Capturing and Representing the Reasoning Processes of Expert Clinical Teachers for Case-Based Teaching

Geneviève Gauthier

Department of Educational and Counselling Psychology

McGill University, Montréal

August 2009

A thesis submitted to McGill University
in partial fulfillment of the requirements for the degree of
Doctor in Philosophy in Educational Psychology

ABSTRACT

Case-based learning has demonstrated its potential in educational and professional environments for its ability to promote reasoning skills and foster independent learning. However, the use of this instructional approach is limited by current assessment practices. Considering that assessment drives and defines the types of learning that can be measured, case-based learning requires a new mode of assessment. Using an evidence-centered design perspective, we propose to explore an assessment task aligned with key instructional objectives of case-based learning.

In the context of medical education, we propose a methodology that addresses the challenge of capturing and representing evolving knowledge into a validation activity. This activity is anchored on case-based teaching practices commonly performed by physicians in medical education. The study examines the reasoning processes of five medical experts while they solve and teach three specific cases. Verbal protocol combined with outcome and process measures lead to an initial visual representation of expert's reasoning processes. These initial visual representations are used with experts for them to validate and evaluate their own reasoning processes for each case. This reflection task informs the design of a combined visual representation for each case that shows similarities and differences in experts' performance. These results can inform the design of complex assessment practices that incorporates models of competent performance and helps to guide curriculum development in medical education.

RÉSUMÉ

L'apprentissage utilisant la méthode de cas appliqués est une approche d'apprentissage qui a fait ses preuves tant dans le milieu professionnel que dans le milieu éducatif car elle stimule le développement du raisonnement et favorise la prise en charge de l'apprentissage chez les apprenants. Toutefois son utilisation dans le milieu éducatif est limitée car les objectifs d'apprentissage clés de cette approche ne sont pas soutenus par les méthodes d'évaluation ayant cours dans le système éducatif. Cette situation est problématique car les méthodes d'évaluation ont un très grand impact sur le type et la nature des apprentissages ayant cours dans les salles de classes.

Dans cette étude nous proposons d'explorer une approche d'évaluation inspirée de la conception basée sur des données probantes (evidence-centered design) dans un contexte d'éducation médicale. La première étape, permettant d'établir les bases d'une telle approche d'évaluation, consiste à construire de manière empirique des modèles permettant l'évaluation et l'interprétation des caractéristiques et paramètres de performance de résolutions de cas. En proposant une méthodologie qui a pour but de capturer et de représenter visuellement le processus de résolution, cette étude analyse la performance de cinq experts cliniciens dans le cadre d'une activité simulant la présentation et la résolution de trois cas de patients. Le protocole verbal combiné à des mesures caractérisant leur processus de raisonnement et leurs solutions sont utilisés pour créer une représentation individuelle pour chacune des résolutions de cas. Ces représentations sont ensuite utilisées avec les experts afin qu'ils valident et évaluent leur propres performances. Cette deuxième étape permet d'établir les étapes importantes communes et les différences menant à une résolution valide de cas. Ces modèles et cette méthodologie ont pour but d'établir une base servant à l'évaluation et au développement de matériel pédagogique pour l'apprentissage par cas.

ACKNOWLEDGEMENTS

Thank you to Susanne P. Lajoie, my supervisor, for believing in me as a person first and foremost. Her support and encouragement have been constant through the challenges of this process.

Thanks to Lysanne and my family who probably thought this would be a never-ending journey but supported me through it just the same. Thank you to Solange, a friend whose lasting intellectual and emotional support was essential to achieving this goal.

And to Sonia, Andrew, Carlos and other colleagues from the research lab for their support and opinions. I wish to acknowledge and thank the passionate educators who participated in this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In today's information age, knowledge is evolving at a rapid pace and the need for learning how-to-learn is gaining importance at all levels of education. It is impossible to design a curriculum that will include everything students need to know. No matter how much material students learn during their schooling and training, they will need to continue developing their ability in order to update and expand their knowledge and competence in their specific areas of specialization given the dynamic growth of information in all areas of specialization. Learners need to develop their abilities to think, solve problems, and become independent learners (Pellegrino, Chudowsky, & Glaser, 2001). As our concept of learning is evolving toward a competency-based model, new instructional approaches such as case-based learning are gaining in popularity since they foster a skill and practice approach to learning (Lundeberg, Levin, & Harrington, 1999; Lundeberg & Yadav, 2006). However, the assessment of the students' knowledge is reported to be a challenging task throughout the literature on case-based learning (Lundeberg & Yadav, 2006; Sudzina, 1999; Sykes & Bird, 1992; Williams, 1992).

Although there is a large amount of research in favour of case-based teaching (Gijbels, Segers, & Struyf, 2008; Savin-Baden, 2004), these instructional approaches pose a critical challenge to current educational assessment practices. Teaching students the correct answer is not the goal of case-based learning; instead, its goal is to teach the reasoning and decision-making processes involved in complex problem solving. The assessment challenge is due to both a philosophical and practical misalignment with current psychometric practices (Savin-Baden, 2004). At the philosophical level, the conceptual learning framework of case-based learning conflicts with the learning framework underlying current assessment practices. At the practical level, current assessment practices encourage students to favour rote learning at the expense of deeper understanding. Assessment practices are largely based on a factual outcome-

based approach to learning, and the psychometric framework that informs many of the quality standards are based on a trait-like approach to learning. This framework limits educational approaches to assessment since it does not take into account that learning is developmental and contextual. Shulman and Lundeberg (Lundeberg et al., 1999; Shulman, 1992) state that there is a need for more empirical research documenting the assessment of case-based education.

