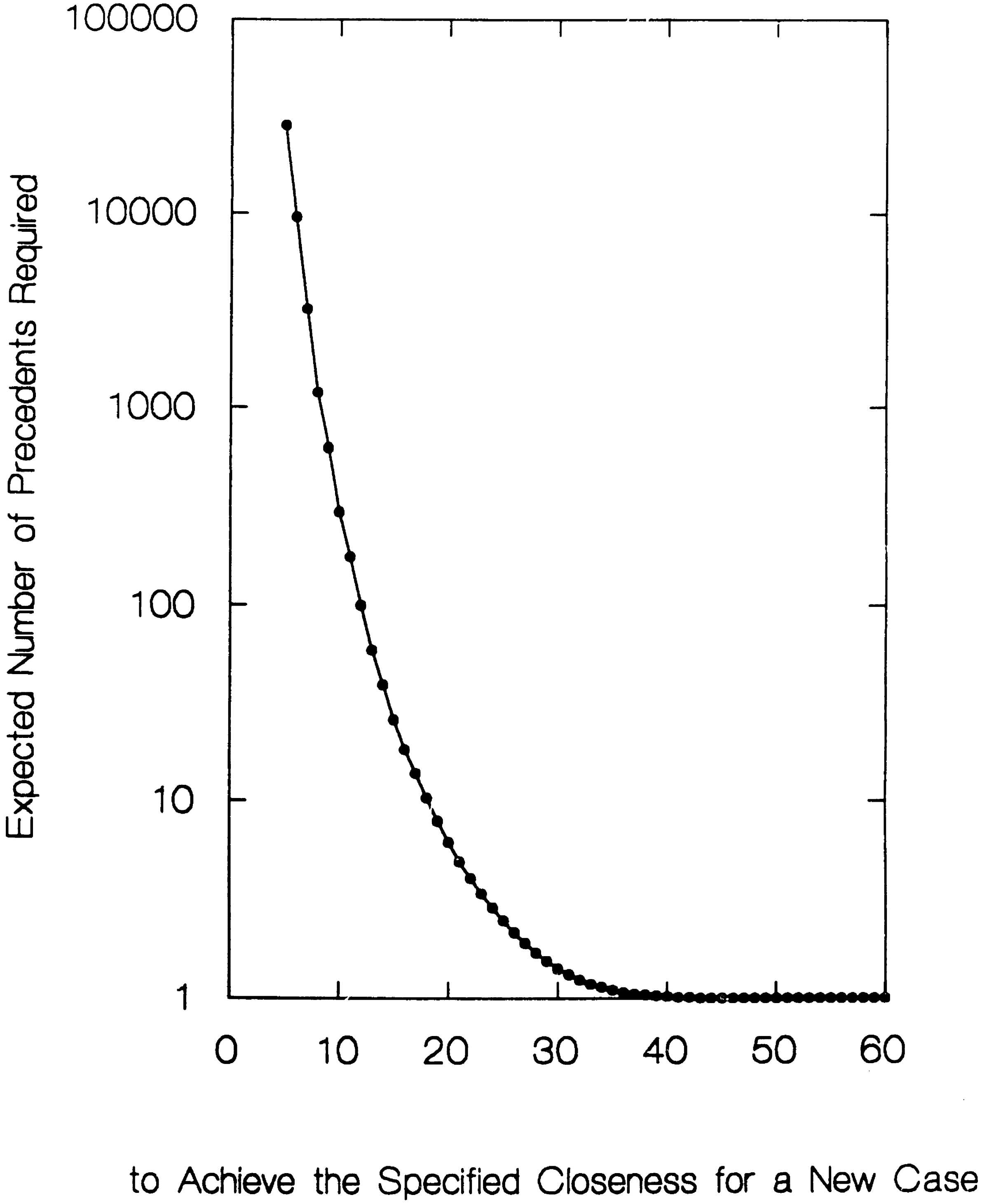


Figure 7.18: Expected number of precedents required to provide a close diagnostic match for the new case

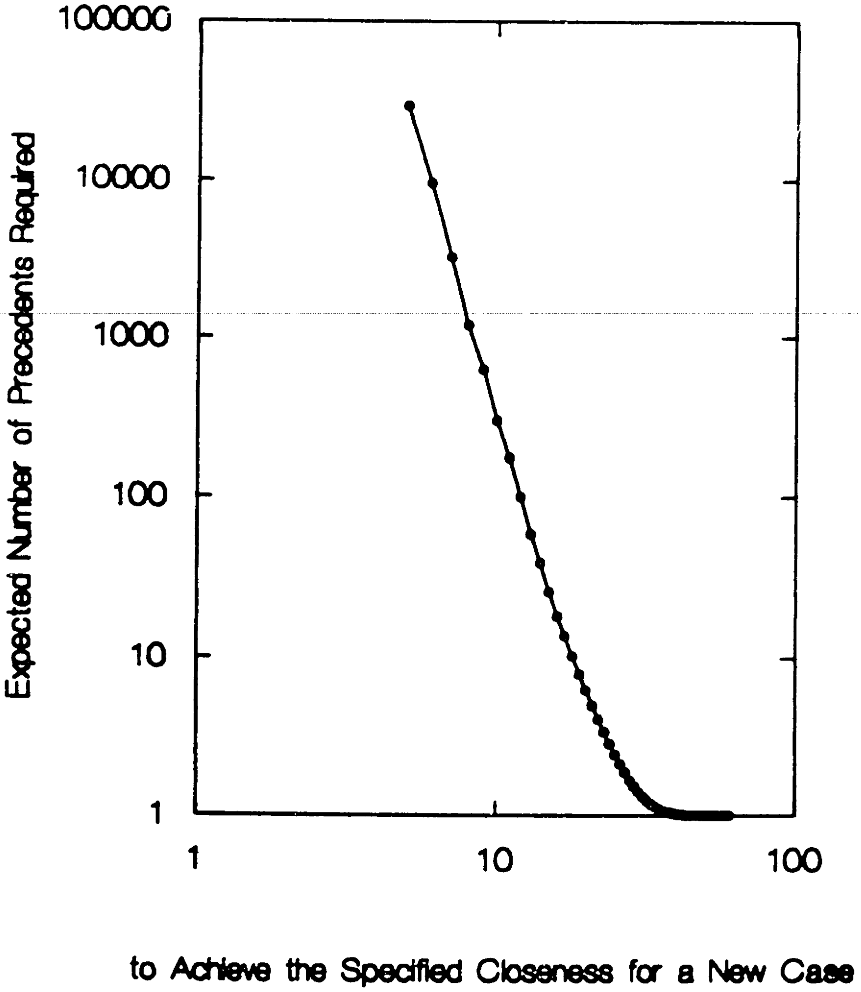


Figure 7.19: Log × log plot of the expected number of precedents required to provide close diagnostic matches

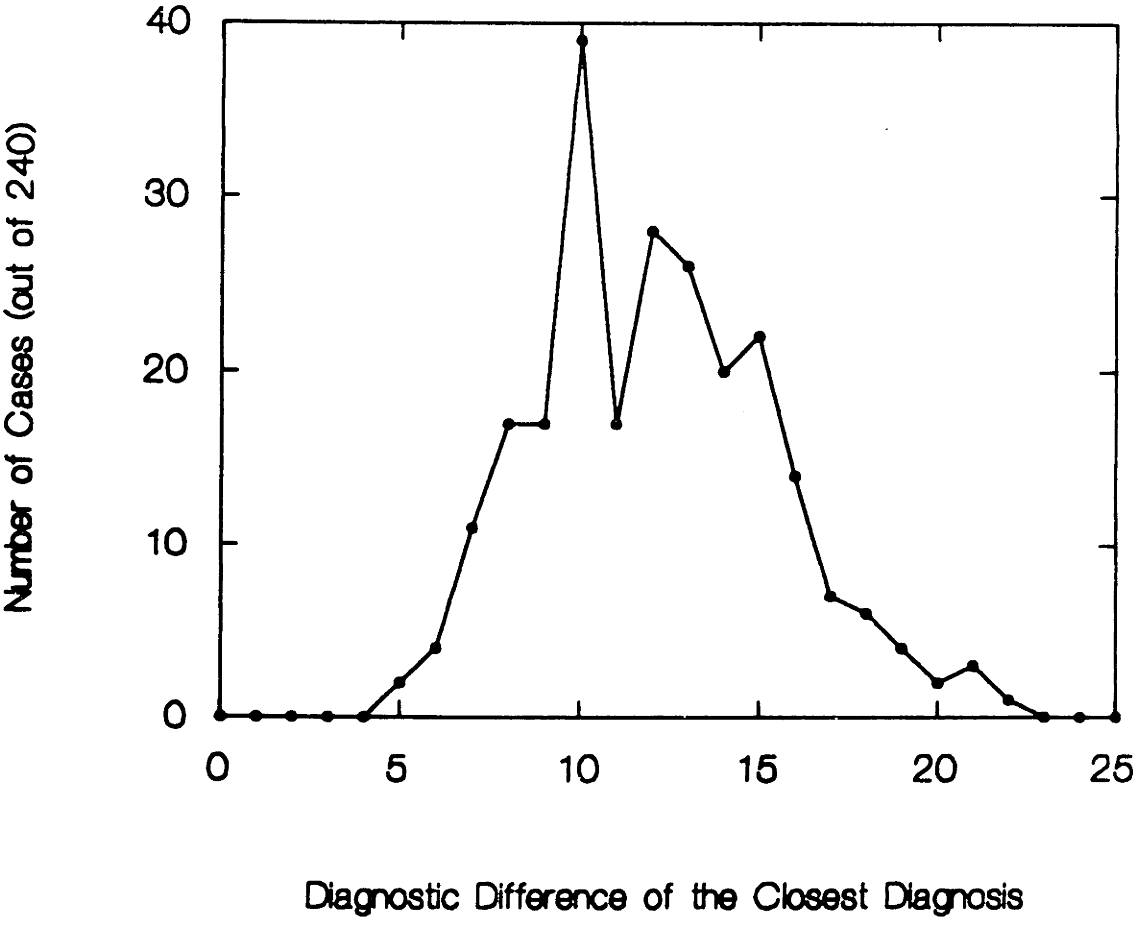


Figure 7.20: Histogram of the diagnostic difference between each of the 240 cases and their closest diagnostic counterparts

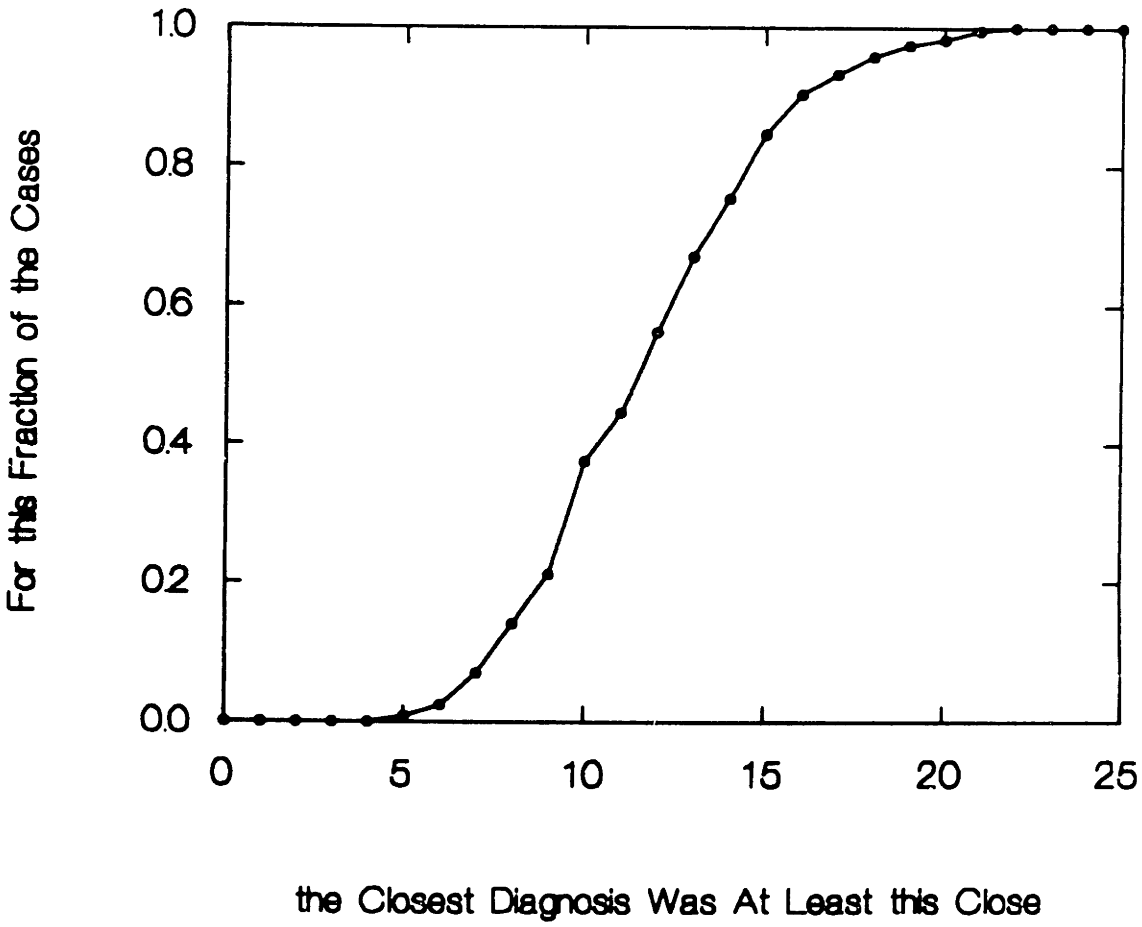


Figure 7.21: Fraction of cases whose closest diagnostic counterparts achieved the specified proximity

Chapter 8

Case Studies

Having discovered the abstract, overall reasons for CBR's inability to master the heart failure domain, we can focus our analysis on concrete, detailed examples. This chapter illustrates the pervasive differences that exist between cases and the effects of this heterogeneity on case-based diagnosis.

8.1 Case Study 1

> PT1060 is an 80 year old white female with known coronary artery disease who is now admitted to the ER with atypical chest-pain. This pain developed the day prior to admission and occurred during several episodes over the last 24 hours. No prolonged episodes associated with nausea, vomiting, or diaphoresis. The pain was relieved in the ER with sublingual nitroglycerin.

> PAST MEDICAL HISTORY: Coronary artery disease, Status post inferior myocardial infarction four years ago, Severe COPD, Hypothyroidism, Hypertension, Status post pacemaker placement
>
> MEDICATIONS ON ADMISSION: Nitroglycerin, Hydralazine, Furosemide, Digitalis
>
> PHYSICAL EXAM: The patient is an elderly female in no acute distress. Blood pressure 130/80. Pulse 80, regular. Afebrile at 98.6. Lungs revealed scattered dry rales at bases. The cardiac exam revealed a regular rhythm with S1, S2, II/VI systolic ejection murmur at the left sternal border. The abdominal exam was normal. Extremities revealed mild pedal edema.
>
> LABORATORY DATA: Na 140, K 4.3, bun 15, creatinine 0.9. CBC within normal limits. EKG revealed a normal sinus rhythm at 75 and first degree AV block.[1]

Now, without glancing back at the previous discharge summary, read the following one:

> This is one of several admissions for PT1128, an 80 year old woman with a history of coronary artery disease and anginal chest-pain. Patient denied diaphoresis, nausea, vomiting, shortness of breath, or syncopal symptoms.

[1] Abbreviated from a New England Medical Center discharge summary dictated by Richard Cottiero, M.D.

> PAST MEDICAL HISTORY: Coronary artery disease, Status post inferior myocardial infarction four years ago, Atrial fibrillation, Severe COPD, Hypothyroidism, Status post pacemaker placement, Hypertension
>
> MEDICATIONS ON ADMISSION: Nitroglycerin, Furosemide, Digitalis
>
> PHYSICAL EXAM: The patient was a pleasant elderly woman in no apparent distress. Blood pressure 132/80, pulse 60 and regular, respiratory rate 18, afebrile at 98.6. Jugular venous pressure 5 cmH20. Lungs: bibasilar rales. Cardiac: regular rate and rhythm, normal S1, S2, II/VI systolic ejection murmur at the left sternal border radiating to the apex and faint S3. Abdomen: normal. Extremities: no edema.
>
> LABORATORY DATA: Na 140, K 4.2, bun 15, creatinine 1.0. CBC: normal. Arterial blood gas on 2 liters: pH 7.39, P02 79, PC02 47. EKG during the evening revealed sinus rhythm with frequent paced beats, lateral inferior ischemia, and an old inferior infarct.[2]

After our cursory perusal of their abbreviated discharge summaries, the two patients PT1060 and PT1128 seem extremely similar, and they are. The Heart Failure system singled them out, from the entire pool of 240 cases, as the two patients with the highest proportion of common abnormal findings. Their commonalities are listed below, abnormal findings in italic.

[2] Abbreviated from a New England Medical Center discharge summary dictated by Amy Kuhlik, M.D.

age 80
female
no nausea/vomiting
no diaphoresis
coronary heart disease, known diagnosis
old myocardial infarction, known diagnosis
hypertension, known diagnosis
copd, known diagnosis
hypothyroidism, known diagnosis
pacemaker
on nitroglycerin
on furosemide
on digitalis
blood pressure $\sim$ 130/80
afebrile at 98.6
appearance in no acute distress
basilar rales
normal pulse
normal S1
normal S2
systolic ejection murmur at the left sternal border, II/VI
normal abdomen
old infarct on ekg
Na 140
K $\sim$ 4.2
bun 15
creatinine $\sim$ 1.0
normal cbc

Even though PT1060 and PT1128 shared a higher proportion of abnormal findings than did any of the other 26859 pairs of cases, they differed by 30%, or 10 abnormal findings. Table 8.1 lists these distinguishing traits.

Figures 8.1 and 8.2 compare their diagnoses, both obtained from the Heart Failure system. Boxes demarcate the distinguishing disease states, while dis-

PT1060	PT1128
atypical chest pain within hours	*history of anginal chest pain*
on hydralazine	*paroxysmal atrial fibrillation, known diagnosis*
mild pedal edema	*low heart rate* 60
first degree block on ekg	*auscultation revealed lv-s3*
	inferior ischemia on ekg
	hypercapnia

Table 8.1: Abnormal findings distinguishing PT1060 and PT1128

tinctive features are circled. The two causal explanations differ by 14 pathophysiologic states, which, relative to the other case pairs, is not a large disparity. However, in absolute terms, 14 states represent a significant difference. Most importantly, PT1128 exhibits only weak evidence for a Myocardial Infarction, whereas PT1060 is clearly experiencing one.[3] Given the mediocrity of what once seemed a perfect match, could CASEY correctly diagnose PT1060 using PT1128 as a precedent? One would certainly hesitate in answering this question affirmatively. Proceeding through the solution transfer process, in detail, will illustrate where case-based reasoning fails.