The study uses an inductive approach to evidence-centered design (Mislevy, Almond, & Lukas, 2003) within a situated perspective on learning (Clancey, 1997; Greeno, 1998). Situated learning theory describes how learning occurs in specific contexts where human thought and action are described as being responsive to the environment. Such contexts provide opportunities for integrating information from multiple sources and promote the social construction of knowledge where knowledge is transformed by integrating the perspectives of those in the group.

The research uses an empirically-based approach to study and model a problem solving task performed by instructional experts. In the context of exploring new modes of assessment (Birenbaum & Dochy, 1996; Segers, Dochy, & Cascallar, 2003), this approach has the potential of informing assessment practices by supporting learner reflection on knowledge representations. This study is anchored in a representative task of knowledge transmission in medicine through the study of competent case resolution protocols. The analysis of the protocol of these expert physicians is used to create blueprints of competent problem solving. These blueprints correspond with the initial requirement of building evidence models in the context of an evidenced-centered assessment system. By gaining a better understanding of the processes and outcomes related to competent performances in context, these evidence models can inform the design of assessment practices aligned with case-based learning goals and purposes.

Structure of the dissertation

The use of cases in teaching offers the potential to meet the challenges of today's evolving knowledge system by showing students how their curriculum-based knowledge can be applied and adapted in practice. The use of cases in teaching encompasses notions of curriculum and instruction; the emphasis is not only on what is taught but how it is taught. Since educators use the term "case" in various ways, the literature review begins by exploring definitions of what cases mean in a variety of contexts. This exploration of the terminology reveals that the choice of what is taught is intrinsically linked to how it is taught to learners; even if case-based learning approaches have differences in meaning, there are common instructional objectives. The review of the literature provides three key instructional objectives: the aim for higher order thinking, the emphasis on the process, and the concept of case as a "messy" or ill-defined problem.

In the literature review these key objectives are discussed in terms of how they pose a critical challenge to current educational assessment practices due to both a philosophical and practical misalignment (Savin-Baden, 2004). At the philosophical level, the conceptual learning framework of case-based learning conflicts with the learning framework underlying current assessment practices. At the practical level, assessment practices occurring at the classroom level have many different purposes prioritized differently by the many stakeholders involved; moreover, these different stakeholders do not share common standards of evaluation for assessment. Current standards to evaluate assessment practices and tools are based on high-stakes testing which often conflicts with supporting the learning occurring in the small-scale context (Shepard, 2006). High stakes assessment refers to a test that has significant consequence for the test-taker. These tests are usually standardized and they play a key role in discriminating who will receive certification, job opportunities and funding opportunities.

These tests have a role to play in the educational system however their purposes and priorities are different from those of classroom assessment.

The present dissertation explores new mode of assessment for classroom context guided by the instructional objectives of case-based learning. More specifically this study explores how expert clinical instructors solve and reflect on specific problem solving tasks, such as diagnosing a patient problem. Informed by a recent model of expertise suggesting that deliberate practice and not time and practice alone, is required to achieve consistent, measurable and reproducible expert performance (Bransford & Schwartz, 2009; Charness, Hoffman, Feltovich, & Ericsson, 2006). By documenting the decision making processes of five clinical teachers on the same cases, the intent is to show the degree of convergence and divergence in the way they diagnose and evaluate the diagnoses and build models that include outcome and process measures to show how and what knowledge experts use to solve case-specific situations.

Definition and Use of Case

In this first section we briefly examine definitions and the etymology of the word “case” to help position the blending of content and method related to its use. The concept of case is used in multiple contexts, so definitions are reviewed to explore the various concepts behind the idea of case. Given a lack of consensus on what constitutes a case, defining the word will help us frame the debate about the different uses of the word within and across domains. These contexts are explored here to provide insight into how the concept of case has been defined and operationalized in practice. Even if much of the focus revolves around the content within a case, looking at definitions and implementations of cases can reveal implications and connection to the organization of material (sequence), the types of knowledge emphasized, the way knowledge is conveyed, and the intended purpose in using cases in the context of learning.

General Definitions of Case

Case, as defined in the Oxford English Dictionary refers to “as an instance or example of something occurring ("Oxford English Dictionary," 2004). The list of synonyms — instance, example, and specimen — accompanying the definition refer to the assembly of concepts it contains, yet the meaning remains quite vague. Further down the page four distinct definitions are detailed: a case as a container (goods, bookcase, storage); a case as a case study (in business, law or computer science); a case as evidence or a set of evidence in argumentation; and a case as a grammatical case (as a linguistic inflection). This refinement can be further narrowed by looking at the etymology of the word which can be traced to two different origins (Etymology Online, 2008). One comes from the Latin “casus” as in a fall, an event or occurrence, while the

other comes from the Latin “capsa” and “capere” that refers to a box or the verb to hold. The first etymology of the word is representative of a dynamic understanding of case as an event situated in a complex context. The other etymology reflects a more static or rigid framework in which one can simply “put” content. These two different etymologies of the word “case” are useful in understanding the continuum of interpretations and conceptual meaning of this term in the literature.