First, CASEY pruned all precedent pathophysiologic states contradicted by new case findings. Figure 8.3 shows the precedent diagnosis being stripped of the state Low Heart Rate, which was rendered false by PT1060's normal heart rate of 80. Secondly, CASEY used the pruned precedent hypothesis to explain each of the new findings (figure 8.4). Not surprisingly, the common abnormal findings were explained correctly. More interesting, however, is the

[3] Although PT1128's *inferior ischemia on ekg* and *low heart rate 60* are more probably explained by Digitalis, there exists the possibility that she is also experiencing a Myocardial Infarction. The Heart Failure system includes this possibility in its differential diagnosis, but CASEY reasons only from the most likely solution.

fate of PT1060's four distinctive symptoms.

Because the finding *on hydralazine* has only one cause, the presence of the drug HYDRALAZINE in the patient's body, CASEY had no choice but to add this cause to the diagnosis. The system subsequently and accurately explained the *mild pedal edema* as a side effect of this drug. However, because the precedent diagnosis did not include MYOCARDIAL INFARCTION (see figures 8.1 and 8.4), *first degree block* was incorrectly ascribed to the patient's OLD MI, and the system had to posit UNSTABLE ANGINA to explain the *atypical chest pain.*

Finally, CASEY pruned unsupported remnants from the causal explanation. These inappropriate disease states are shown crossed out in figure 8.5, and the final diagnosis appears in figure 8.6.

The most significant difference between CASEY's solution for case PT1060 (figure 8.6) and Heart Failure's ideal solution for the case (figure 8.7), is the conspicuous absence of the diagnosis MYOCARDIAL INFARCTION from CASEY's hypothesis. The patient's recent *atypical chest pain* and the laboratory finding *first degree block* together strongly suggest that an infarction took place. Biased by the precedent diagnostic scheme, however, CASEY attributed these findings to plausible, but less likely causes. Furthermore, when it might have explained the patient's HIGH LA PRESSURE[4] via a high probability pathway from MYOCARDIAL INFARCTION, the system linked it instead, with less likelihood, to the HIGH LV PRESSURE CHRONIC, simply because the precedent hypothesis already contained that disease state.

[4] LA is an abbreviation of LEFT ATRIAL; similarly, LV abbreviates LEFT VENTRICULAR.

Overall, the difference between the two solutions is one of severity rather than one of disparity. However, because MYOCARDIAL INFARCTION is both acute and irreversible, CASEY's somewhat reasonable estimation compares unfavorably to Heart Failure's more discriminating accuracy. The case-based solution contained relative error of $\epsilon = 0.35$ and was estimated to be almost four times less likely than the model-based ideal. Accordingly, PT1060 and PT1128 were not similar enough.

8.2 Case Study 2

Considering CASEY's unexceptional performance with respect to two cases *chosen* for their similarity, we understand why its average accuracy is so poor. This case study presents one of the system's more serious blunders.

CASEY diagnosed PT1033 with a case base of 200 patients at its disposal. The best matched precedent that the system could find was PT1097. Assuredly, the two cases shared several common findings, which are listed below, abnormal findings in italic.

orthopnea
hypertension, known diagnosis
blood pressure ~ 115/75
afebrile at ~ 98.6
rales
auscultation revealed lv-s3
normal abdomen
mild pedal edema
hypocapnia

However, the two patients were also markedly different. Table 8.2 itemizes

24 distinguishing abnormal findings.

PT1033	PT1097
nocturnal dyspnea	*dyspnea on exertion*
history of anginal chest pain	*copd, known diagnosis*
anginal chest pain within hours	*on furosemide*
high heart rate 110	*on digitalis*
systolic ejection murmur	*on bronchodilator*
atrial fibrillation on ekg	*wheezes*
inferior ischemia on ekg	*distended jugular venous pressure* 12
generalized cardiac enlargement on xray	*holosystolic murmur*
congestive failure on chest xray	*left bundle branch block on ekg*
pleural effusion on chest xray	*left ventricular cardiac enlargement on xray*
high bun level 21	*vascular redistribution on chest xray*
	low sodium 134
	low potassium 3.6

Table 8.2: Abnormal findings distinguishing PT1033 and PT1097

As a result, their diagnoses have very little in common. Figures 8.8 and 8.9 illustrate the differences by boxing distinguishing disease states and circling distinctive findings. Figures 8.10 through 8.12 detail a highly inappropriate solution transfer.

Comparing CASEY's hypothesis for PT1033 (figure 8.13) with the Heart Failure ideal (figure 8.14), we notice a fundamental diagnostic difference between the aortic disease underlying the ideal explanation and the mitral disease derived from the precedent and used in CASEY's hypothesis. Although PT1097's *holosystolic murmur* clearly indicated Mitral Regurgitation Chronic, PT1033's *systolic ejection murmur* was considerably more characteristic of Aortic Stenosis. However, the precedent scheme induced CASEY to impute the murmur to its less likely cause. Consequently, the *history of* *anginal ·hest pain*, which should have been ascribed to EXERTIONAL ANGINA stemming indirectly from the AORTIC STENOSIS, was ascribed instead to a rather improbable CORONARY SPASM. The unsuitable precedent hypothesis was also at fault for offering LV HYPERTROPHY to explain a *generalized cardiac enlargement.* Furthermore, COPD OR CHRONIC BRONCHITIS, no longer a known diagnosis, remained in the hypothesis in order to rationalize TACHYPNEA and, indirectly, HIGH BLOOD VOLUME, both of which could have been imputed to more probable physiology.

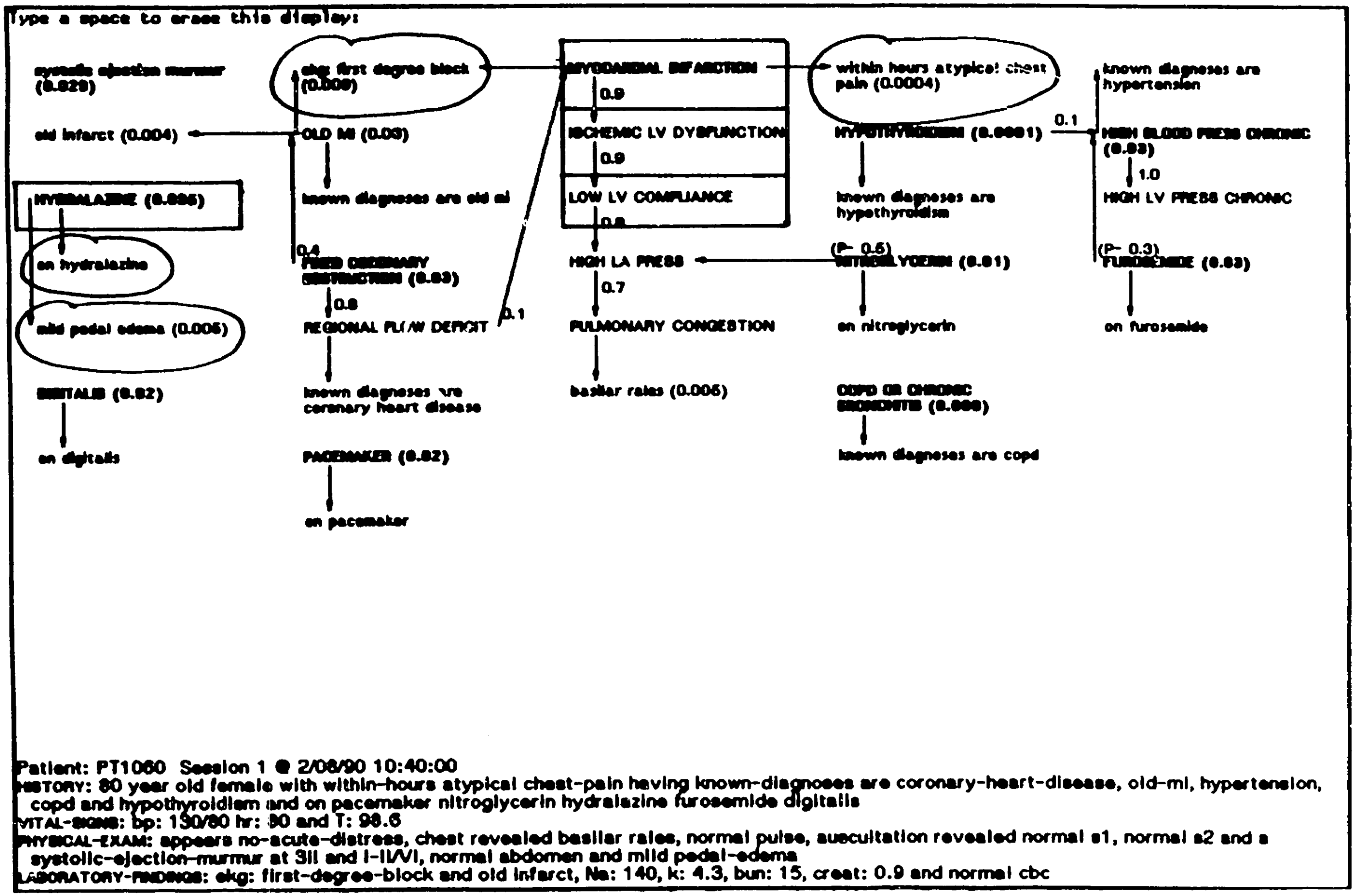
Thus, even CASEY's best match for PT1033 failed disastrously, yielding a relative error of $\epsilon = 0.68$ and a relative likelihood of nearly 10^{-13}.

8.3 Case Study 3

How close must two solutions be in order to justify transfer? Figures 8.15 and 8.16 highlight the six pathophysiologic states that differentiate two of the closest solutions in the case pool, PT1205 and PT1192. Although transfer from precedent PT1192 to new case PT1205 (detailed in figures 8.17 through 8.19) is tenable, it is by no means perfect.

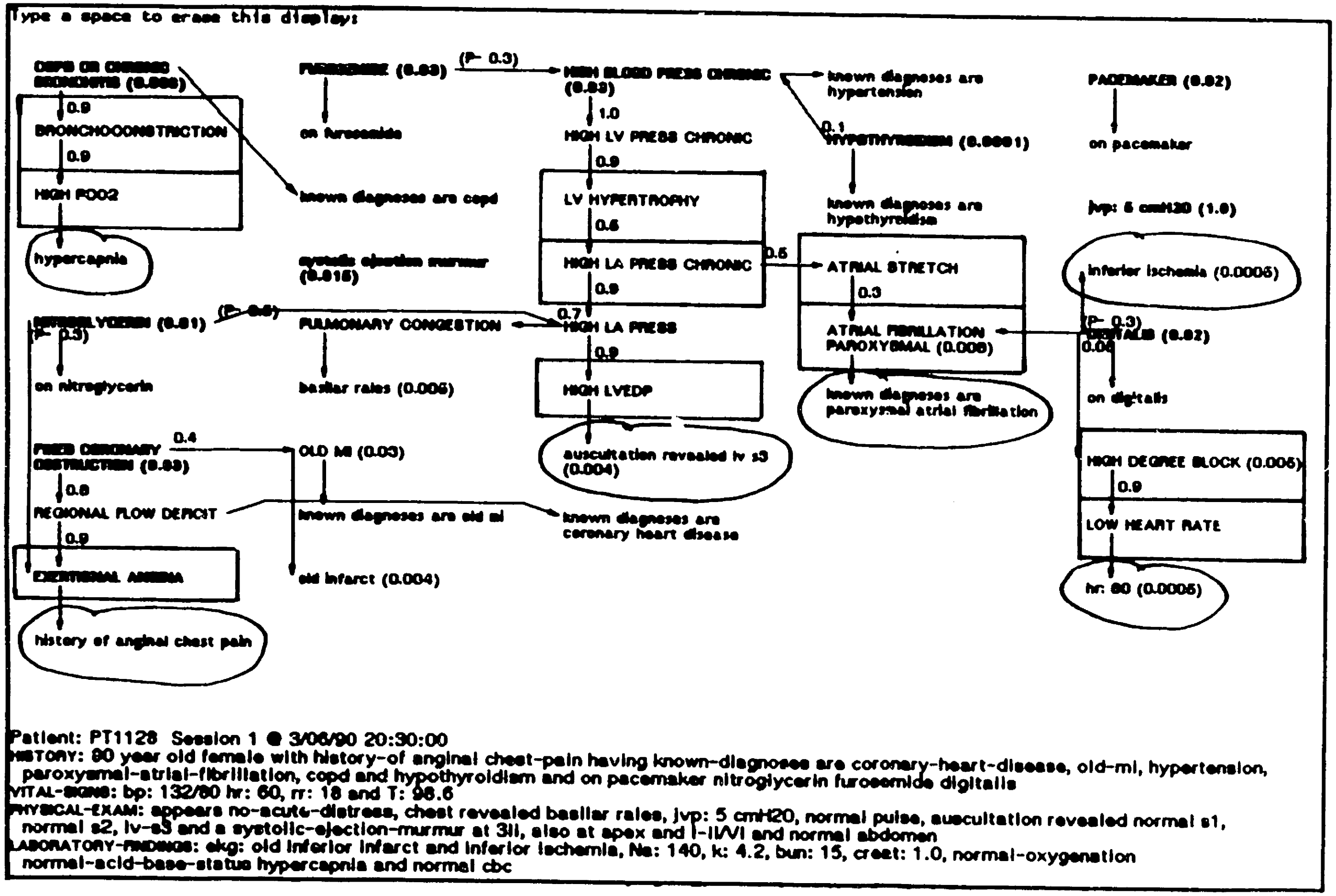
CASEY's solution, shown in figure 8.20, retains MITRAL REGURGITATION CHRONIC as a remnant from the precedent hypothesis, even though PT1205 does not share PT1192's strong support for this pathophysiologic state. Using the inappropriate and unlikely disease to explain CARDIAC DILATATION and LOW CARDIAC OUTPUT, CASEY allowed the more probable causes LOW LV SYSTOLIC FUNCTION CHRONIC and LOW LV EMPTYING to be pruned from the causal explanation. Consequently, the case-based solution was judged to be almost a hundred times less likely than the model-based ideal (figure 8.21), despite a relative error of only 0.13.

If CASEY must locate precedents whose diagnoses differ from that of the new case in no more than *five* pathophysiologic states, in order to derive truly accurate solutions, then by figure 7.18, it will require a case base with nearly a hundred thousand patients. Could the reasoning from experience that human beings rely upon so heavily be so infeasible on a computer?



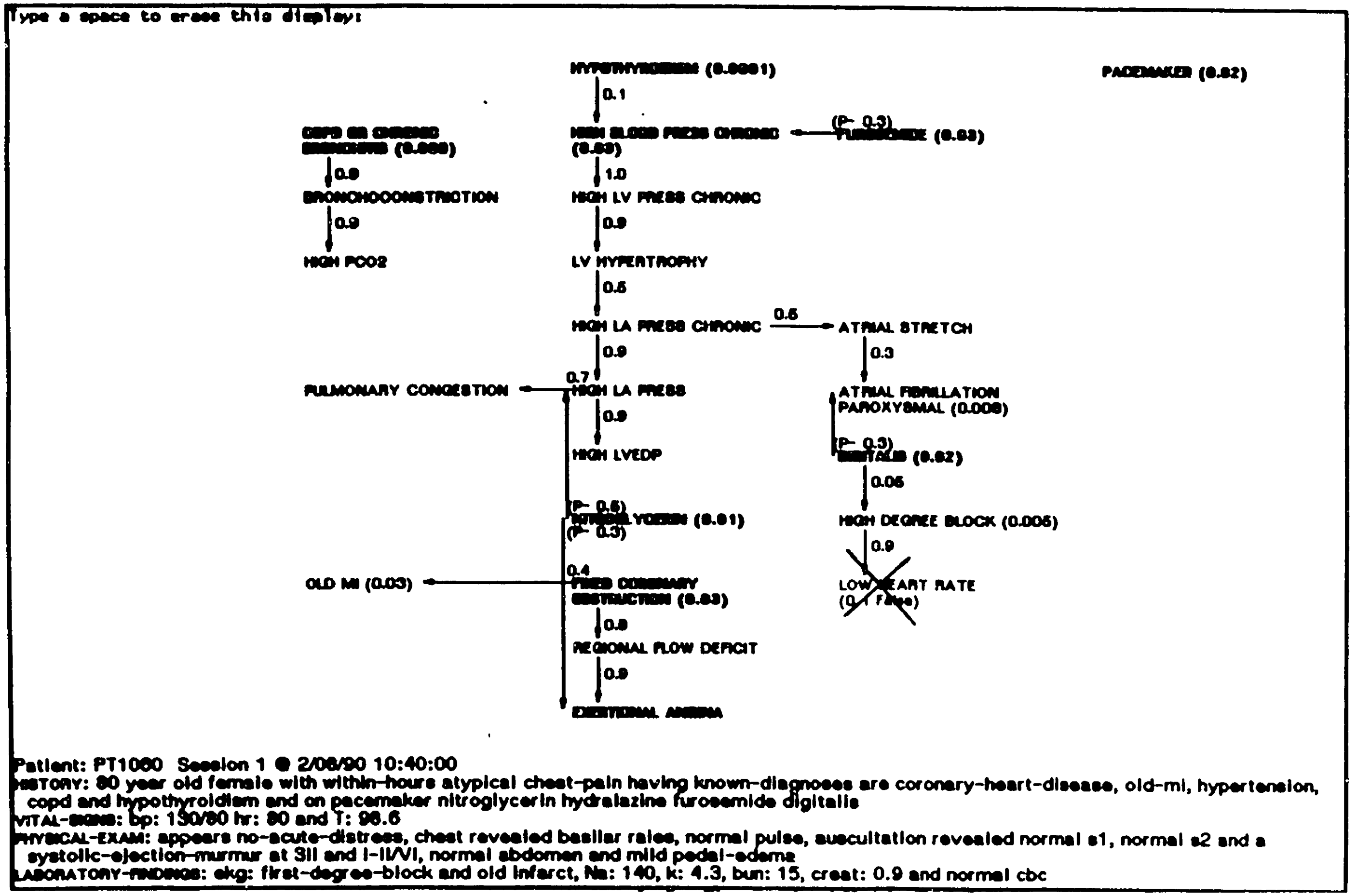
Patient: PT1060 Session 1 @ 2/08/90 10:40:00
HISTORY: 80 year old female with within-hours atypical chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension, copd and hypothyroidism and on pacemaker nitroglycerin hydralazine furosemide digitalis
VITAL-SIGNS: bp: 130/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, normal pulse, auscultation revealed normal s1, normal s2 and a systolic-ejection-murmur at 3il and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: first-degree-block and old infarct, Na: 140, k: 4.3, bun: 15, creat: 0.9 and normal cbc