Definitions of Case in the Literature

Although authors from different disciplines have elaborated on what cases are and how to use them to teach in different contexts, this review surveyed definitions of cases in the areas of teacher education and professional fields such as medicine, law, and business, which have been using cases in their formal teaching curriculum for over twenty years. The goal is to review concepts and themes associated with the use of cases in teaching within a set of different domains in order to identify commonalities in their conceptualization and in the transmission of knowledge.

Cases in Education

The review begins with literature in the field of education that considers the teaching of K-12 curriculum to future teachers. Table 1 is a summary of the definitions and related terminology from this review.

Table 1

Definition of Case in Education

Authors	Definition and Related Terminology
Shulman	"A case has a narrative, a story, a set of events that unfolds over time in a particular place" (p.21). Three types of cases proposed: prototypes, precedents and parables (1986).
Merseth	"A case is a descriptive research document, often presented in narrative form, that is based on a real-life situation or event" (p.2). Purposes of cases: exemplar, opportunities to practice analysis and problem solving, to trigger personal reflection (1994).
Carter	"Consist of events, characters, and settings arranged in a temporal sequence implying both causality and significance" (1993, p.6). Four categories of cases: exemplar, problem situation, story, narrative (1999).
Kolodner	Cases are like movie scripts that include "a setting, the actors and their goals, a sequence of events, the expected results, what actually happened (if it is known), and explanations linking outcomes to goals and means" (2006, p.226).

In Education, early references to the use of cases in teaching can be traced back to the 1930s (Sperle, 1933); however, growing interest in the topic came in response to Shulman's presidential address at the American Educational Research Association (AERA) conference in Chicago in 1985. In the publication of his speech the following year (Shulman, 1986), Shulman proposed that a case needed to be more than a report of an event, but a theoretical claim that implies that the instance of case has generalizable content worth teaching. Shulman later provided a more general definition: "A case has a narrative, a story, a set of events that unfolds over time in a particular place" (Shulman, 1992, p. 21). He argued for the power of case literature to shed light on interwoven practical and theoretical knowledge and proposed three types of cases: prototypes, precedents, parables. Prototype cases mainly exemplify theoretical principles, precedents capture and communicate principles of practice, and parables are used to convey norms and values of a community of practice. Both Shulman and Sperle

argue that the power of using cases lies in the way in which a case is explicated, argued, dissected, and reassembled. This approach of blending the case as content and as a method of instruction might explain why so few proponents of case-based instruction make a clear distinction between what constitutes a case and how it is used.

In a review of the topic in Education, Merseth defines a case as "a descriptive research document, often presented in narrative form, that is based on a real-life situation or event" (Merseth, 1994, p. 2). This definition highlights three fundamentals of cases: (a) they originate from real-life situations, (b) they are the product of meticulous research, and (c) they allow learners to develop multiple perspectives of classroom reality. Merseth also proposes three categories of cases that are classified by purpose: cases as exemplars, cases as opportunities to practice analysis and problem solving, and cases to trigger personal reflection. Her classification of cases is similar to the one proposed by Shulman, but her definition and emphasis on case needing to be based in real event contrast with work of previous authors (Shulman & Colbert, 1987; Shulman, 1992; Shulman, 1986; Sperle, 1933). Merseth emphasizes the research-based nature of cases yet acknowledges that this base is very small.

By contrasting teacher education with other fields that use cases for instruction, Carter (1999) has attempted to define the term case more precisely. In earlier work, she defined cases as stories that "consist of events, characters, and settings arranged in a temporal sequence implying both causality and significance" (Carter, 1993, p. 6). Later, she proposed four different views of what a case can be: a case as an exemplar, a case as a problem situation, a case as a story, and a case as a narrative. Her cases as exemplars are similar to Shulman's prototype as an application to show concrete examples of theories or principles. Cases as problem situations are stories that examine and illustrate the complex contexts of teaching practices. Though the difference between cases as exemplars and problems is not well elaborated, Carter refers to these two

types as conventional uses of cases where the cases are subordinate to propositional knowledge. She characterizes these uses of cases as artificial since they are primarily created and selected to be clear and are classified as a segment of the curriculum while referring to narrative and stories as more authentic and rich. The narratives are lived experiences or personal life stories that could be told from first-hand experience, yet the definition of cases as stories does not exclude narratives. When referring to stories and narratives, she refers to “storied knowledge” (p.7) that is richer and more natural than the other types of cases and notes the autonomous status of story as an expression of knowledge. She advocates against the selection of cases on the basis of stories that fit into the curriculum, and advises a broader use of cases to free the stories of the framework. Carter’s argument is analogous to the dichotomous conception of cases as identified from the two main etymologies: cases as either free form narrative or rigid exemplar made to “fit” the curriculum.