Figure 8.1: Diagnosis for case PT1060



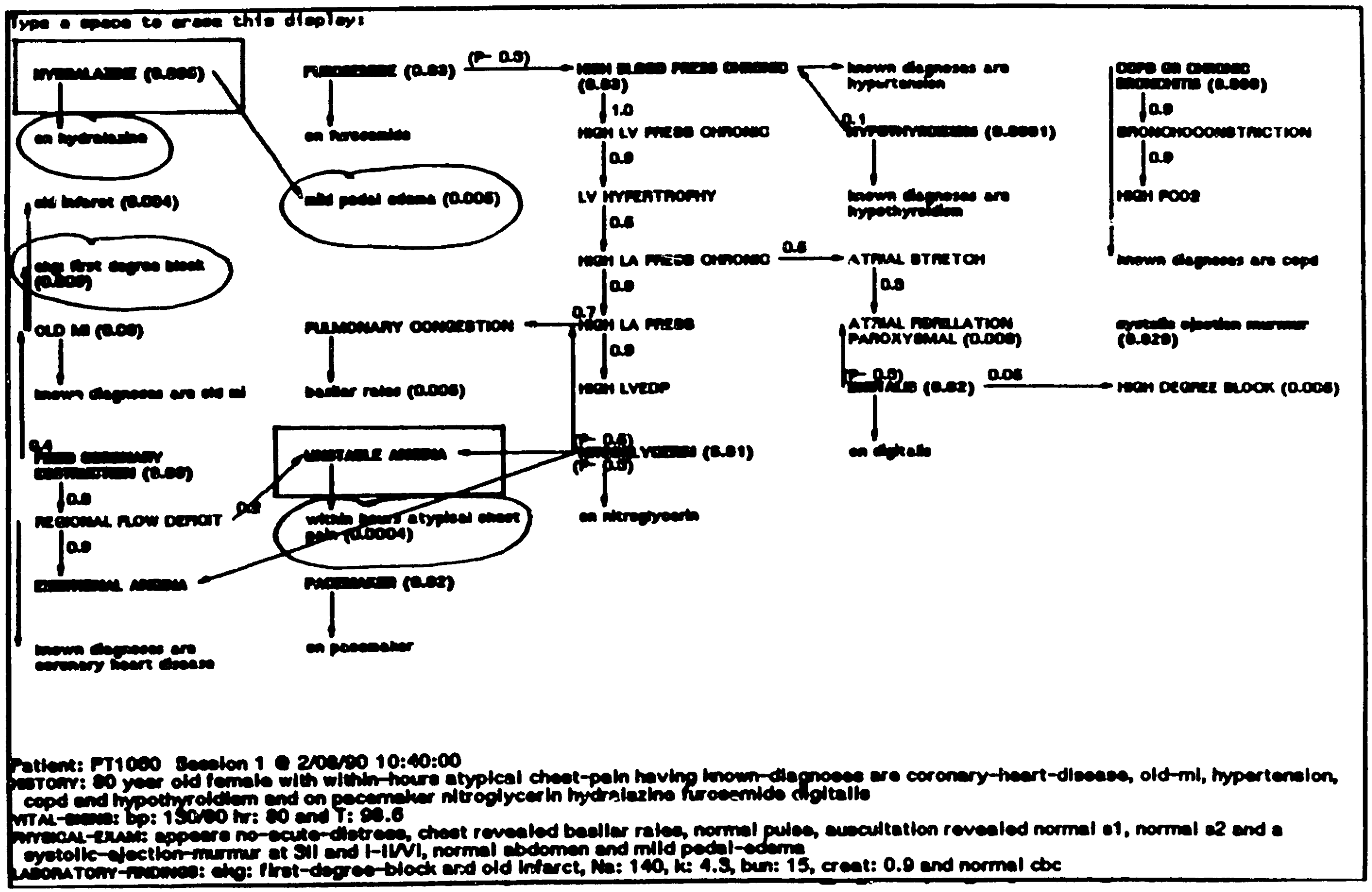
Patient: PT1128 Session 1 @ 3/06/90 20:30:00
HISTORY: 90 year old female with history-of anginal chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension, paroxysmal-atrial-fibrillation, copd and hypothyroidism and on pacemaker nitroglycerin furosemide digitalis
VITAL-SIGNS: bp: 132/80 hr: 60, rr: 18 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, normal pulse, auscultation revealed normal s1, normal s2, lv-s3 and a systolic-ejection-murmur at 3ll, also at apex and l-II/VI and normal abdomen
LABORATORY-FINDINGS: ekg: old inferior infarct and inferior ischemia, Na: 140, k: 4.2, bun: 15, creat: 1.0, normal-oxygenation normal-acid-base-status hypercapnia and normal cbc

Figure 8.2: Diagnosis for case PT1128



Patient: PT1060 Session 1 @ 2/08/90 10:40:00
HISTORY: 80 year old female with within-hours atypical chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension, copd and hypothyroidism and on pacemaker nitroglycerin hydralazine furosemide digitalis
VITAL-SIGNS: bp: 130/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, normal pulse, auscultation revealed normal s1, normal s2 and a systolic-ejection-murmur at 3ll and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: first-degree-block and old infarct, Na: 140, k: 4.3, bun: 15, creat: 0.9 and normal cbc

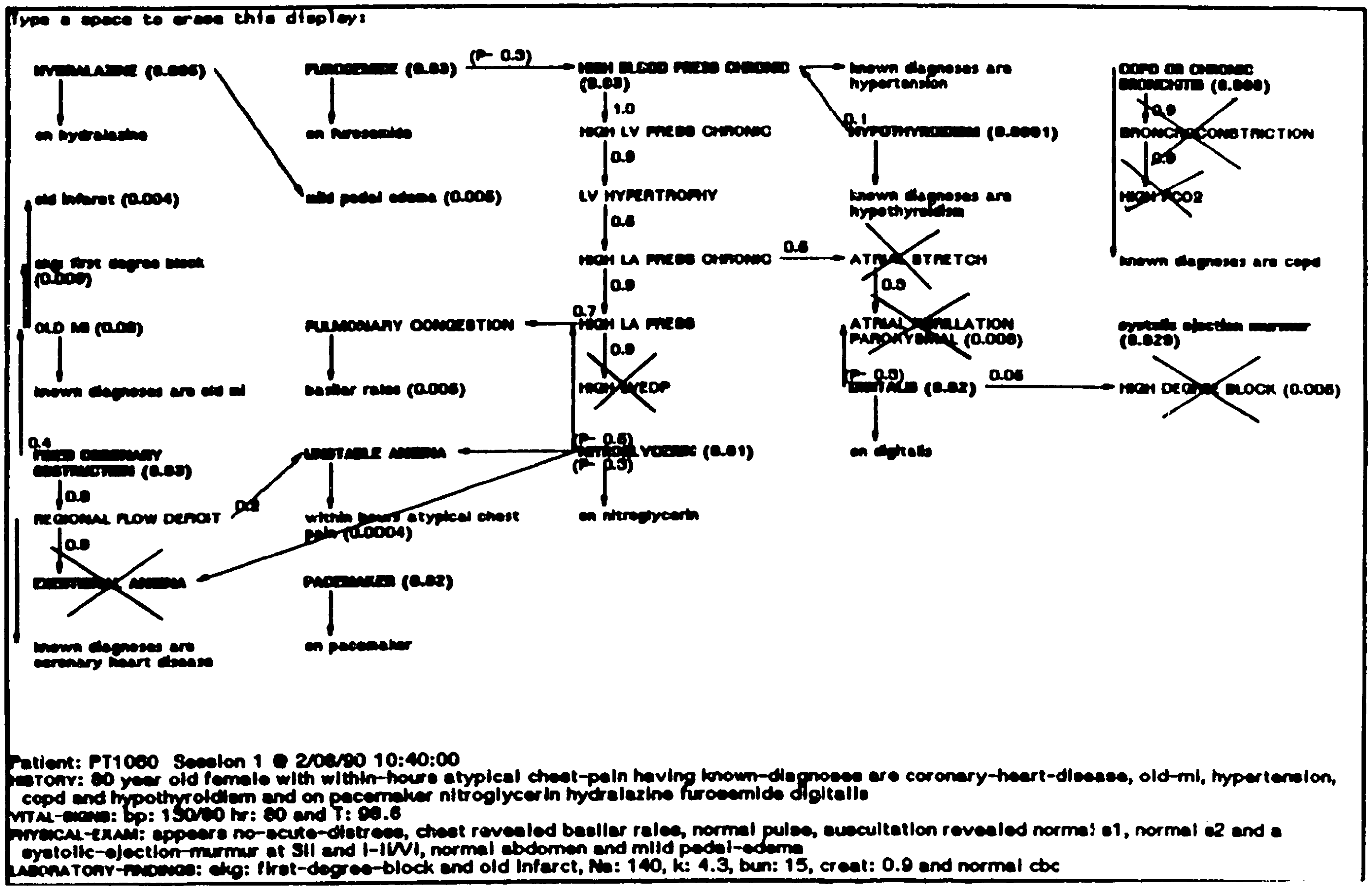
Figure 8.3: Transfer PT1128 ⇒ PT1060: pruning false states



Type a space to erase this display:

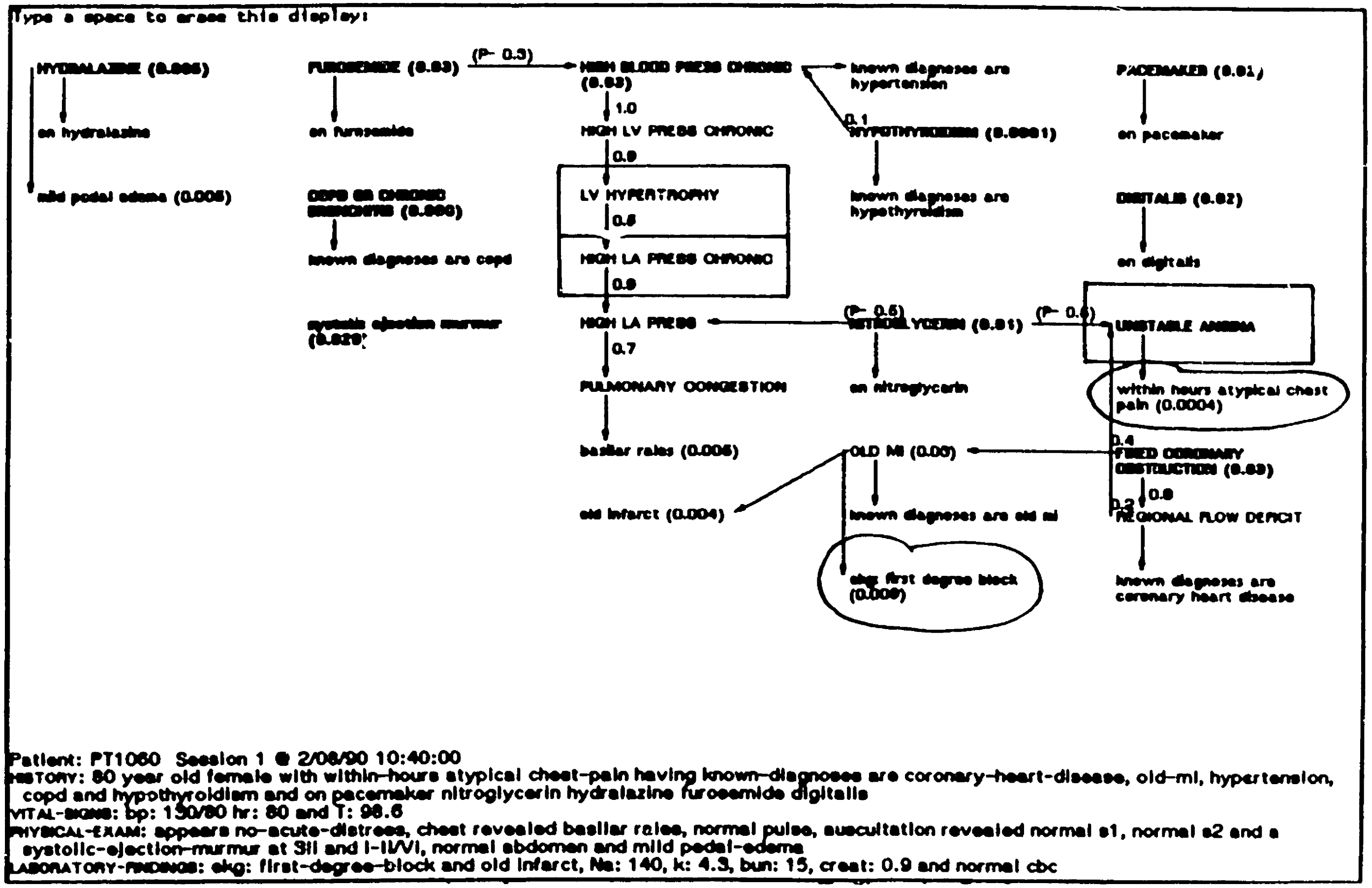
Patient: PT1060 Session 1 @ 2/08/90 10:40:00
HISTORY: 80 year old female with within-hours atypical chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension, copd and hypothyroidism and on pacemaker nitroglycerin hydralazine furosemide digitalis
VITAL-SIGNS: bp: 130/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, normal pulse, auscultation revealed normal s1, normal s2 and a systolic-ejection-murmur at 3il and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: first-degree-block and old infarct, Na: 140, k: 4.3, bun: 15, creat: 0.9 and normal cbc

Figure 8.4: Transfer PT1128 ⇒ PT1060: explaining new findings



```
Patient: PT1060  Session 1 @ 2/08/90 10:40:00
HISTORY: 80 year old female with within-hours atypical chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension,
  copd and hypothyroidism and on pacemaker nitroglycerin hydralazine furosemide digitalis
VITAL-SIGNS: bp: 130/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, normal pulse, auscultation revealed normal s1, normal s2 and a
  systolic-ejection-murmur at SII and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: first-degree-block and old infarct, Na: 140, k: 4.3, bun: 15, creat: 0.9 and normal cbc
```

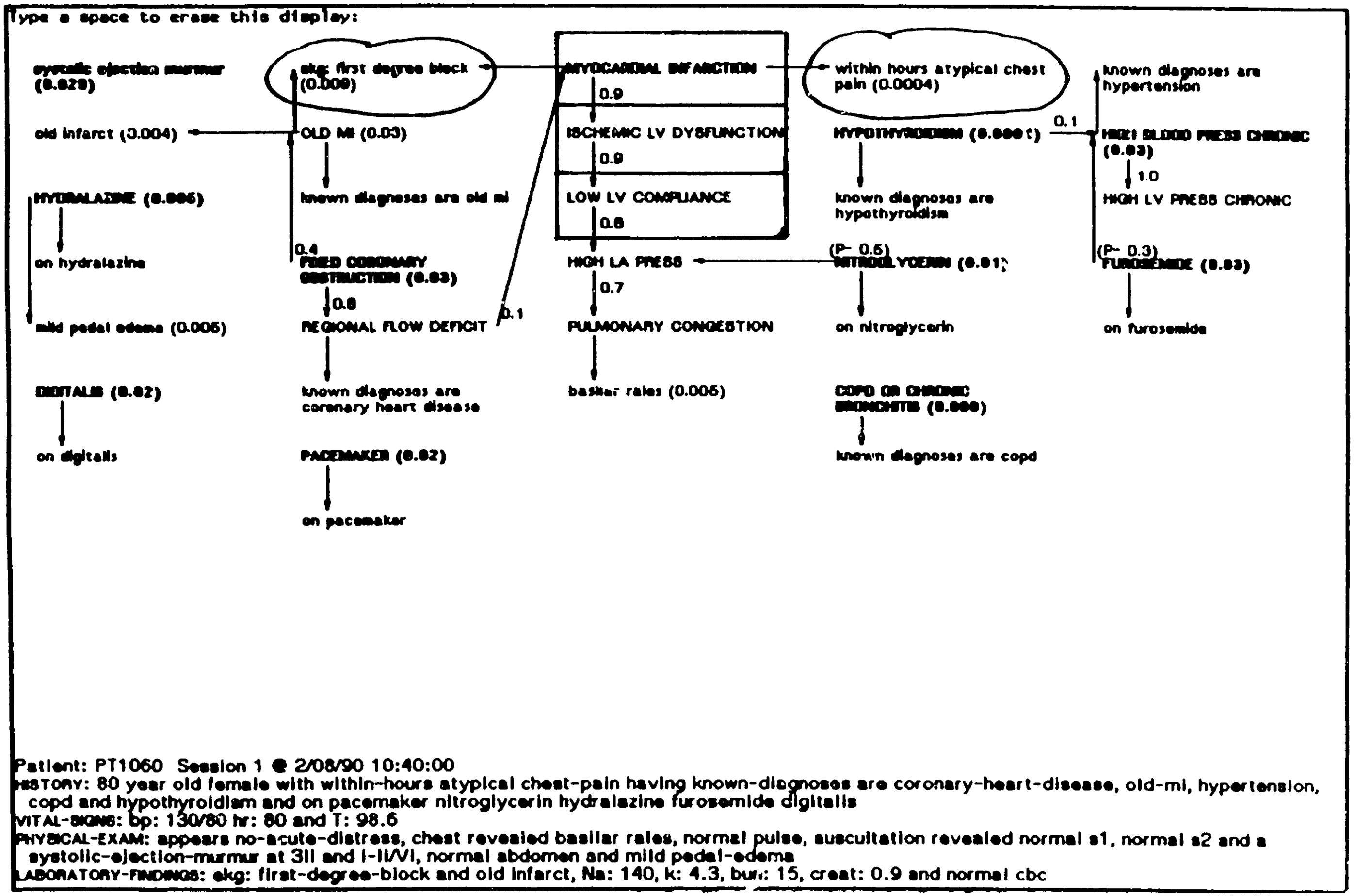
Figure 8.5: Transfer PT1128 ⇒ PT1060: pruning unnecessary states



Type a space to erase this display:

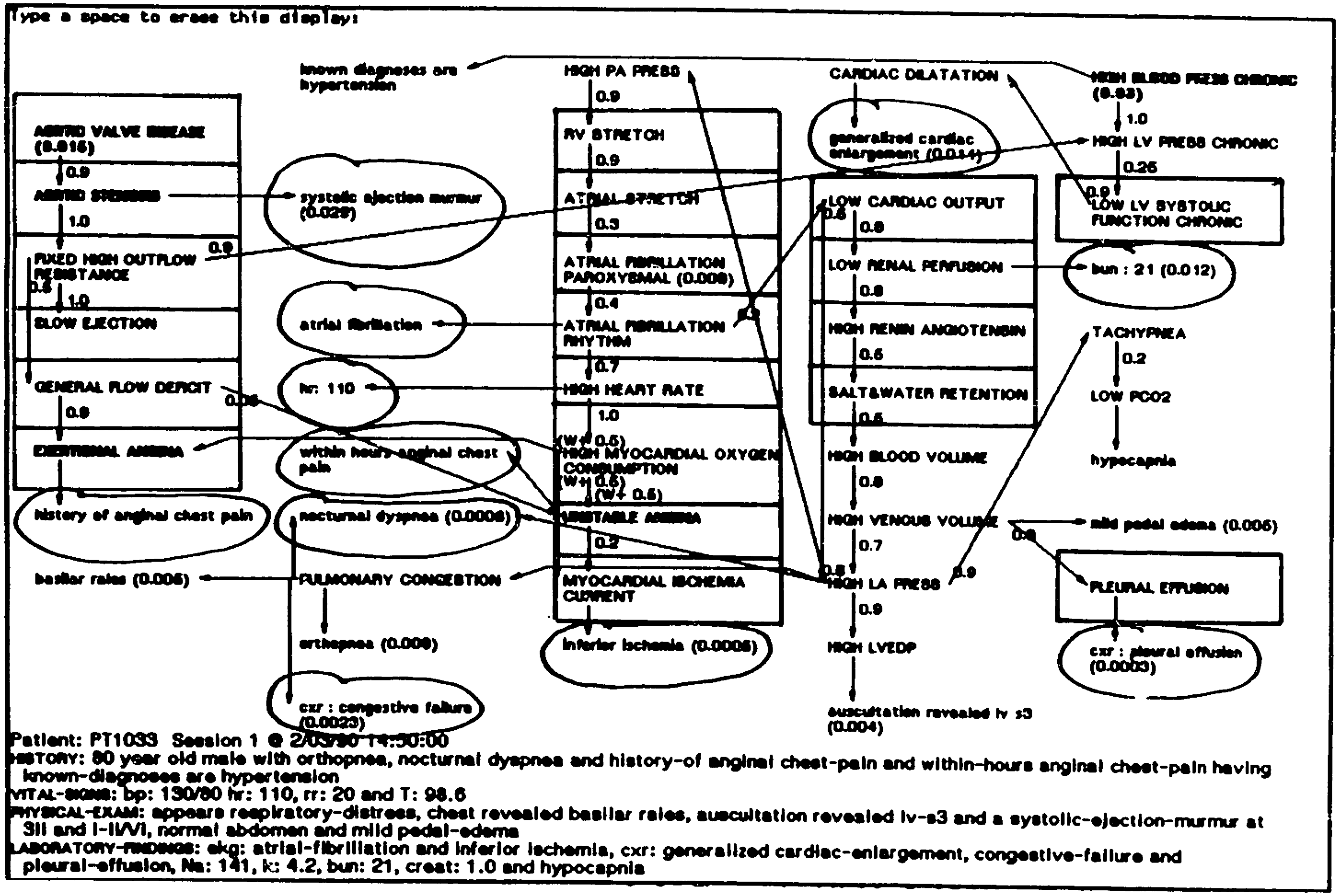
Patient: PT1060 Session 1 @ 2/08/90 10:40:00
HISTORY: 80 year old female with within-hours atypical chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension, copd and hypothyroidism and on pacemaker nitroglycerin hydralazine furosemide digitalis
VITAL-SIGNS: bp: 150/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, normal pulse, auscultation revealed normal s1, normal s2 and a systolic-ejection-murmur at SII and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: first-degree-block and old infarct, Na: 140, k: 4.3, bun: 15, creat: 0.9 and normal cbc

Figure 8.6: CASEY's solution for PT1060



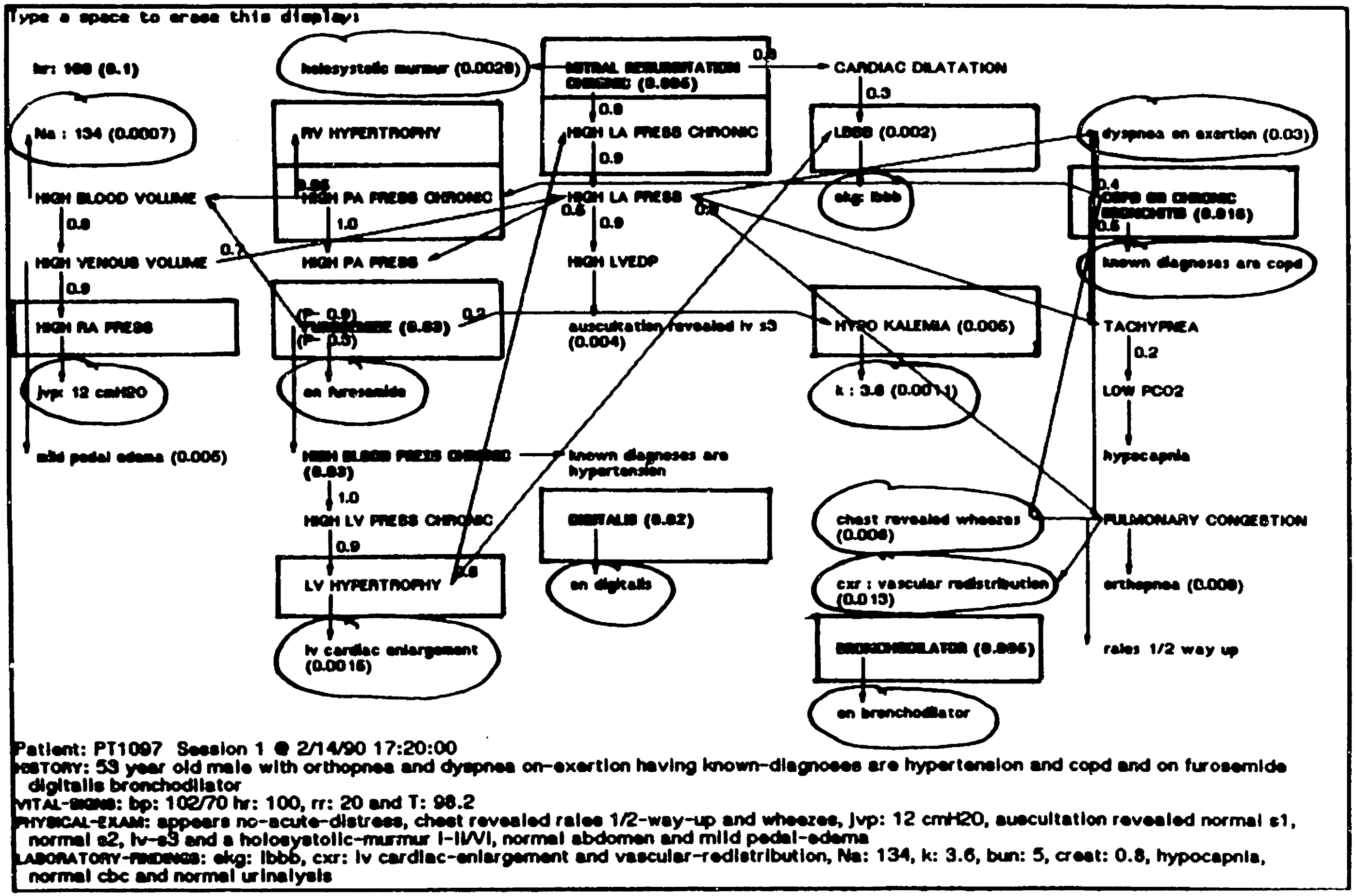
Patient: PT1060 Session 1 @ 2/08/90 10:40:00
HISTORY: 80 year old female with within-hours atypical chest-pain having known-diagnoses are coronary-heart-disease, old-mi, hypertension, copd and hypothyroidism and on pacemaker nitroglycerin hydralazine furosemide digitalis
VITAL-SIGNS: bp: 130/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, normal pulse, auscultation revealed normal s1, normal s2 and a systolic-ejection-murmur at 3ll and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: first-degree-block and old infarct, Na: 140, k: 4.3, bun: 15, creat: 0.9 and normal cbc

Figure 8.7: Ideal solution for PT1060


Type a space to erase this display:
known diagnoses are hypertension
HIGH PA PRESS
0.9
RV STRETCH
0.9
ATRIAL STRETCH
0.3
ATRIAL FIBRILLATION PAROXYSMAL (0.008)
0.4
ATRIAL FIBRILLATION RHYTHM
0.7
HIGH HEART RATE
1.0
(W+ 0.5)
HIGH MYOCARDIAL OXYGEN CONSUMPTION
(W+ 0.5)
(W+ 0.5)
UNSTABLE ANGINA
0.2
MYOCARDIAL ISCHEMIA CURRENT
inferior ischemia (0.0005)
CARDIAC DILATATION
generalized cardiac enlargement (0.014)
LOW CARDIAC OUTPUT
0.8
LOW RENAL PERFUSION
0.8
HIGH RENIN ANGIOTENSIN
0.5
SALT&WATER RETENTION
0.5
HIGH BLOOD VOLUME
0.8
HIGH VENOUS VOLUME
0.7
HIGH LA PRESS
0.9
HIGH LVEDP
auscultation revealed lv s3 (0.004)
HIGH BLOOD PRESS CHRONIC (0.93)
1.0
HIGH LV PRESS CHRONIC
0.25
0.9
LOW LV SYSTOLIC FUNCTION CHRONIC
bun : 21 (0.012)
TACHYPNEA
0.2
LOW PCO2
hypocapnia
mild pedal edema (0.005)
0.8
PLEURAL EFFUSION
cxr : pleural effusion (0.0003)
AORTIC VALVE DISEASE (0.015)
0.9
AORTIC STENOSIS
1.0
FIXED HIGH OUTFLOW RESISTANCE
0.9
0.5
1.0
SLOW EJECTION
GENERAL FLOW DEFICIT
0.05
0.9
EXERTIONAL ANGINA
history of anginal chest pain
basilar rales (0.005)
systolic ejection murmur (0.029)
atrial fibrillation
hr: 110
within hours anginal chest pain
nocturnal dyspnea (0.0006)
PULMONARY CONGESTION
orthopnea (0.009)
cxr : congestive failure (0.0023)
0.5
0.8
0.9

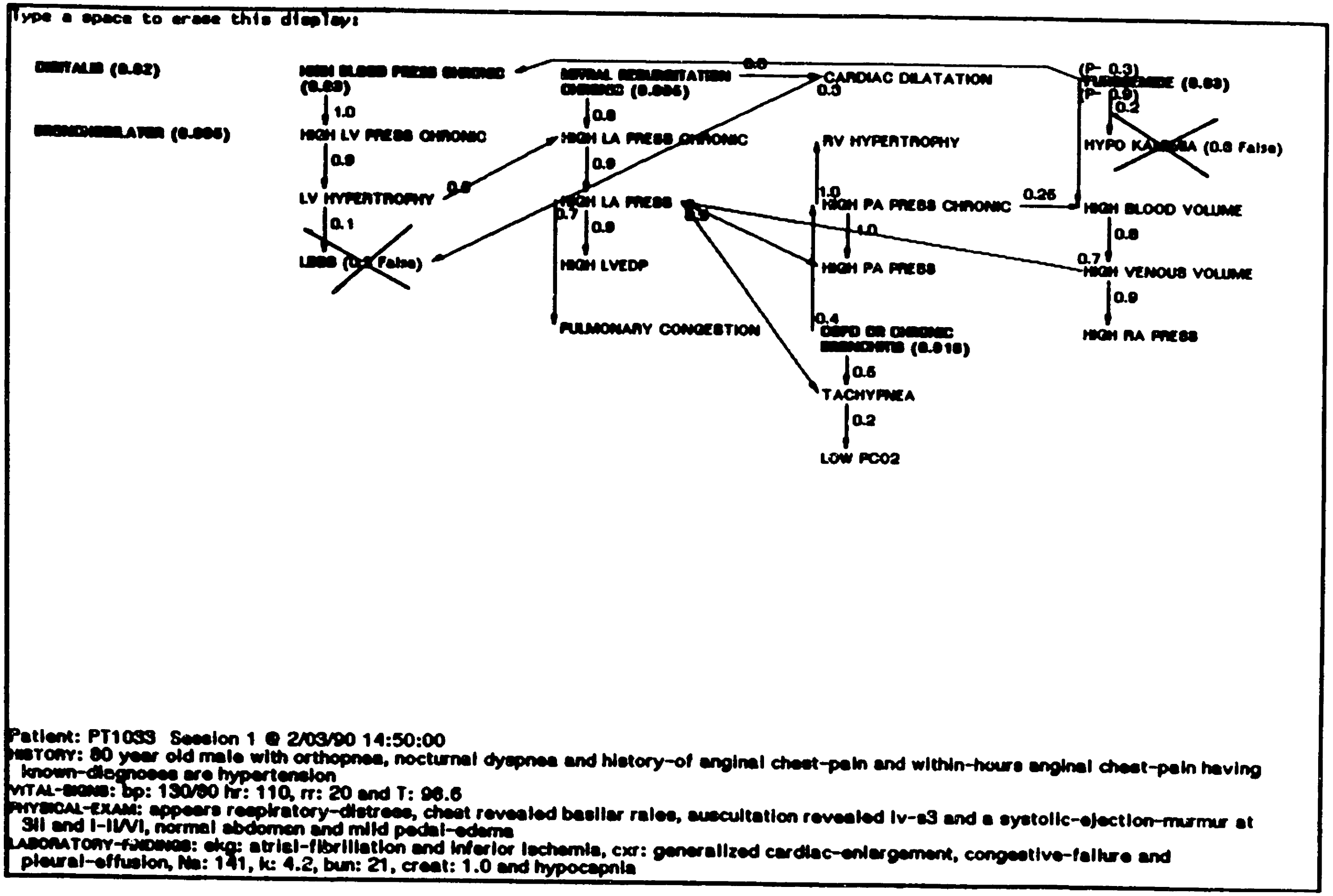
Patient: PT1033 Session 1 @ 2/03/90 14:50:00
HISTORY: 80 year old male with orthopnea, nocturnal dyspnea and history-of anginal chest-pain and within-hours anginal chest-pain having known-diagnoses are hypertension
VITAL-SIGNS: bp: 130/80 hr: 110, rr: 20 and T: 98.6
PHYSICAL-EXAM: appears respiratory-distress, chest revealed basilar rales, auscultation revealed lv-s3 and a systolic-ejection-murmur at 3ll and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: atrial-fibrillation and inferior ischemia, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 141, k: 4.2, bun: 21, creat: 1.0 and hypocapnia

Figure 8.8: Diagnosis for case PT1033



Patient: PT1097 Session 1 @ 2/14/90 17:20:00
HISTORY: 53 year old male with orthopnea and dyspnea on-exertion having known-diagnoses are hypertension and copd and on furosemide digitalis bronchodilator
VITAL-SIGNS: bp: 102/70 hr: 100, rr: 20 and T: 98.2
PHYSICAL-EXAM: appears no-acute-distress, chest revealed rales 1/2-way-up and wheezes, jvp: 12 cmH20, auscultation revealed normal s1, normal s2, lv-s3 and a holosystolic-murmur I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: lbbb, cxr: lv cardiac-enlargement and vascular-redistribution, Na: 134, k: 3.6, bun: 5, creat: 0.8, hypocapnia, normal cbc and normal urinalysis

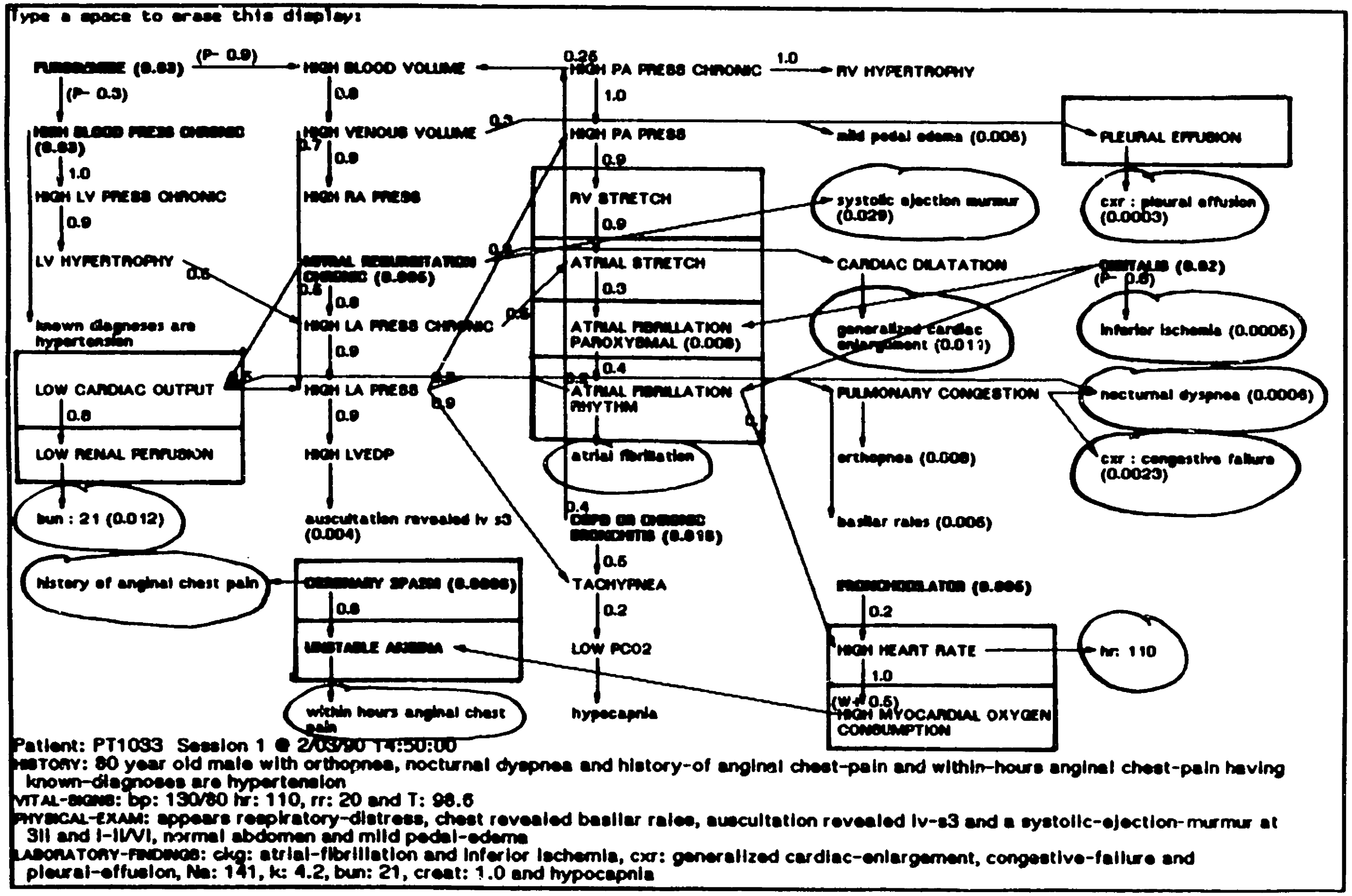
Figure 8.9: Diagnosis for case PT1097



Type a space to erase this display:

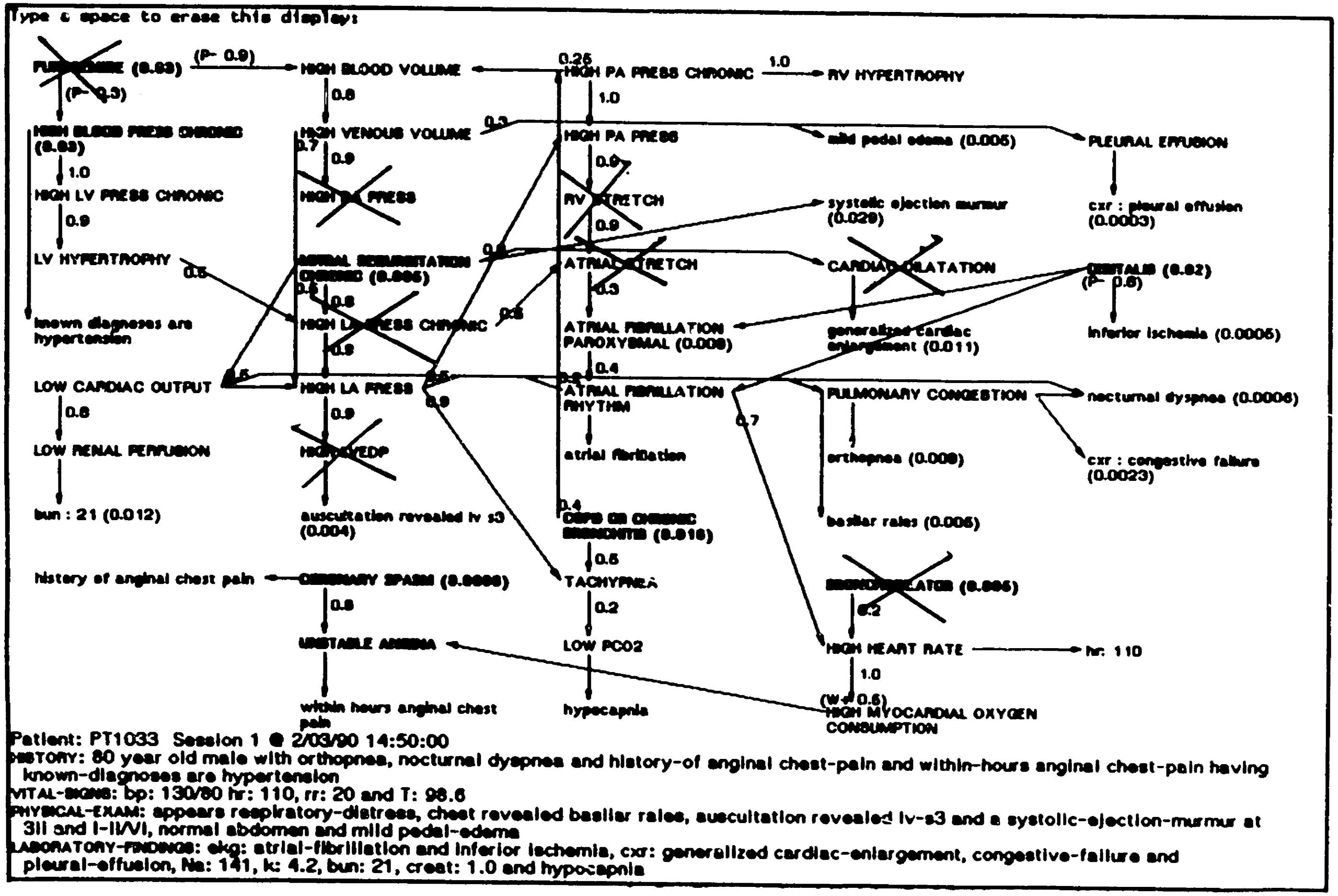
Patient: PT1033 Session 1 @ 2/03/90 14:50:00
HISTORY: 80 year old male with orthopnea, nocturnal dyspnea and history-of anginal chest-pain and within-hours anginal chest-pain having known-diagnoses are hypertension
VITAL-SIGNS: bp: 130/80 hr: 110, rr: 20 and T: 98.6
PHYSICAL-EXAM: appears respiratory-distress, chest revealed basilar rales, auscultation revealed lv-s3 and a systolic-ejection-murmur at SII and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: atrial-fibrillation and inferior ischemia, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 141, k: 4.2, bun: 21, creat: 1.0 and hypocapnia

Figure 8.10: Transfer PT1097 ⇒ PT1033: pruning false states



Patient: PT1033 Session 1 @ 2/03/90 14:50:00
HISTORY: 80 year old male with orthopnea, nocturnal dyspnea and history-of anginal chest-pain and within-hours anginal chest-pain having known-diagnoses are hypertension
VITAL-SIGNS: bp: 130/80 hr: 110, rr: 20 and T: 98.6
PHYSICAL-EXAM: appears respiratory-distress, chest revealed basilar rales, auscultation revealed lv-s3 and a systolic-ejection-murmur at 3II and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ckg: atrial-fibrillation and inferior ischemia, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 141, k: 4.2, bun: 21, creat: 1.0 and hypocapnia

Figure 8.11: Transfer PT1097 ⇒ PT1033: explaining new findings



Patient: PT1033 Session 1 @ 2/03/90 14:50:00

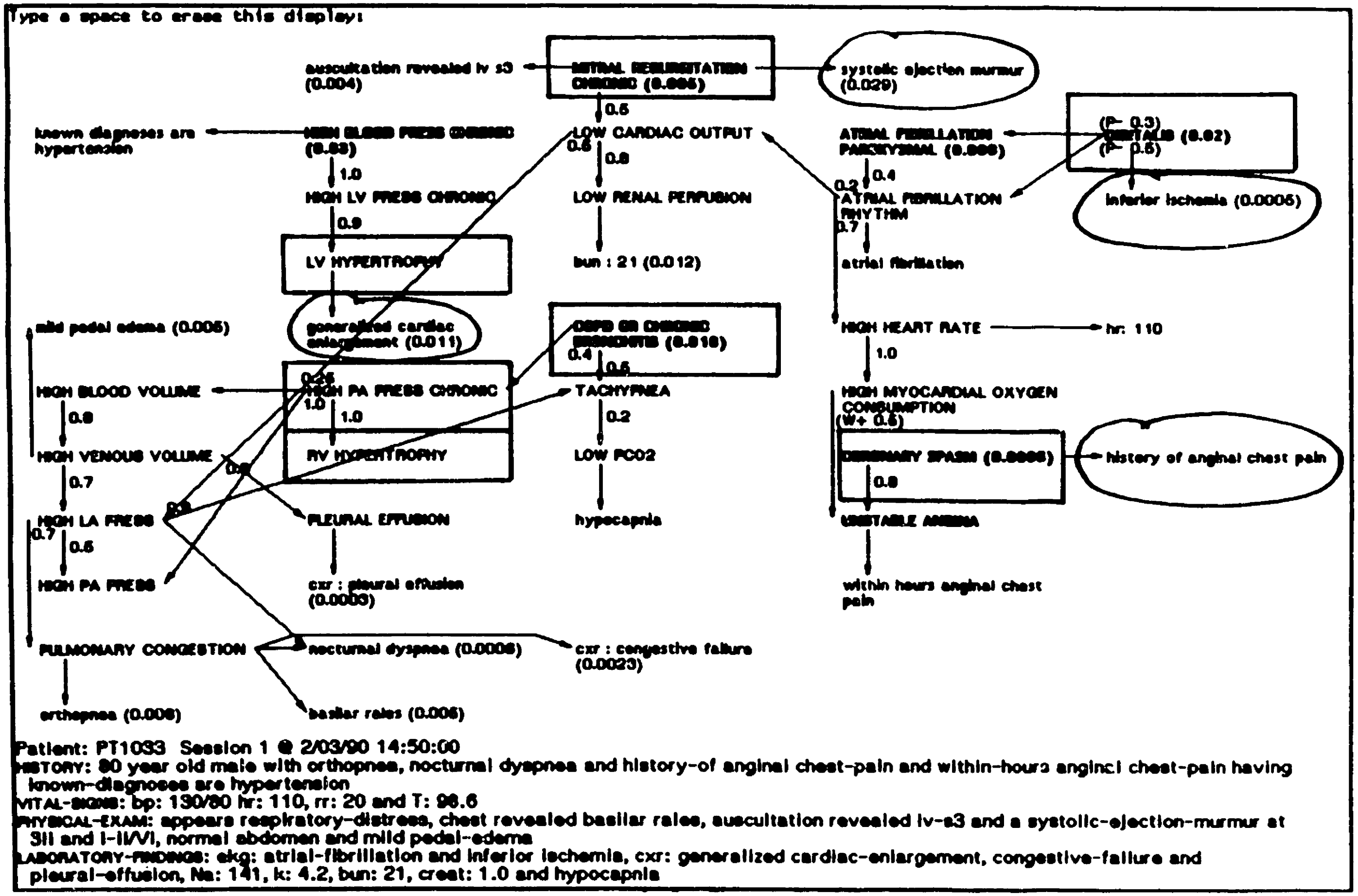
HISTORY: 80 year old male with orthopnea, nocturnal dyspnea and history-of anginal chest-pain and within-hours anginal chest-pain having known-diagnoses are hypertension

VITAL-SIGNS: bp: 130/80 hr: 110, rr: 20 and T: 98.6

PHYSICAL-EXAM: appears respiratory-distress, chest revealed basilar rales, auscultation revealed lv-s3 and a systolic-ejection-murmur at 3ll and I-II/VI, normal abdomen and mild pedal-edema

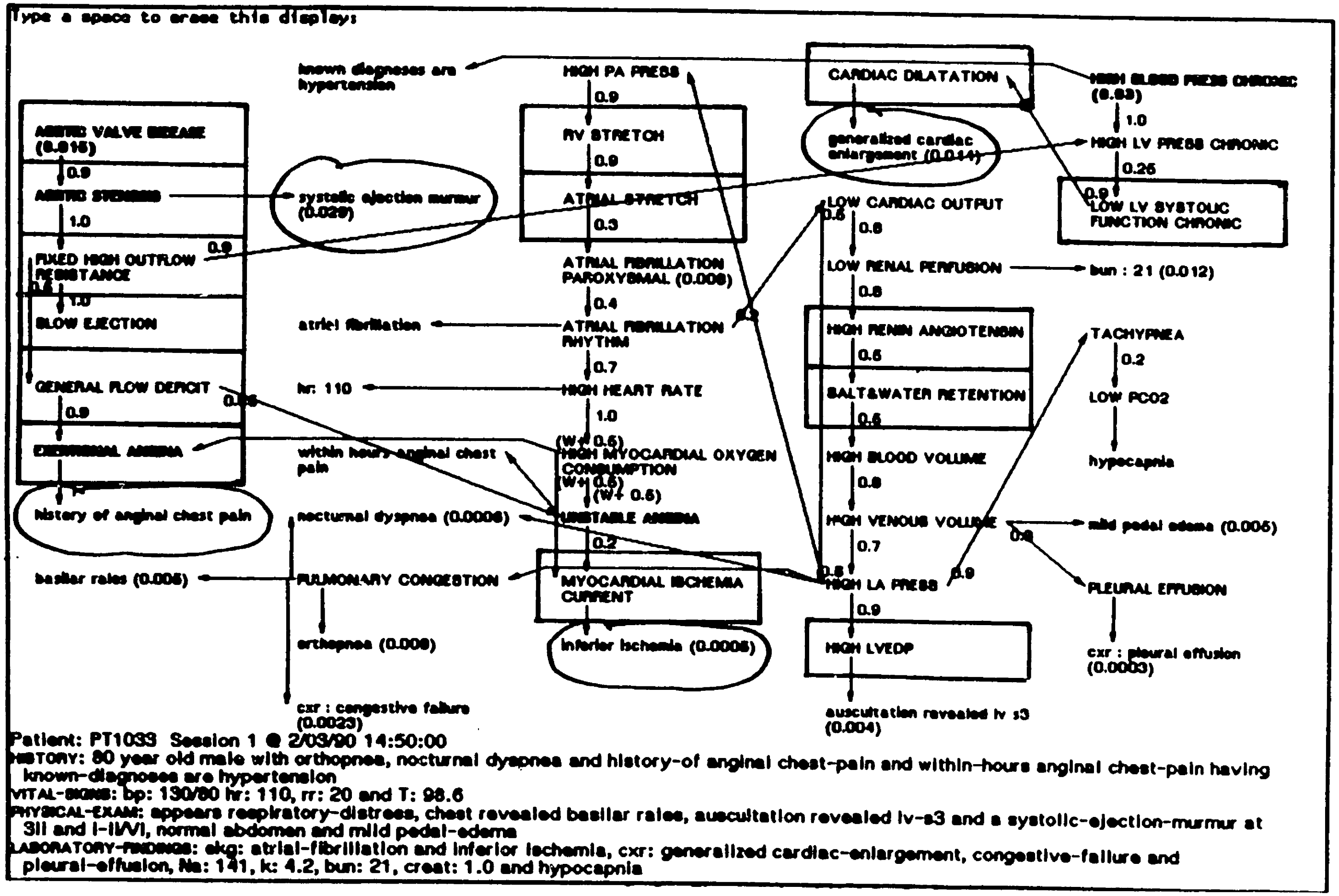
LABORATORY-FINDINGS: ekg: atrial-fibrillation and inferior ischemia, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 141, k: 4.2, bun: 21, creat: 1.0 and hypocapnia

Figure 8.12: Transfer PT1097 ⇒ PT1033: pruning unnecessary states


Type a space to erase this display:
auscultation revealed lv s3 (0.004)
MITRAL REGURGITATION CHRONIC (0.[illegible])
systolic ejection murmur (0.029)
known diagnoses are hypertension
HIGH BLOOD PRESS CHRONIC (0.83)
1.0
HIGH LV PRESS CHRONIC
0.9
LV HYPERTROPHY
generalized cardiac enlargement (0.011)
0.5
LOW CARDIAC OUTPUT
0.6
0.8
LOW RENAL PERFUSION
bun : 21 (0.012)
ATRIAL FIBRILLATION PAROXYSMAL (0.[illegible])
(P- 0.3)
DIGITALIS (0.02)
(P- 0.5)
inferior ischemia (0.0005)
0.4
0.2
ATRIAL FIBRILLATION RHYTHM
0.7
atrial fibrillation
HIGH HEART RATE
hr: 110
1.0
HIGH MYOCARDIAL OXYGEN CONSUMPTION
(W+ 0.5)
CORONARY SPASM (0.[illegible])
history of anginal chest pain
0.8
UNSTABLE ANGINA
within hours anginal chest pain
COPD OR CHRONIC BRONCHITIS (0.018)
0.4
0.5
TACHYPNEA
0.2
LOW PCO2
hypocapnia
mild pedal edema (0.005)
HIGH BLOOD VOLUME
0.25
HIGH PA PRESS CHRONIC
1.0
1.0
RV HYPERTROPHY
0.8
HIGH VENOUS VOLUME
0.7
0.8
0.8
HIGH LA PRESS
0.7
0.5
HIGH PA PRESS
PLEURAL EFFUSION
cxr : pleural effusion (0.0003)
PULMONARY CONGESTION
nocturnal dyspnea (0.0008)
cxr : congestive failure (0.0023)
orthopnea (0.008)
basilar rales (0.006)

Patient: PT1033 Session 1 @ 2/03/90 14:50:00
HISTORY: 80 year old male with orthopnea, nocturnal dyspnea and history-of anginal chest-pain and within-hours anginal chest-pain having known-diagnoses are hypertension
VITAL-SIGNS: bp: 130/80 hr: 110, rr: 20 and T: 98.6
PHYSICAL-EXAM: appears respiratory-distress, chest revealed basilar rales, auscultation revealed lv-s3 and a systolic-ejection-murmur at 3II and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: atrial-fibrillation and inferior ischemia, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 141, k: 4.2, bun: 21, creat: 1.0 and hypocapnia