In contrast to holistic and nonspecific views of cases as stories, case-based reasoning researchers have tried to define and build intelligent systems using very structured cases. Influenced by research on experts’ knowledge structure and memory, the case-based reasoning approach was initiated by computer scientists (Schank & Abelson, 1977) who were trying to design expert computer systems and Intelligent Tutoring Systems (ITS). These systems are computer-based problem-solving monitors, coaches, laboratory instructors, or consultants’ tools that provide adapted instruction and feedback to learners (Sleeman & Brown, 1982). Many of these systems use a case-based reasoning approach to store and represent domain knowledge. In case-based reasoning, the information exists in a library of past cases, rather than being encoded in classical rules. Proponents of this approach see the mind as a record of thousands of cases (Kolodner, 1993; Schank, 1998) and propose that humans make sense of new cases by comparing and matching characteristics and features of these new cases with

previous ones. Case based reasoning is also an approach to learning and problem solving (Aamodt & Plaza, 1994). This approach uses the case as a structured way of representing a story or a problem. Kolodner (2006) compares these "real" or "made-up" cases to movie scripts which include "a setting, the actors and their goals, a sequence of events, the expected results, what actually happened (if it is known), and explanations linking outcomes to goals and means" (p. 226). Through experience, people can generalize from cases to new situations that are similar and create scripts and schema for particular types of events. Authors in the case-based reasoning literature emphasize the importance of learning by doing (Kolodner et al., 2003) and the role of experience in anchoring knowledge.

Cases in the Professions

To build on the definition and use of cases in education, it is helpful to look at professions that have used cases for teaching purposes. Exploring the strengths and weaknesses that have been reported over and extended period of time can reveal what is and is not effective in specific instructional settings. Table 2 gives an overview of the definitions surveyed in three disciplines: Law, Medicine and Business. The work of Donald Schön (1983, 1987), in which he examines the use of cases in three other professions, completes the review.

Table 2

Definitions of Case in the Professions

Authors	Fields	Definition and Related Terminology
Moskovitz ,M.	Law	Students study good arguments presented by good lawyers to good judges who write good opinions (1992).
Barrows & Tamblyn	Medicine	Complex, authentic situations that do not have a single “correct” solution (1980).
Christensen, C. R., Hansen, A. J., & Moore, J. F.	Business	Real-world problems and challenges faced by a protagonist or a firm that involve a decision making process (1987).
Schön, D.A.	Architecture, Urban planning, Psychotherapy, etc…	Cases as units of practice that are organized into sets that share family resemblance and cases as units of deliberation and action from which practitioners build up expertise (1987).

The use of cases for teaching Law was presented as the scientific method of teaching by Dean Christopher Langdell of the Harvard Law School in 1870 (Patterson, 1951). The use of the case method is widespread in the US and its adoption can be linked to dissatisfaction with the previous legal education system and with the epistemology of the Common Law system in most of North America (Williams, 1992). The method of casuistry, on which the practice of modern law in many English-speaking countries is based on, dates back from Aristotle (384–322 B.C.) and is defined as the use of reasoning to resolve ethical issues by applying general rules of religion and moral to particular instances ("Oxford English Dictionary," 2004). The use of cases or descriptions of specific situations is at the core of the legal profession; under this system, laws are made by judges using previous cases and decision rather than by only applying specific laws and rules voted by the legislation. This legal system originates from the English tradition where judges were given authority to make decisions supported by the traditions and customs of the community in which they were living.

The case method uses legal documentation and process to build teaching cases around specific events from which students can extract and learn rules of law by working on a hypothetical extension of the case (Llewellyn, 1948; Patterson, 1951). A judicial record consists of a summary of the important facts for the case, the decision of the court, and justifications for that judgment. Instruction in Law is teacher-centered and characterized by large group discussion and casebooks are created or selected by instructors for teaching purposes. The selection and sequencing of cases in casebooks varies, but cases represent good examples of judicial reasoning presented to convey principles and rules for specific topics of the domain. A case can be either the original judgment from the court, or an abridged or annotated version of the court judgment (Moskovitz, 1992). Over the years the case method has been adopted in most Law schools in North America, but little empirical research has been done to document the benefits of the method (Teich, 1986). Limitations of the case method have been identified as presenting a fragmented perspective of the field, putting students through a long initial period of confusion, taking too much time, and not addressing skills like problem identification or interviewing that lawyers face on a daily basis by focusing solely on judgments (Llewellyn, 1948; Moskovitz, 1992; Williams, 1992).

In Medicine, patient cases are at the core of the apprenticeship model of teaching once novices are on the wards, but the emphasis on cases as the primary vehicle for curriculum is more recent. In the traditional medical school model, which became the standard after the Flexner report in 1919, students spend two years attending lectures on basic sciences followed by two years of clinical work in which they learn to apply their knowledge to the treatment of patients. Other medical schools use a problem based learning (PBL) model in which the use of cases or problem scenario is central from the first year on. Approaches and methods related to problem-based learning vary depending on context and purposes (Barrows, 1986), but the original

approach to problem-based learning is characterized by an inquiry type of learning where a small cooperative team is guided by tutors that act as facilitators with the goal of developing students' clinical problem solving and self-directed learning abilities. Under this approach, a good case is defined as a complex, authentic situation that does not have a single "correct" solution (Savin-Baden & Howell Major, 2004). Though some authors emphasize the need for problems to combine both theory and practice (Boud, 1985), there is a consensus that the focus should not be on the product, but instead, on the process leading to the solution. The problem-based learning approach aims to promote students' engagement in their learning process and prepare them to be able to handle the uncertainty and challenges that characterize the treatment of real patients.