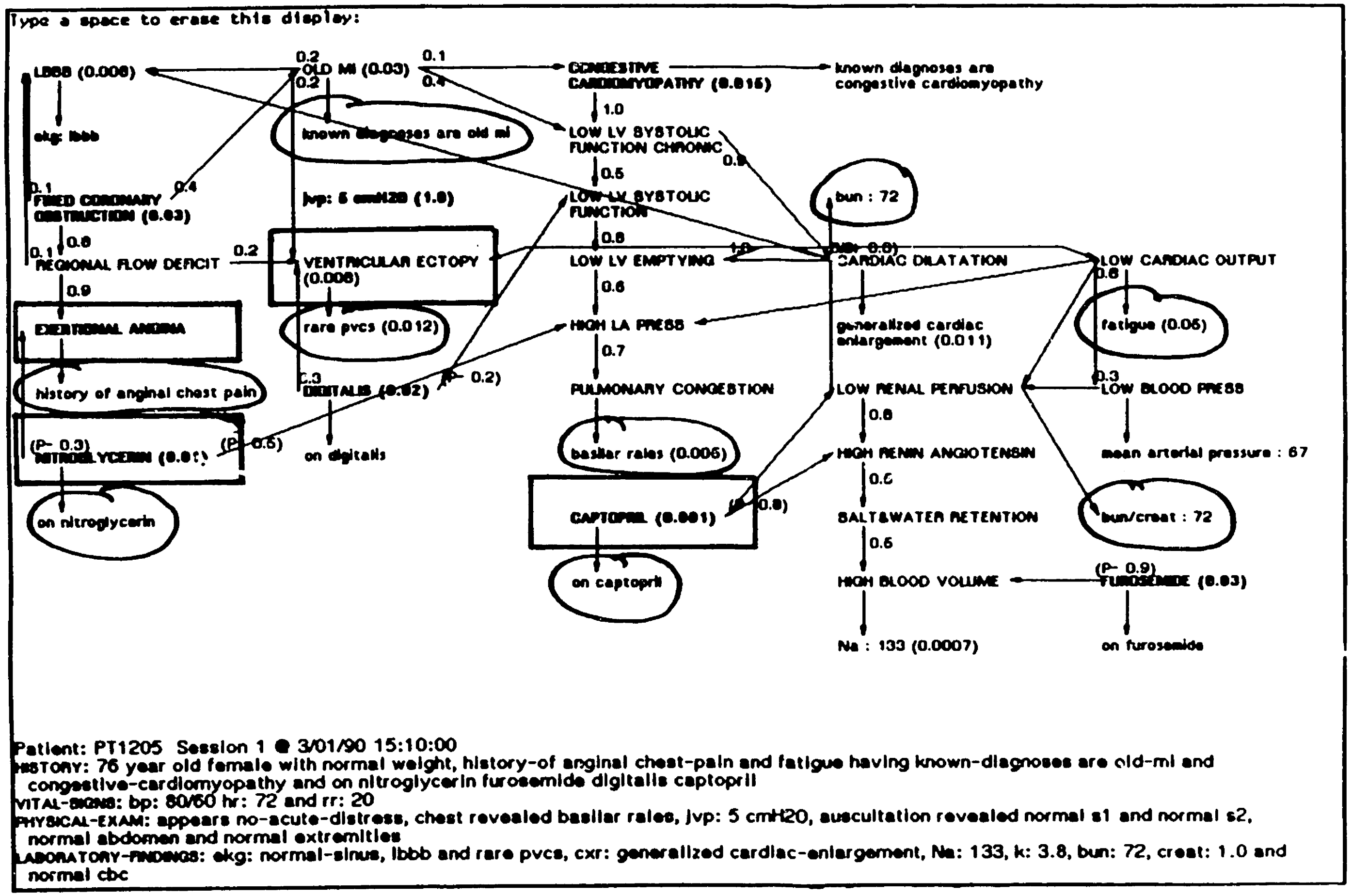
Figure 8.13: CASEY's solution for PT1033



Type a space to erase this display:

Patient: PT1033 Session 1 @ 2/03/90 14:50:00
HISTORY: 80 year old male with orthopnea, nocturnal dyspnea and history-of anginal chest-pain and within-hours anginal chest-pain having known-diagnoses are hypertension
VITAL-SIGNS: bp: 130/80 hr: 110, rr: 20 and T: 98.6
PHYSICAL-EXAM: appears respiratory-distress, chest revealed basilar rales, auscultation revealed lv-s3 and a systolic-ejection-murmur at 3ll and I-II/VI, normal abdomen and mild pedal-edema
LABORATORY-FINDINGS: ekg: atrial-fibrillation and inferior ischemia, cxr: generalized cardiac-enlargement, congestive-failure and pleural-effusion, Na: 141, k: 4.2, bun: 21, creat: 1.0 and hypocapnia

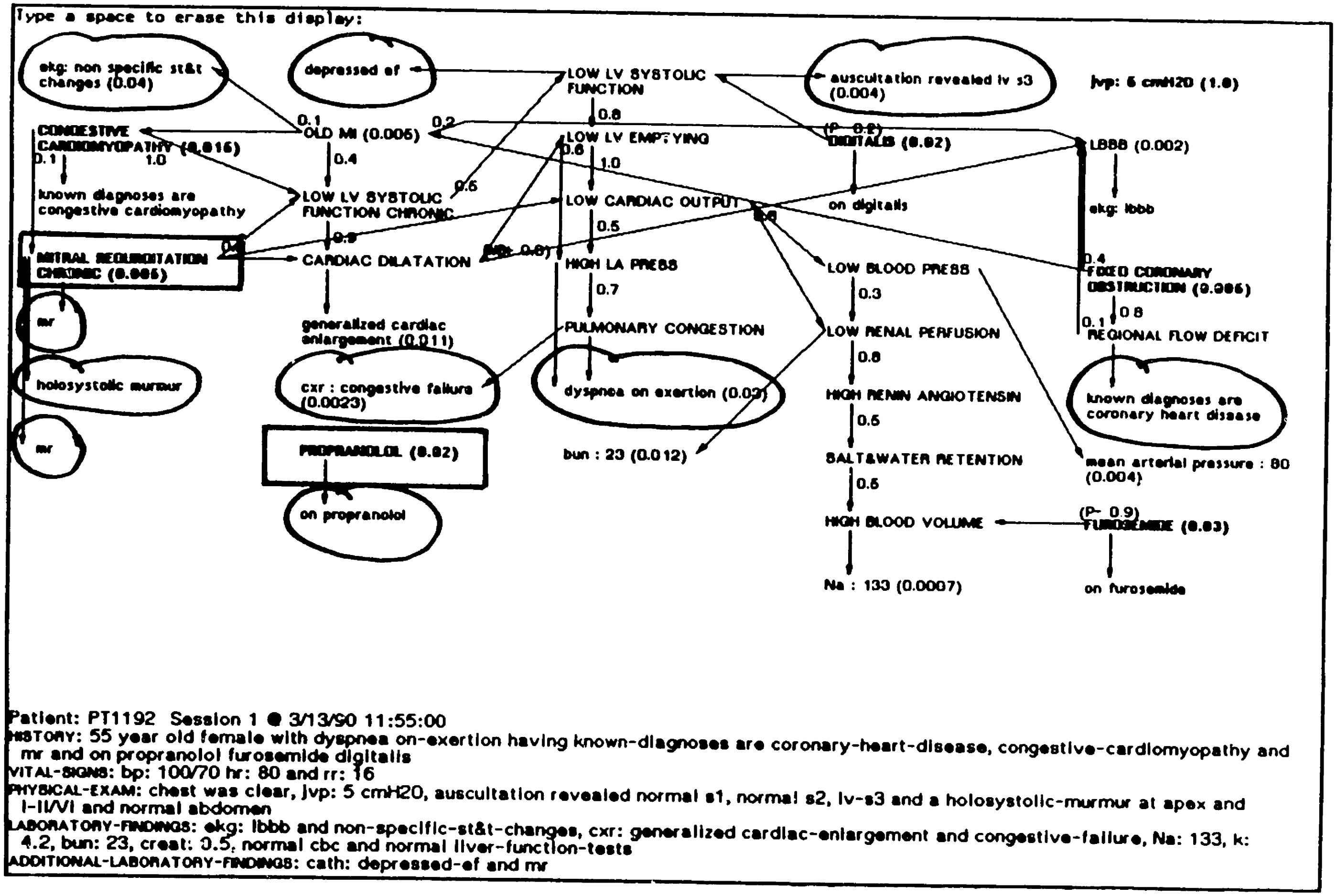
Figure 8.14: Ideal solution for PT1033



Type a space to erase this display:

Patient: PT1205 Session 1 @ 3/01/90 15:10:00
HISTORY: 76 year old female with normal weight, history-of anginal chest-pain and fatigue having known-diagnoses are old-mi and congestive-cardiomyopathy and on nitroglycerin furosemide digitalis captopril
VITAL-SIGNS: bp: 80/60 hr: 72 and rr: 20
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, auscultation revealed normal s1 and normal s2, normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: normal-sinus, lbbb and rare pvcs, cxr: generalized cardiac-enlargement, Na: 133, k: 3.8, bun: 72, creat: 1.0 and normal cbc

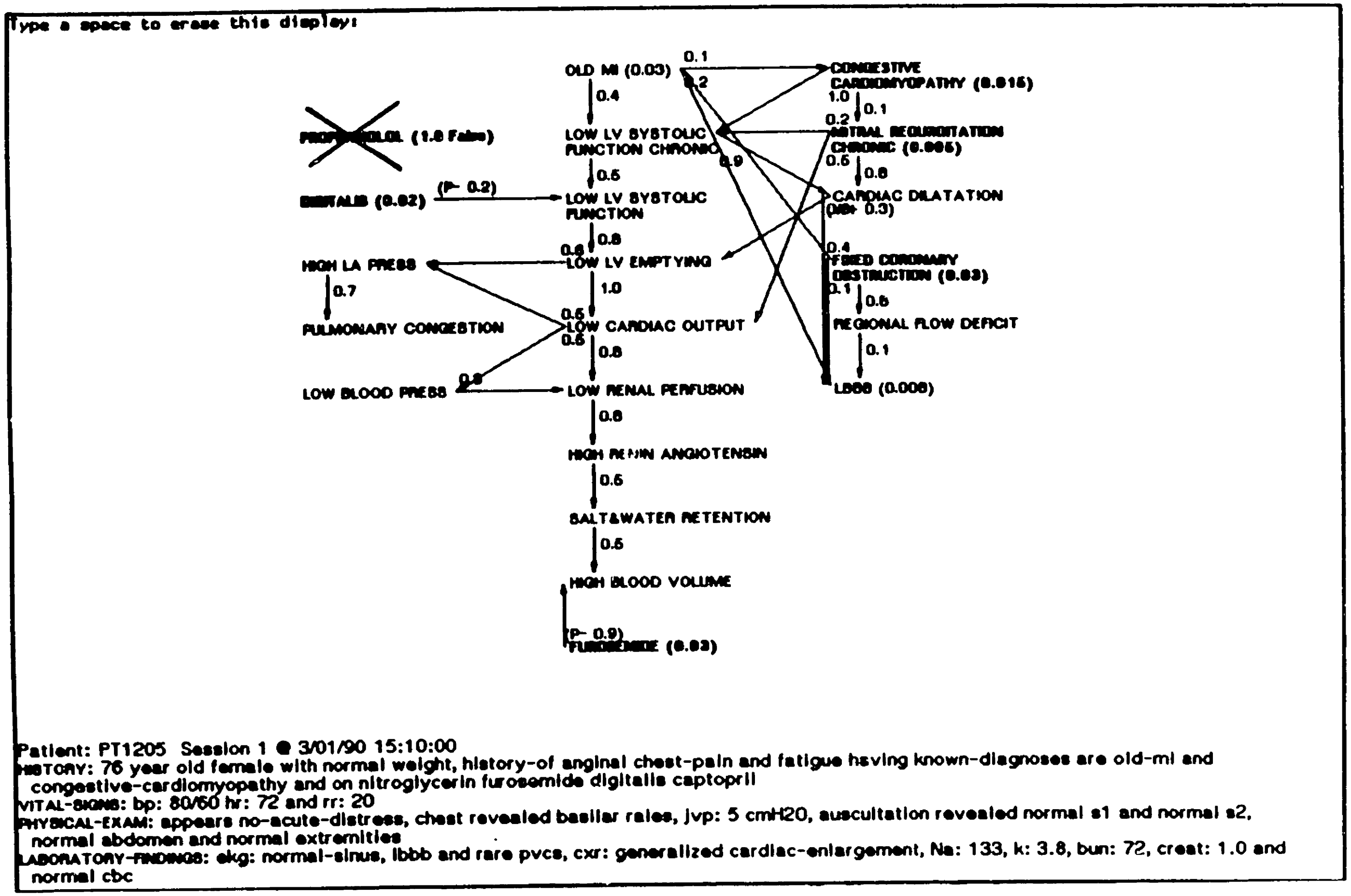
Figure 8.15: Diagnosis for case PT1205



Type a space to erase this display:

Patient: PT1192 Session 1 @ 3/13/90 11:55:00
HISTORY: 55 year old female with dyspnea on-exertion having known-diagnoses are coronary-heart-disease, congestive-cardiomyopathy and mr and on propranolol furosemide digitalis
VITAL-SIGNS: bp: 100/70 hr: 80 and rr: 16
PHYSICAL-EXAM: chest was clear, jvp: 5 cmH20, auscultation revealed normal s1, normal s2, lv-s3 and a holosystolic-murmur at apex and I-II/VI and normal abdomen
LABORATORY-FINDINGS: ekg: lbbb and non-specific-st&t-changes, cxr: generalized cardiac-enlargement and congestive-failure, Na: 133, k: 4.2, bun: 23, creat: 0.5, normal cbc and normal liver-function-tests
ADDITIONAL-LABORATORY-FINDINGS: cath: depressed-ef and mr

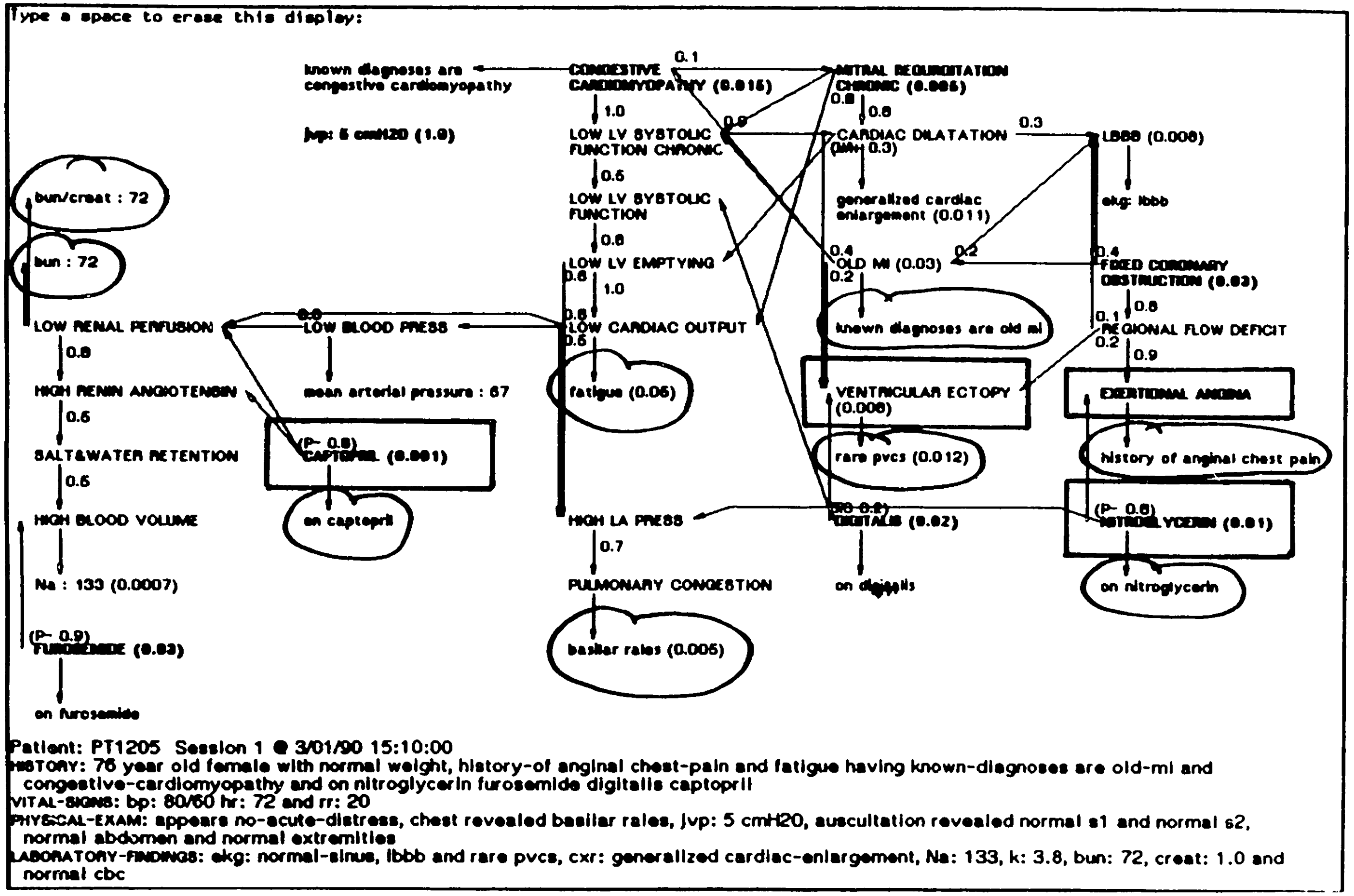
Figure 8.16: Diagnosis for case PT1192



```
Type a space to erase this display:
```

```
Patient: PT1205  Session 1 @ 3/01/90 15:10:00
HISTORY: 76 year old female with normal weight, history-of anginal chest-pain and fatigue having known-diagnoses are old-mi and
  congestive-cardiomyopathy and on nitroglycerin furosemide digitalis captopril
VITAL-SIGNS: bp: 80/60 hr: 72 and rr: 20
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, auscultation revealed normal s1 and normal s2,
  normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: normal-sinus, lbbb and rare pvcs, cxr: generalized cardiac-enlargement, Na: 133, k: 3.8, bun: 72, creat: 1.0 and
  normal cbc
```

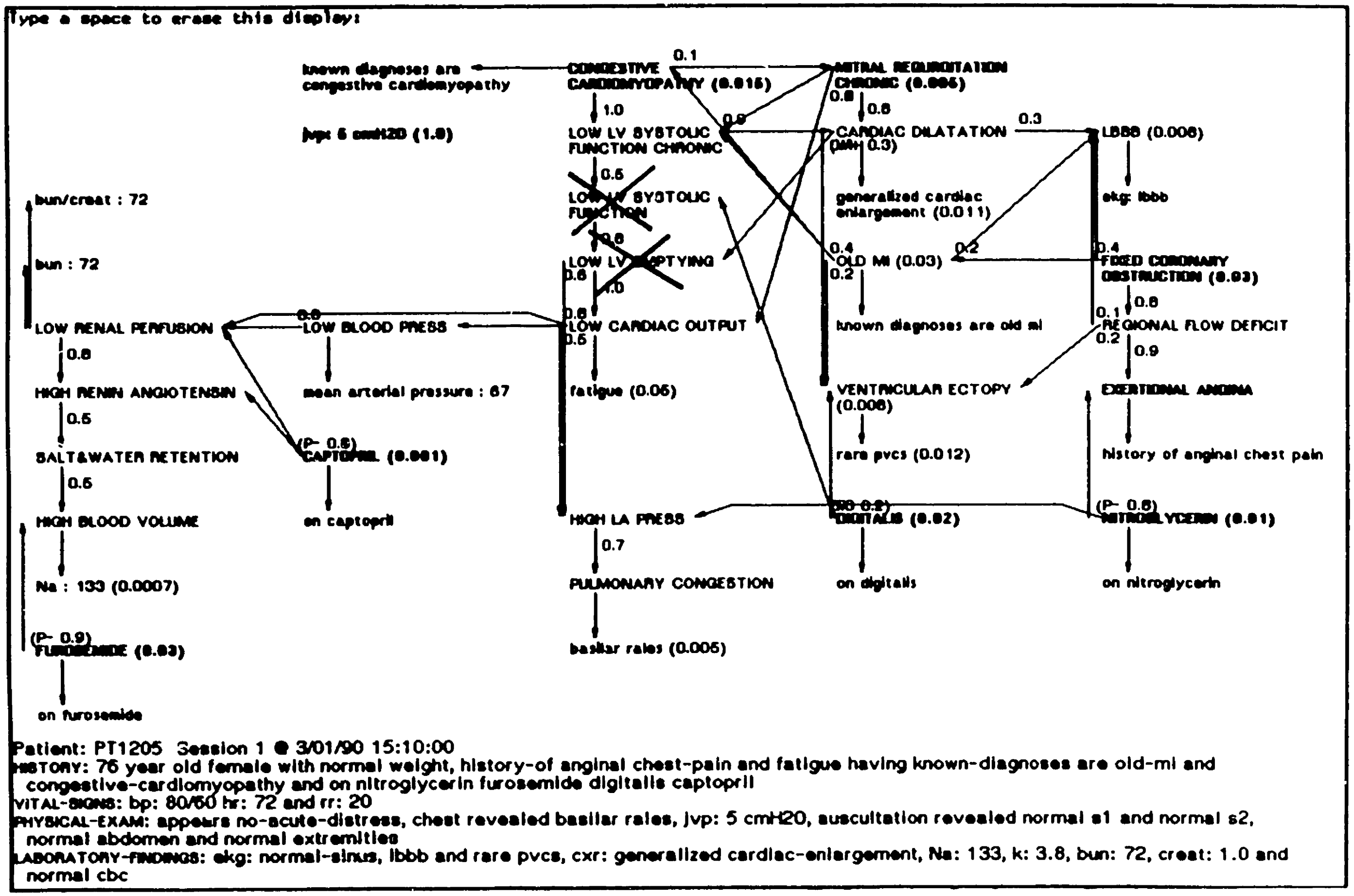
Figure 8.17: Transfer PT1192 ⇒ PT1205: pruning false states



Type a space to erase this display:

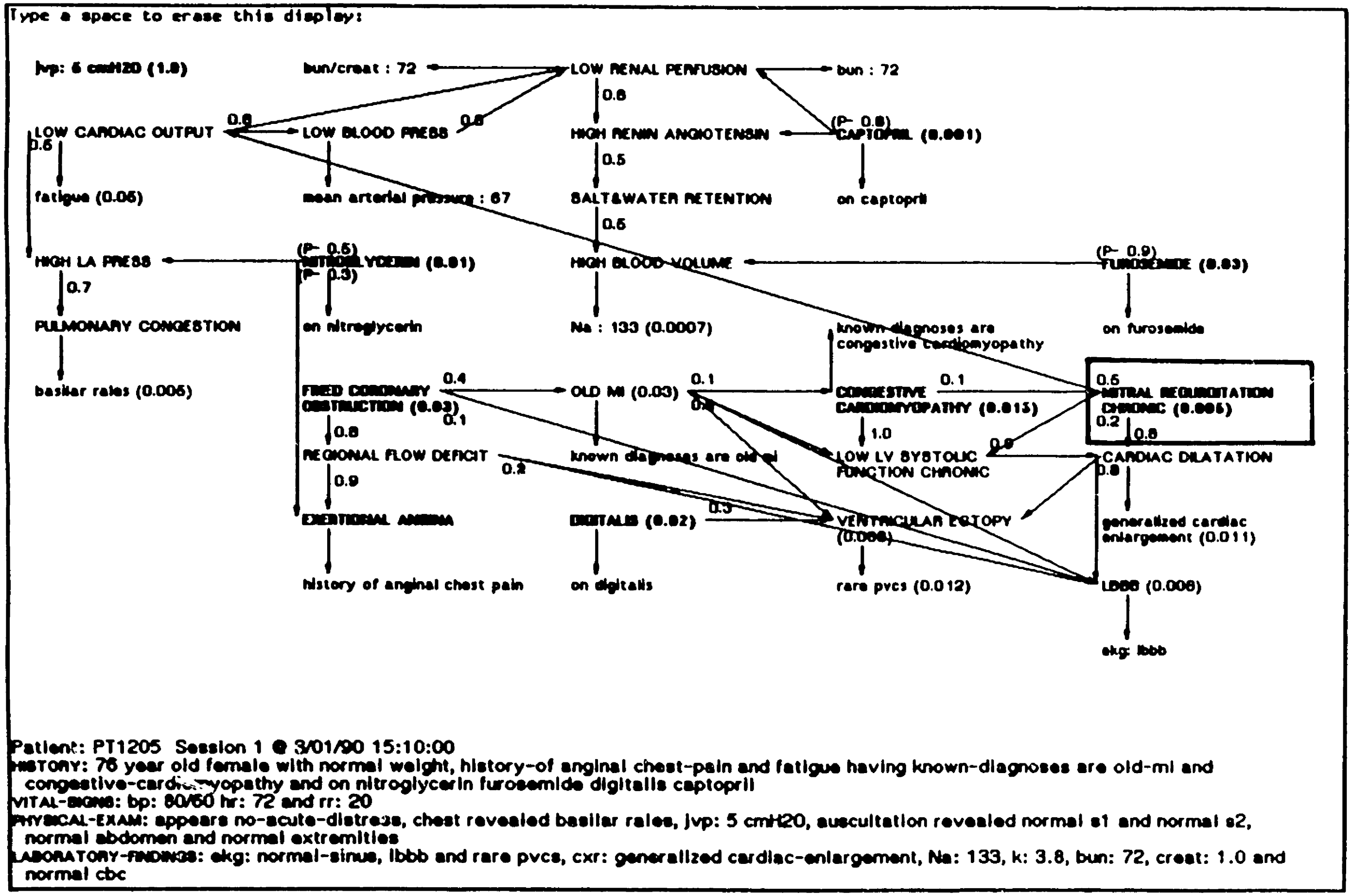
Patient: PT1205 Session 1 @ 3/01/90 15:10:00
HISTORY: 76 year old female with normal weight, history-of anginal chest-pain and fatigue having known-diagnoses are old-mi and congestive-cardiomyopathy and on nitroglycerin furosemide digitalis captopril
VITAL-SIGNS: bp: 80/60 hr: 72 and rr: 20
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, auscultation revealed normal s1 and normal s2, normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: normal-sinus, lbbb and rare pvcs, cxr: generalized cardiac-enlargement, Na: 133, k: 3.8, bun: 72, creat: 1.0 and normal cbc

Figure 8.18: Transfer PT1192 ⇒ PT1205: explaining new findings



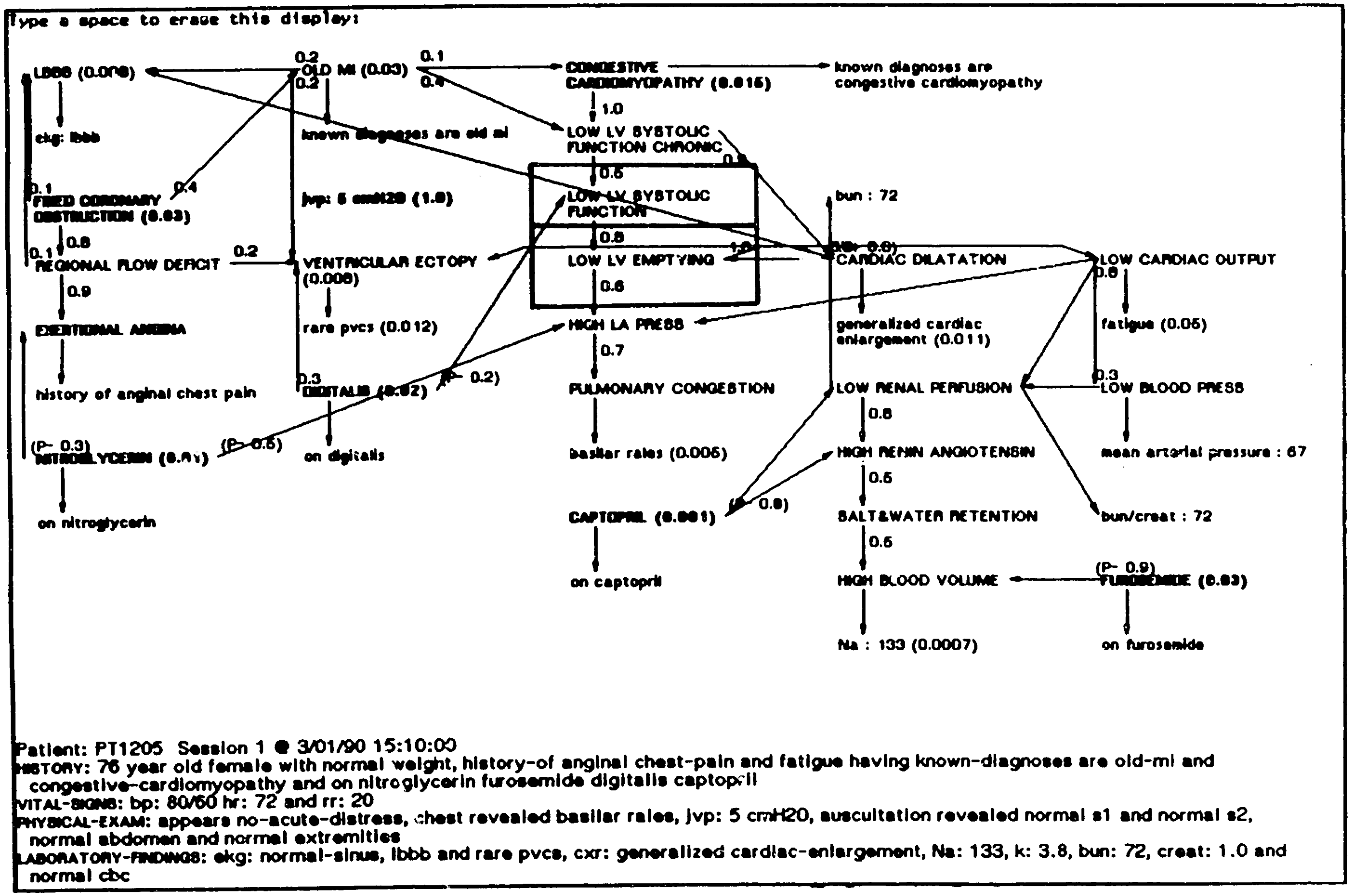
Patient: PT1205 Session 1 @ 3/01/90 15:10:00
HISTORY: 76 year old female with normal weight, history-of anginal chest-pain and fatigue having known-diagnoses are old-mi and congestive-cardiomyopathy and on nitroglycerin furosemide digitalis captopril
VITAL-SIGNS: bp: 80/60 hr: 72 and rr: 20
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, auscultation revealed normal s1 and normal s2, normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: normal-sinus, lbbb and rare pvcs, cxr: generalized cardiac-enlargement, Na: 133, k: 3.8, bun: 72, creat: 1.0 and normal cbc

Figure 8.19: Transfer PT1192 ⇒ PT1205: pruning unnecessary states



Patient: PT1205 Session 1 @ 3/01/90 15:10:00
HISTORY: 76 year old female with normal weight, history-of anginal chest-pain and fatigue having known-diagnoses are old-mi and congestive-cardiomyopathy and on nitroglycerin furosemide digitalis captopril
VITAL-SIGNS: bp: 80/60 hr: 72 and rr: 20
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, auscultation revealed normal s1 and normal s2, normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: normal-sinus, lbbb and rare pvcs, cxr: generalized cardiac-enlargement, Na: 133, k: 3.8, bun: 72, creat: 1.0 and normal cbc

Figure 8.20: CASEY's solution for PT1205



Type a space to erase this display:

Patient: PT1205 Session 1 @ 3/01/90 15:10:00
HISTORY: 76 year old female with normal weight, history-of anginal chest-pain and fatigue having known-diagnoses are old-mi and congestive-cardiomyopathy and on nitroglycerin furosemide digitalis captopril
VITAL-SIGNS: bp: 80/60 hr: 72 and rr: 20
PHYSICAL-EXAM: appears no-acute-distress, chest revealed basilar rales, jvp: 5 cmH20, auscultation revealed normal s1 and normal s2, normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: normal-sinus, lbbb and rare pvcs, cxr: generalized cardiac-enlargement, Na: 133, k: 3.8, bun: 72, creat: 1.0 and normal cbc

Figure 8.21: Ideal solution for PT1205

Chapter 9

Discussion

9.1 Generalizing the Results

What can we infer from case-based reasoning's ineffectiveness for heart failure diagnosis? Fundamentally, all inferences are limited by the domain specificity of our evaluation. However, this study went beyond mere evaluation to identify the dependence of performance on the domain. Having characterized the aspects of heart failure that allow it to elude case-based reasoning, we can project this elusiveness onto other potential applications that share the inauspicious traits.

Analyzing the distribution of patients and their diagnoses exposed a three-fold lack of regularity, which neatly corresponds to failure at the three paradigmatic stages of case-based reasoning. First, because similar cases do not recur routinely, precedent retrieval fails in its search for patients that match the new case. Secondly, case similarity is correlated only weakly with diagnostic similarity; thus, precedent evaluation often misjudges the suitability of cases retrieved. Finally, the rarity of similar solutions precludes successful transfer almost entirely, given anything less than an immense pool of patients.

Conclusively, then, the following conditions are necessary for the method's success:

- Similar cases recur
- Similar cases require similar solutions
- Similar solutions recur

Without them, reasoning would fail at one of the three corresponding stages of the paradigm. Because the first two conditions entail the third, however, they appear to be sufficient, but empirical evidence of *failure* can not verify this assessment. Proof of sufficiency would require carefully controlled experiments that demonstrate *success*.

Without sufficient conditions, we cannot delineate the class of applications that admit CBR. However, each necessary condition circumscribes a category of domains that do not support the method. For example, similar situations are presumably rare in air traffic control, as the number of possible aircraft locations, directions, and destinations can combine to form innumerable novel traffic situations. If controllers constantly maneuver through unforeseen circumstances, they would not benefit from CBR and, instead, must apply general principles to each problem anew. Similar EKG signals, in the presence of noise, might not necessarily appear similar. Without an adequate metric for selecting precedents, signal detection must resort to more sophisticated, context sensitive analysis. Heart failure violates both conditions of regularity, as this study demonstrates. Furthermore, discussions with several physicians reveal that cardiology is no more heterogeneous than other medical specialties. Therefore, medical diagnosis in general appears to lie outside the scope cf the case-based approach.

9.2 Heterogeneity or Homogeneity?

The major outcome of this work is an elucidation of the pervasive and unexpected heterogeneity inherent to the heart failure domain. Such a result directly contradicts not only our own limited intuitions, but moreover, the understanding of cardiologists who spend their careers diagnosing patients with heart failure. How could the domain experts themselves insist that similar cases recur routinely, when an analysis of the case distribution illustrates conclusively that they do not?

The issue is one of perspective, "as the waves shape themselves symmetrically from the cliff top, but to the swimmer among them are divided by steep gulfs, and foaming crests"[26]. Cardiologists may espouse the looser, panoramic definition of similarity, calling cases routine when they are *familiar* in a number of respects. The patients selected for the case studies of the previous chapter, for example, fit this interpretation. Conversely, case-based reasoning relies critically on a closer, stricter diagnostic correspondence. For its purposes, the notion of similarity is futile unless it signifies approximate identity.

Because almost all heart failure cases involve multiple, interacting diseases, it is difficult to find two patients with exactly the same set of complaints. The pool of 240 cases, which represents all heart failure patients diagnosed at Tufts New England Medical Center over approximately two years, barely samples the nearly infinite variety of *combinations* that exist among findings, disease states, and diagnostic syndromes. A practicing cardiologist, over the course of ten years, may see only 1000 cases; yet, even 100,000 would not exhaust the vast diversity of possibilities.[1] Human beings, however, maintain the results of induction over their experiences, and may reason at a different level of abstraction. Thus, when cardiologists encounter a "routine" case, they perceive a collection of findings, of diseases, of syndromes that, *individually*, all look familiar, while in actuality, the particular combination may be unique to their experience.

For example, the expert chess player may come to regard most moves, with a few exceptions, as straightforward and formulaic. Notwithstanding, has (s)he *ever* played the same game of chess twice? Detective stories also seem inevitably conventional to the mystery aficionado. Each character is standard in some way, perhaps resembling another character from a novel read previously. Even the once surprising twists of plot become commonplace. Comparing two particular books closely, however, uncovers not only a number of similarities, but also quite a few differences, and if we had to identify the villain of one novel given the clues of the other, we would fail.

[1] Considering the 7000 pathways that the Heart Failure system uses to construct its hypotheses, considering that a typical diagnosis contains *at least* 8 pathways, and considering that a particular pathway rules out *at most* 100 other pathways, the number (7000)(6900)(6800)(6700)(6600)(6500)(6400)(6300) $\approx 10^{30}$ is a *conservative* lower bound on the total number of causal explanations possible.

Correspondingly, it would seem that people do not acquire language by assembling a case base of utterances. Instead, they derive principles of grammar and semantics, which they subsequently employ in constructing and understanding any of the infinitely many expressions that the language makes possible. Because the same principles apply time after time, our linguistic endeavors become routine, and therefore automatic.

Apparently, although somewhat paradoxically, we can perform complex reasoning within a heterogeneous domain and still consider each problem unremarkable.

9.3 The Nature of Experience

What is the nature of experience that it affords us this computational versatility? The answer lies in the subtle distinction between such abstract experience and concrete experience*s*.