In business schools, the case method was first used in Harvard Business School around the 1920s (Barnes, Christensen, & Hansen, 1994). The aim of using cases is to provide students with practical experience to help them develop decision-making abilities for the real world. Cases employed in the classroom can be told in narrative form and are usually presented to students in an incomplete state. They are used to trigger discussion and to determine what decisions and actions should be taken in hypothetical contexts. This method aims at fostering the students' construction of their own framework to approach, understand and reason with business problems (Christensen et al., 1987). According to these authors, the case method highlights the numerous and constantly changing conditions that influence business processes and emphasizes the contextual and subjective perspective of the different business parties involved. When using cases in Business, it is important to acknowledge the different social perspectives, as any given problem can be understood and framed in a different way by individuals and groups and to consider that these perceptions may evolve (Herreid, 2007).

Schön's work examines multiple professional disciplines; therefore, we present his work in this section, but because his work is also essential to the education literature,

it could be classified in either section. Trained in philosophy, his approach to the topic of learning is rooted in epistemology and based on the study of practitioners' performance and reflection. He refers to cases as instances of messy problems found in practice and he contrasts them with school-based problems that have well defined and clear solutions. He criticizes the school-based approach to learning as one that promotes the uses of defined methods on well-formed problems to produce predetermined solutions. Schön advocates for case-based instruction believing it facilitates the inquiry process, or as Dewey (1916, 1933) termed, the "learning by doing" process. Cases can facilitate the inquiry process, where learners engage in informal experiments and conduct trials and revisions based on their hypotheses and data collection. Schön describes reflection in and on action during problem solving that requires the interplay of thought and actions to develop understanding and generate context appropriate solution(s). Schön also proposes cases as units of practice from which practitioners organize their experience with variations of sets of case types. This process leads practitioners to develop routines and repertoires of cases. This "knowing in practice" usually leads to tacit pattern recognition found in more expert practitioners. Schön cautions that this automatization without proper reflection can lead to over-learning, rigidity, and boredom. Both the process of inquiry and the interaction between thought and action are essential to continuous improvement and learning through cases.

Case as Content or Method?

The various understandings of the concept of case as an element of curriculum design can be placed on a continuum related to the two etymologies of the word. Priming on one end of that continuum is the concept of case as a container or fixed structure, and on the other end of the continuum, we have the concept of case as an event in flux. This review however, also shows how cases as content are closely related

to their use in a given context: the nature and function of cases cannot easily be taken apart. In Law and Medicine, references are more clearly made to the use of cases as a method for teaching or learning (i.e., case method or problem-based learning) whereas in other fields like Education and Business, cases were emphasized as stories or problems in narrative form presented to students for analysis.

This tendency to integrate the function of cases in each context shows a clear connection between the curriculum, as a fixed container, and the method of instruction, as event in flux. Cases are tools that can convey domain knowledge and the critical processes needed to translate disciplinary knowledge into every day practice (Shulman, 1992). The next section identifies common functions of case across learning situations to shed light on shared underlying constructs of learning.

Common Instructional Goals of Case Based Learning

Previous literature on case based learning has documented the differences between the use of cases across different domains and even within individual fields (Savin-Baden & Howell Major, 2004; Williams, 1992). Instead of focusing on the differences between the uses of cases, we want to focus on what these approaches have in common. The identification of these common recurrent themes across contexts can provide insight into shared underlying constructs of learning. This work focuses on three common goals that are aligned with current constructivist view on learning: the aim for higher order thinking; the emphasis on the process, not solely on the outcome leading to the answer; and the nurturing of the development of multiple perspectives with no unique “right” answers.

Aiming for Higher Order Thinking

Many authors state that the ultimate goal of using cases in teaching is to promote higher order thinking skills with the emphasis and perspectives varying depending on the author. In Education, Merseth (1994) acknowledges that cases are intended to trigger personal reflection, promote the learner's development of multiple perspectives of reality, and provide practice at analyzing and generating solution(s) for specific situations. Shulman (1992) promotes the use of cases because it help students to develop skills of critical analysis and problem-solving skills and Schön supports this approach to learning as it enables the inquiry process and the thinking and reflection in and on action. In Law, the case method aims at teaching basic knowledge of Law along with essential thinking skills that are required for students to analyze, summarize, and communicate relevant information related about cases (Williams, 1992). In Medicine, the role of cases in developing the learner's clinical problem-solving abilities is central to promoting students' engagement in their own learning process (Barrows & Tamblyn, 1980), while in Business, a similar argument revolves around developing learners' decision-making abilities for the real world (Barnes et al., 1994). In other words, authors across domains refer to higher order thinking as abilities that can be grouped into four main categories: problem solving abilities, learning to learn, the ability to reflect on situations, and to reflect on one's own performance. This enumeration of goals is aligned with current cognitive research about learning and expertise; learning, as suggested from research in cognitive sciences, "is an active process of mental construction and sense making" (Shepard, 2000, p. 6). The concept of learning does not solely focus on the memorization and recall of information, but also involves the ability to use, adapt and apply this information to new contexts (Herman, 1997). Other important aspects of

learning involve learners' organization of knowledge, problem representations, use of strategies, and self-monitoring skills (Pellegrino, Chudowsky, & Glaser, 2001).

Emphasis on the Process

Another common element mentioned in almost all case definitions is the value of a decision-making process leading to possible resolution or outcomes for a case. In both Shulman's (1986) and Sperle's (1933) work, the discussions in which the case is dissected and explained are just as or even more important than the problem itself. The role and importance of the inquiry process in Schön's (1983) work is evident; when discussing how to design cases for repertoire-building, he includes the inquiry process: what arguments were used, what other competing options and moves were considered to the initial statement, action taken, and results achieved.

In Law, the importance of the reasoning and argument process is emphasized by professors' in-class interactive presentations of the case to model argumentation and use of key evidence that lead to defensible motion for a case (Patterson, 1951). Even if few students actually get involved with the professor's teaching in class, every student is expected to be prepared. This process of analyzing and preparing arguments for each case being discussed in class is essential to the success of learners in the case method (Llewellyn, 1948). In Medicine and Business, the role of the instructor in guiding the discussion and integrating more than one perspective for any given case is also emphasized (Barrow, 1988; Christensen et al., 1987). Shifting perspectives of different groups in business and evolving variables in patients' conditions in Medicine also emphasize the need for flexible reasoning skills and the ability to conceptualize and arrive at more than one solution.

This emphasis on the process resonates with the perspective that learning is an active process in which learners and context need to interact. Case-based learning

focuses on the design of interactions that are structured to model and discuss how and why certain decisions are made in specific contexts. Learning is no longer oriented on the content of the telling that teachers do; it now revolves around designing task and situation in which learners develop their understanding and skills around specific situations. Under the socio-constructivist perspective, the notion of learning has shifted from being focused on the content delivered by the teacher to the development of meaningful opportunities for learners to build and improve their knowledge and skills (Dochy & McDowell, 1997).

Multiple Perspectives and no Unique "Right" Answer

Another shared theme in the use of cases in teaching is the indeterminate nature of the final solution for any given case. In Law, the concept of different possible answers is formalized in all higher court instances through the documentation of judges' disagreement on cases. It is also emphasized by teachers' interactive presentations of hypothetical extensions of the case when they defend different perspectives or change key elements of the case to show why different solutions are defensible (Patterson, 1951). In Medicine and Business, the importance of integrating more than one "good" perspective for any given case is also emphasized (Barrow, 1988; Christensen et al., 1987). Shifting perspectives, whether it is about taking the perspective of the medical patient or different client groups in business, provides opportunities to show how different understandings and priorities influence the possible, potential acceptable solutions. Merseth's description emphasizes that one of the three fundamental goals of cases is to allow learners to develop multiple perspectives of the reality (Merseth, 1994) additionally arguing against using the case's stories to make them fit specific, predetermined "good" answers that are aligned with a rigid curriculum.

The acknowledgment of multiple perspectives and the concept of having many possible answers for a case can be linked to discussions about what counts as knowledge. These discussions on the nature of knowledge are linked to the topic of epistemology in philosophy. Epistemology defines not only the nature of the knowledge but also how it is acquired (Fenstermacher, 1994). According to Fenstermacher (1994), there are three types of knowledge: theoretical comprehension of scientific arguments, practical competence of general craft, and practical wisdom and insight to handle particular legal and medical cases. Discourse on epistemology in case-based learning can be found explicitly in Schön's work (Schön, 1983). He refers to practical knowledge as a different type of knowledge requiring a different epistemology, and he warns against the prevalent use of 'technical rationality" as the prevalent paradigm that tries to apply research-based theories to problems and tasks of everyday practitioner's work. For Schön, knowledge is acquired through action, both by doing and by reflecting on experiences. The current constructivist theory of learning is philosophical — one that views knowledge as a human construction (Knorr, 1981). Under a constructivist paradigm, there is an acknowledgement that social, cultural and personal factors influence knowledge. Learning is defined as active sense-making, and learners are encouraged to discover concepts and facts for themselves (Brown, Collins, & Duguid, 1989).

These three themes of fostering higher order thinking, emphasizing the process and the multiples perspectives rather than a unique reality, underlie the use of cases across contexts. They reveal the types of knowledge emphasized and the way knowledge is conveyed using cases in learning situations. The concept of case, as surveyed in education and professional training, goes beyond the notion of simple curriculum representation: it incorporates elements of enactment of that curriculum that are compatible with the current constructivist theory on learning; yet, these instructional

goals support learning principles and perspectives that pose a challenge to current assessment practices. In other words, the correspondence or link between these instructional goals and the way learning is assessed is problematic.

Assessment in Case Based Learning: The Missing Link

Assessment of student's knowledge is reported to be a challenging task throughout the literature on cases (Lundeberg & Yadav, 2006; Sudzina, 1999; Sykes & Bird, 1992; Williams, 1992). This challenge is due to both a philosophical and practical misalignment (Savin-Baden, 2004). At the philosophical level, the conceptual learning framework of case-based learning conflicts with the learning framework underlying current assessment practices. At the practical level, current assessment practices encourage students to exercise rote learning at the expense of deeper understanding. This misalignment hinders the achievement of learning outcomes endorsed by the case-based learning approach, sending contradictory messages to learners about the depth of knowledge necessary to fully comprehend a subject while hindering the credibility of the instructor's delivery of instruction. It is not a secret that assessment dictates what is valued and rewarded in the classroom; it is a powerful motivational tool that influences behaviours of students.