Until recently, semantic memory, concerning abstract *experience*, and episodic memory, concerning concrete *experiences*, were though to represent identical phenomena. Both involve the storage of information, both expedite the retrieval of this information when it is evoked or activated by relevant thought, and both consequently facilitate the partial replication of a former state of mind, whose content can be used concertedly with current reasoning.

> For a long time the prevalent view was that memory is essentially unitary and that different forms of retrieval represent one and the same set of underlying processes, differing only in the kind of factual information retrieved: episodic remembering is the retrieval

> of personal, temporally dated, and self-relevant facts, whereas semantic knowing is the retrieval of impersonal, undated, and world relevant facts[22].

Endel Tulving and his colleagues have demonstrated not only that the two types of memory are functionally distinct, but that they are local to structurally disparate regions of the brain[22]. Tulving also relates fascinating accounts of the amnesiac K.C., whose frontal lobe injuries virtually extinguished his ability to retrieve episodic information, while leaving his semantic memory largely intact. Although K.C. cannot remember a single personal past experience, he exhibits no language impairment and can discuss with ease sundry topics such as history, geography, politics, and music. K.C. knows exactly how to change a flat tire, but he cannot recollect ever having performed the sequence of steps he recalls, nor having watched another person do so. Similarly, he is adept at chess, without remembering ever having played the game with anyone[22].

From "these kinds of contrasts between what K.C. does not remember *of* his past and what he knows *from* it [Tulving's italics]" emerges a compelling distinction between remembering and knowing[22]. Schank's notion of reminding, therefore, is different from the more general knowing by which people bring concise, distilled experience to bear in *routine* problem solving. Although reminding may initiate induction, it is quickly superseded by the more automatic, abstract, semantic retrieval that results when experiences are analyzed and compiled. Obviously, K.C. can not, and therefore does not, reason from any previous chess game, or even any particular move, as an entity. Instead, he must have abstracted his general knowledge of the game from them, relegating them to oblivion without allowing his playing ability to decrease.

I do not claim that case-based reasoning is a misguided approach to problem solving. It offers enormous advantages to domains with smooth and tractably finite (as opposed to systematically enumerable) solution spaces. Domains like heart failure diagnosis, however, are too complex and varied, conducive more to the analytic learning of modular abstractions, rather than to a strict memory based approach, which loses its efficiency when an impractical number of cases are required.[2]

9.4 Ongoing and Future Research

Acquiring Experience When human beings learn problem-solving tasks such as diagnosing heart failure patients, they do not store cases and solutions in their entirety, but instead analyze them, incorporating the knowledge they derive inductively into abstract, general principles. How computers might implement this process is an important research issue, raised by the limitations of the pure case-based approach for domains that lack sufficient regularity.

Yeona Jang proposes a solution within the heart failure domain, in terms of a system that dissects precedent diagnoses into *generalized causal units*[9]. While each unit includes only a single primary disease and its pathophysiologic effects, it also records the co-occurring diseases that directly influenced this pathophysiology, to account for disease interactions. Units compiled from dif-

[2] The Heart Failure system's causal model may very well represent induced, semantic knowledge; recall that this model-based system can solve most cases, with laudable accuracy, in under a minute.

ferent precedents identify the various symptom contexts associated with each primary disease, maintaining and updating frequencies according to which each implicated finding is representative of the context. By reducing diagnostic hypotheses in this manner, Jang deals systematically with the complexity generated by multiple diseases and their interactions. Presumably, the concise, modular constituents will recur more often than do cases themselves.

Case-based reasoning learns at the level of cases and their solutions, which for heart failure diagnosis and other complex domains, constitute too broad a level of detail. Clearly, a more reductive approach, one that induces smaller, modular principles, is required to exploit the fine-grain regularity that these domains possess. Yet without knowing exactly the level of abstraction at which this regularity occurs, how can a machine learning system isolate the desired recurrences? Arguably, by analyzing the statistical correlations that exist among atomic units and assembling commonly occurring clusters of them, a system could dynamically induce the most suitable level of abstraction along with the recurrences it encompasses.

For example, in the heart failure domain, such a system might derive likely conglomerations of co-occurring pathophysiologic states from the Heart Failure system's causal model. Co-occurrence could be defined initially as the transitive closure of a correlative equivalence relation, determined by the probability on the link between two states. Clusters, presumably, would represent diagnostic syndromes, such as aortic disease (including AORTIC VALVE DISEASE, AORTIC STENOSIS, FIXED HIGH OUTFLOW RESISTANCE, and SLOW EJECTION; see figure 8.8) or hypertensive left ventricular hypertrophy (including

HIGH BLOOD PRESSURE CHRONIC, HIGH LV PRESSURE CHRONIC, and LV HYPERTROPHY; see figures 8.2 and 8.9). Representative symptoms would affiliate with the clusters according to analogous computation.

Subsequently, precedent diagnoses could be used to refine the clustering. For every pathophysiologic link implicated in the precedent causal explanation, the system either adds statistical support to an *intra*-cluster link or provides an *inter*-cluster link with suggested support. Conversely, the system should weaken every cluster correlation pairing a precedent pathophysiologic state with a state not included in the precedent hypothesis. Cluster merging occurs when the links between two clusters accumulate sufficient support, while cluster fragmentation occurs when the links between two or more cluster components become sufficiently tenuous.

The system diagnoses a new case by activating the clusters supported by findings. Intermediate and primary clusters are activated recursively by the evidence provided by other clusters. Finally, selected clusters combine to form a hypothesis, which can be adjusted locally using the Heart Failure system's causal model.

Although CASEY's generalization mechanism may eventually create these clusters, it treats them as self-sufficient units, confusing them with GENs at less suitable levels of generality and confounding their compositional potential. By dynamically discovering the appropriate level of abstraction for each cluster, the alternate system could exploit maximum regularity while keeping combinatorial complexity to a minimum.

Combining Semantic Knowledge and Episodic Memory Phyllis Koton, in collaboration with Marc Vilain and Melissa Chase, has developed a representation language for hybrid reasoning involving both semantic knowledge and episodic memories [23]. Through this representation, the learner can be primed with the analytic, or semantic, knowledge necessary for routine problem solving, expressing this knowledge as paradigmatic generalizations in a discrimination network. Subsequently, as its practical experience grows, the system can incorporate statistical, or episodic, memories, which specify exceptions to the initial defaults and are indexed beneath them. The inductive bias provided by the original knowledge not only facilitates the interpretation of new experiences, but in addition reduces the sample complexity of the learning task, defined by the number of cases required to achieve proficiency.

David Spiegelhalter, Nomi Harris, and their colleagues have also investigated the use of case experience to augment semantic, expert-derived knowledge[20,8]. Their Bayesian belief network for diagnosing congenital heart disease adjusts subjective probabilities estimated by physicians, using the statistical evidence provided by actual patients. According to the quantified uncertainty with which the experts qualify their judgments, the system posits a hypothetical sample of patients upon which these judgments might have been based, with greater uncertainty indicating a smaller sample size. New cases are added incrementally to these hypothetical samples, thus modifying the probability estimates according to experience. While episodic learning allows the system to overcome its initial biases and uncertainty, the semantic foundation decreases the amount of experience it requires to derive accurate diagnoses.

Memory Design The memory structure designed by Kolodner[10] and implemented in CASEY exhibits substantial complexity problems. Although it *might* fare better in a simpler, more regular domain, this study clearly illustrates its propensity for rapid growth, challenging researchers to design memories that are more reliably tractable. Unfortunately, the problem of redundant indexing is a difficult one to overcome, considering the combinatorial number of ways in which two cases can match. CBR research has taken recourse alternately to selecting specific features for indexing *a priori*, to performing retrieval in parallel[13,21,24], and to employing connectionist approaches to precedent matching through constraint satisfaction[21].

Solution Adaptation As the case studies of Chapter 8 testify, CASEY often makes the mistake of attributing a finding to a less likely cause, simply because this cause already exists in the precedent hypothesis. To improve solution adaptation, a newer implementation of the system might consider all possible explanations of a finding, within the context of the rest of the hypothesis, before assigning it to a particular pathophysiologic state; thus, the system would insure that the finding is ascribed to its most likely cause under the circumstances outlined by the precedent diagnostic scheme.[3] Furthermore, CASEY could take likelihood into account while pruning "unnecessary" states, and might include among a new patient's generalized causal features only those

[3] However, this "enhanced" implementation could no longer claim to restrict modification entirely to the local differences between the new case and the precedent. Instead, it would undertake more thorough computation comparable to some of the reasoning that the Heart Failure system performs. We must be careful, when striking the balance between reasoning from similar cases and reasoning from domain knowledge, not to defeat CASEY's purpose of avoiding the reconstruction of solutions from scratch.

potential causes whose causal likelihood exceeds a prespecified threshold.[4] An even more rigorous use of probability might involve retrieving and evaluating precedents on the basis of *strongly indicative findings*, whose specificity can be calculated using the prevalences of these findings and their pathophysiologic causes, as well as the conditional probabilities embodied in the causal links between them.[5]

Exploiting the Differential Diagnosis While the Heart Failure system constructs a differential diagnosis for each patient, CASEY stores only the *most* likely hypotheses with its precedents. By ignoring alternate solutions, the case-based system may discard valuable clues for constructing new causal explanations. For example, a precedent and new case might be similar in that they both include diagnoses A and B in their differentials. However, subtle, contextual clues might render A more likely for the precedent, while identifying B as more probable for the new case. Since CASEY focuses on only the best hypothesis for each patient, it regards the two solutions as different, and therefore, transfer would fail. Instead CASEY could transfer the entire differential, reorganizing it according to the suitability of each implicated hypothesis for the new case.[6]

Evaluation Across Domains The major limitation of this work is its domain specificity. While evaluating other domains may broaden or restrict the compass of the results presented here, it will in either case contribute to a

[4] Suggested by Peter Szolovits.
[5] Suggested by William Long.
[6] Suggested by William Long.

fuller understanding of the case-based approach.

In order to delimit the scope of the method systematically, future research might test hypothetical domains, each created and controlled to meet different criteria of regularity in varying degrees. Experiments involving these domains could then pinpoint the sufficient conditions for CBR's success and verify the necessary conditions identified here. For consistency, the hypothetical domains should share the same feature and solution spaces. Random functions could then generate test cases easily, according to distributions that meet prespecified regularity criteria. For example, one such distribution might generate case configurations with little variation (similar cases recur) as well as solution configurations that do not differ substantially (similar solutions recur), while pairing cases and solutions haphazardly (similar cases do not require similar solutions).

As the results of this study run so contrary to its initial assumptions, the evaluation of other case-based applications becomes even more imperative. Current research should be either vindicated or redirected, but under no circumstances should it continue to rely on the intuition that proved so misleading for the heart failure domain.

9.5 Summary

Case-based reasoning attempts to exploit the regularity inherent to routine problem solving, by solving problems from experience. Given a new "case," the reasoner searches its collection of previously solved cases, called precedents, and derives, from the most suitable precedent, a response to the new situation. The reasoner gains efficiency by remembering the solutions to recurrent problems rather than reconstructing them from scratch. Furthermore, the experiential approach forfeits no accuracy, as long as similar problems require similar solutions consistently.

Although these intuitive claims have remained largely untested, they continue to be the *primary* justification for research in CBR. The case-based approach has been implemented in various domains, but almost none of these applications have undergone thorough empirical evaluation. Few researchers have investigated the scaling effects of experience on accuracy and efficiency. Moreover, no research at all has endeavored to explore the method's scope.

This study answers many of these concerns, by evaluating case-based reasoning in the context of heart failure diagnosis, using the existing CASEY system and a pool of 240 patients. Because cardiologists describe most of their cases as routine, and because the pathophysiology of heart failure guarantees systematic, causal relationships between diseases and their symptoms, I expected to confirm the conjecture that the regularity inherent to heart failure diagnosis would allow CBR to succeed.

First, I investigated the scale of the problem, varying the number of precedents and measuring the effects of increasing experience on both accuracy and efficiency. Presumably, increasing the number of precedents available would enhance the reasoner's chances of locating an appropriate match for the new case, thus causing accuracy to improve. Efficiency, I expected, would decline only moderately. Instead, my results were surprising:

- Accuracy experienced no significant change, remaining unacceptably low, even as experience increased to 200 precedents.

- Efficiency, meanwhile, degraded significantly as experience increased: memory grew almost quadratically with the number of precedents, while precedent retrieval searched approximately six times as many memory nodes as there were patients in the case base.

Upon analyzing the distribution of patients, I discovered that they were, contrary to initial presumption, extremely heterogeneous. Thus, the method failed inevitably for the following reasons:

- Similar cases did not recur

  The similarity of the *best matched* precedent, after peaking rapidly, remained mediocre as case base size grew from 1 to 200 patients.

- Similar cases did not necessarily require similar solutions

  Case similarity was only weakly correlated, if at all, with diagnostic similarity and the accuracy of solution transfer.

- Similar solutions did not recur

  The 240 patient diagnoses were all mutually distinct; moreover, statistical calculations revealed that over a hundred thousand patients might be required to provide decent precedents for most typical cases.

In conclusion, my research reveals several apparently unobvious truths concerning the case-based approach and the domains, such as heart failure diagnosis, which are unsuited to it.

Case-based reasoning makes strong assumptions concerning domain regularity, with respect to the size of the entities whose common occurrence and correlation it governs. By defining the unit of experience to be an *entire* problem-solving situation and its solution, the method restricts itself to domains exhibiting *coarse-grain* homogeneity. Domains like medical diagnosis involve situations that combine, in infinitely many possible ways, smaller recurring elements such as diagnostic syndromes. Thus, despite the uniform fine-grain patterns, these domains do not support case-based reasoning, because the larger problem solving units are infinitely various and, therefore, largely distinct. To learn these complex, but undoubtedly regular, domains requires the induction of smaller, modular principles, which can be used interactively to solve the less homogeneous, larger problems.

9.6 Intuition and Evaluation

Perhaps as significant as anything this study demonstrates about the heart failure domain and case-based reasoning is its lesson concerning intuition and evaluation. Human intuition, or common sense, is ordinarily dependable. It derives from our general experience, and we learn to rely upon it in general. However, there have been many situations, this study among them, during which we become disillusioned with intuition, during which "All that in idea seemed simple became in practice immediately complex"[26]. Because of the inevitable contexts that we have not experienced and because of the inevitable factors that we have not considered, intuition is always more simplistic than the unfathomed truth and, in occasional consequence, leads us to be naive.

Over the centuries, science has assumed the responsibility for dispelling this naiveté, providing us with more accurate, more detailed, and often more complex visions of our existence. For example, not until 1867 did Pasteur refute the simple notion that life is generated spontaneously. In 1916, Einstein showed the long accepted laws of Newton to be mere approximations to a subtler truth, and shortly thereafter, Heisenberg exposed the uncertainty underlying what we thought was a deterministic universe.

Because our intuitions in artificial intelligence are similarly subject to question, it is our duty as scientists to question them and, more productively, to subject them to evaluation.

Bibliography

[1] Ray Bareiss. The experimental evaluation of a case-based learning apprentice. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 162-167, 1989.

[2] Jaime G. Carbonell. Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning*, volume 2. Morgan Kaufmann, Los Altos, CA, 1986.

[3] Paul R. Cohen and Adele E. Howe. How evaluation guides AI research. *AI Magazine*, 9(4):35-43, 1988.

[4] Paul R. Cohen. Evaluation and case-based reasoning. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 168-172, 1989.

[5] Marc Goodman. CBR in battle planning. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 264-269, 1989.

[6] Kristian J. Hammond. CHEF: A model of case-based planning. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages

267-271, 1986.

[7] Kristian J. Hammond. Case-based planning. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 264-269, 1988.

[8] Nomi L. Harris, David J. Spiegelhalter, Kate Bull, and Rodney C. G. Franklin. Criticizing conditional probabilities in belief networks. Unpublished manuscript, 1990.

[9] Yeona Jang. Diagnostic system capable of compiling knowledge through experience. Unpublished manuscript, 1990.

[10] Janet L. Kolodner. Maintaining organization in a dynamic long-term memory. *Cognitive Science*, 7:243-280, 1983.

[11] Janet L. Kolodner. Reconstructive memory: A computer model. *Cognitive Science*, 7:281-328, 1983.

[12] Janet L. Kolodner, Robert L. Simpson, Jr., and Katia Sycara-Cyranski. A process model of case-based reasoning in problem solving. In *Proceedings of the National Conference on Artificial Intelligence*, pages 284-290. American Association for Artificial Intelligence, 1985.

[13] Janet L. Kolodner. Retrieving events from a case memory: A parallel implementation. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 233-249, 1988.

[14] Phyllis A. Koton. Using experience in learning and problem solving. TR 441, Massachusetts Institute of Technology, Laboratory for Computer Science, 545 Technology Square, Cambridge, MA, 02139, 1989.

[15] Phyllis A. Koton. Applications and validation: Case-based reasoning in the real world (Panel overview). In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, page 160, 1989.

[16] Phyllis A. Koton, Evaluating case-based problem solving. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 173-175, 1989.

[17] William J. Long. Medical diagnosis using a probabilistic causal network. *Applied Artificial Intelligence*, 3:367-383, 1989.

[18] William S. Mark. Case-based reasoning for autoclave management. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 176-180, 1989.

[19] Roger C. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People.* Cambridge University Press, Cambridge, 1982.

[20] David J. Spiegelhalter, Rodney C. G. Franklin, and Kate Bull. Assessment, criticism, and improvement of imprecise subjective probabilities for a medical expert system. In *Proceedings of the Workshop on Uncertainty in Artificial Intelligence*, pages 335-342, 1989.

[21] Paul Thagard and Keith J. Holyoak. Why indexing is the wrong way to think about analog retrieval. In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 36-40, 1989.

[22] Endel Tulving. Remembering and knowing the past. *American Scientist*, 77(4):361-367, 1989.

[23] Marc Vilain, Phyllis Koton, and Melissa P. Chase. On analytical and similarity-based classification. Unpublished manuscript, to appear in *Proceedings of the National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, 1990.

[24] David L. Waltz. Is indexing used for retrieval? In *Proceedings of the DARPA-Sponsored Case-Based Reasoning Workshop*, pages 41-44, 1989.

[25] Patrick H. Winston. Learning new principles from precedents and exercises: The details. AIM 678, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139, 1982.

[26] Virginia Woolf. *To the Lighthouse*. Harcourt Brace Jovanovich, New York, 1927.