The nature of tasks and requirements embedded in the assessment process influences students' approach to learning (Beckwith, 1991; Collins, 1990). Case-based learning promotes a competence-oriented learning where learners are acknowledged as active players in the learning process, while the development of their own learning goals is influenced by their perceived goals for the class. These perceptions, which shape their learning behaviours, are strongly influenced by outcome measures (Boekaerts, 1996). The current assessment practices do not measure outcomes of learning that support the

goals and instruction of case-based learning approach. The need to send a coherent message to learners is embodied by the idea of "constructive alignment" (Biggs, 1996). Biggs argues that a shift toward constructivist instructional goals requires a congruent change in corresponding assessment methods. Even if this need is well understood from a theoretical perspective (Cizek & Gary, 1996; Dochy & Moerkerke, 1997; Shepard, 2006), its implications at all levels of the educational system remain a challenge. A brief review on the origin and different purposes of assessment will illustrate the complexity and subtleties of the present day's heterogeneous understanding of assessment and related standards.

Origin and Development of Assessment

The practice of assessment was first introduced in China in 200 B.C. as a means of selecting bureaucrats. In this context, the development of standardized assessments or "tests" was a groundbreaking way of making selections that were previously based solely on genealogy (DuBois, 1964; Gipps, 1999). The origin of the term assessment resonates with the two paths of historical development identified by Glaser and Silver (1994) in the context of education: 1) for selection and placement, and 2) for measuring educational outcomes. From an operational perspective, the former focuses on identifying capabilities at an individual level prior to instruction, while the other aims at assessing educational outcomes after instruction. In both situations, assessment is conceived as tasks of evaluation separate from instruction (Cizek & Gary, 1996). In these assessment contexts, the instruction and assessment tasks are separate activities whose responsibilities are given to different people. Teachers are responsible for instruction whereas assessment belongs to the realm of measurement experts (Birenbaum, 2003). This conception of assessment is rooted in earlier theories of learning and measurement techniques (Pellegrino et al., 2001). Cizek (1996) posits that

the current constructs of learning and achievement are no longer aligned and synonymous due to each one's different socio-historical development; however, while the relationship to learning of these two constructs has evolved differently, it is important to understand that they are both intrinsically related to instruction in any learning context (Birenbaum, 2003).

Stakeholders and their Distinct Purposes

Each learning situation involves the students who are taught, the instructor who instructs, the subject matter presented, and the context in which the learning takes place (Posner, 1985; Schwab, 1971). Learners, teachers, administrators, certification boards, and the general public all have different views on how and what should be learned and assessed; they do not share a common understanding of the primary purposes of assessment, nor do they prioritize the same quality standards. Assessment is defined as the process of making a judgment according to specific goals, criteria and standards (Scriven, 1967). A distinction is often made between summative and formative assessment. Summative assessment occurs after learning has taken place and is used to document performance outcomes for grading purposes (Black & William, 1998). Formative assessment occurs during the instructional process, and it typically aims at improving teaching and learning (Sadler, 1998; Shepard, Hammerness, L., & Rust, 2005). However, the lack of clear theoretical distinction between these two types of assessment (Taras, 2005) has led to a greater attention to different forms and purposes of assessment (Knight, 2006).

Assessment in the educational system serves multiple purposes: measuring prior knowledge, assisting the learning process, measuring individual achievement, selection of individuals, and program evaluation. Most educators and researchers believe that assessment's primary role should be to support learning (Boud & Falchikov, 2006;

Dochy & McDowell, 1997; Pellegrino, Chudowsky, & Glaser, 2001; Shepard, 2000), but the classroom is not an independent unit; uses and decisions about classroom assessments happening at different administrative levels (program, department, university, ministry) impact what happens in the classroom (Joughin & Macdonald, 2004). Additionally, many programs lead to certification exams designed by external agencies. These licensure exams also influence teaching and assessment practices in the classroom and the many levels at which assessment take place all impact each other directly and indirectly. For example, in Medicine all Canadian students are required to pass the Medical Council of Canada Qualifying Examination at the end of their training. Ignoring the impact that each level has on the other ones can lead to a problematic situation like the impact that assessments at the ministry level has over instruction delivered by teachers in the United States (Pellegrino & Chudowsky, 2003; Smith, 1991). Much time and attention is spent by instructors and students on practicing for upcoming tests even when teachers perceive these tests as being invalid (Smith, 1991); moreover, the notion of accountability, implemented at many levels of the educational system where the performance of educators and schools is evaluated based on the results of the students, influences instruction delivered by teachers. These negative consequences of the “hyper-test consciousness” created by “ institutional test anxiety” limit flexibility and sovereignty for classroom instruction to focus on learning (Baker, 2007). The educational system is complex and interconnected; therefore, change in the classroom cannot be successful without conscious alignment of the purposes of assessment (Pellegrino et al., 2001).

Standards and Criteria

Standards and criteria have been developed to help ensure the quality and the fairness of assessments. Traditionally, the criteria for judging test material are validity

and reliability. Validity and reliability are not only measurement principles, they represent social and scientific values used to make judgments (Messick, 1995).

Reliability refers to the extent to which an assessment tool measures consistently — that is, measuring the same thing, the same way, each time it is used in the same context. While reliability is sometimes included within the concept of validity, it is necessary but not sufficient to ensure the validity of a test (Herman, 1997). Traditionally the goal of reliability is to quantify and predict the precision of the results (Haertel & Herman, 2005). Depending on the context, purpose, and type of instrument used, the estimation of reliability will consist of 1) the test-retest reliability, 2) the inter-rater reliability, or 3) the internal consistency reliability (Trochim, 2000). True score measurement theory, underlying the concept of reliability, assumes that any measurement reflects the score and some error of measurement. Under the classical test theory, assessment tools and tasks had to meet strict standards of measurement error. However, these strict criteria have been revised through the performance assessment movement of the 1980s and 1990s where variability was inherent to the nature of the context (Linn, 1994).

The other criterion, validity, refers to the accuracy with which an assessment is measuring or capturing what it claims be measuring (Cronbach, 1971). At the general level, valid assessment is defined as a task(s) that generates accurate inferences about students' learning and accomplishments which can be generalized to a larger domain of knowledge or skill (Herman, 1997). The estimation of the validity of an instrument is concerned with the constructs that the test measures as well as how well it measures them (Anastasi, 1996). Validity is a well debated concept (Smith & Fey, 2000) and one that has many dimensions; however, for practitioners who are responsible for using standards for classroom assessment, it is usually divided into three categories: content validity, criterion validity, and construct validity. Content validity refers to how well the

assessment measures a representative sample of the domain; criterion validity refers to the predictive value of the test or how well it can predict other related variable; and construct validity refers to how well the assessment measures the ability, behaviour, skill, and related theoretical constructs that are intended to test.

Validity and reliability are concepts that are continually examined and debated by researchers and while it is the practitioners who mainly guide the empirical use, the administrators and teachers have limited resources and time to dedicate to philosophical and theoretical debate of these standards. Sireci and Hambleton (1997) confirm this in stating that while validity is not the sole property of the test, it incorporates the meaningfulness of the inferences derived from the assessment score which involves teachers and administrators. Aligned with the consequential aspect of validity, Frederiksen and Collins (1989) propose the idea of systematic validity to support behaviours and inferences from both teachers and learners through the assessment process. In their system, the concept of transparency and openness advocates that the assessment process and criteria should be clear enough to enable learners to understand and be able to assess their work or performance to some degree. While it is worth noting that researchers are making considerable effort to present a simplified and usable framework for evaluating assessment practices to practitioners (Lissitz & Samuelsen, 2007), the debate between completeness and simplicity of use is still ongoing (Kane, 2008). The purpose here is not to provide details on the debate regarding technical qualities, but instead to convey the complexity and theoretical implications of these standards.

Challenges Posed by Current Assessment Practices

Though the use of cases in teaching offers the potential to meet the challenges of today's evolving knowledge, its key instructional characteristics pose a challenge to

current assessment practices by challenging measurement theory assumptions and underlying theories of learning. I discuss how the higher order thinking skills and problem solving abilities at the heart of case-based teaching cannot be identified nor measured by standard measurement practices (Pellegrino, Chudowsky, & Glaser, 2001). Another shortcoming of current assessment practice is related to the lack of attention given to the argumentation process that leads to the resolution of a case. Additionally, teaching students the correct answer within cased-based instruction is not the goal of this learning system; however, without "right" or "wrong" answers, the concept of variability, as conceived by current psychometric paradigm, is treated as a measurement error.

Defining and Measuring Higher Order Thinking Skills

The concept of "knowing" goes beyond memorizing pieces of information and involves being able to use, adapt, and integrate knowledge in new situations (Herman, 1997). In other words, authors in case-based learning across domains refer to higher order thinking as abilities that can be grouped into four categories: problem solving abilities, learning to learn, and the ability to reflect on situations and to reflect on one's own performance. While the focus on development of theoretical foundation of complex cognitive skills has taken us beyond what behaviorism would have promoted, it has not addressed the challenge of defining, identifying, and measuring higher order thinking (Airasian, 1997). Higher order thinking skills (HOTs) are often referred to as the three upper levels of Bloom's taxonomy: analysis, synthesis and evaluation (Ennis, 1993; Savin-Baden & Howell Major, 2004); the lower three levels are: knowledge, comprehension, and application (Bloom & Krathwohl, 1956). This distinction between higher and lower levels however, is somewhat artificial as the lower level is both a pre-requisite to the higher level and cyclical in nature. Bloom's taxonomy has been criticized

as vague, especially for assessment purposes (Ennis, 1993). Ennis has proposed a list of eleven steps enabling the evaluation of critical thinking. He argued that this list of eleven steps combined with an essay type test target generalizable skills across domains. However, the focus on reading tasks in Ennis's framework might have limited its impact in other domains. Despite the numerous limitations of Bloom's taxonomy written in 1957, it has still retained its influence on administrative, practitioners and research communities, and is still one of the most commonly cited works. The wide ranges of research and methods available to identify and judge higher order thinking reflect diverse perspectives on current theories of learning and human performance as well as the varied practical demands of contexts in which they are used. Most of the references to learning and thinking listed above integrate elements from both the situative and cognitive theoretical perspectives on learning; however, many of the concepts used in assessment practices originate from the behaviorist and differential approaches to learning, implying direct measurement of behaviour (Pellegrino et al., 2001; Shepard, 2006).

To measure, in science, means to link observations to theoretical terms or constructs like weight (Haig & Borsboom, 2008). Output is generated through observational procedure and the interpretation of these outputs informs us about the concepts of interest. The conceptual framework of measurement procedures of physical concepts that can be observed directly, like duration or height, are usually not questioned in the uses of most contexts with the exception of particle physics. However, in psychology, unobservable psychological constructs like learning — for which measurement procedures like multiple-choice tests, essays, or portfolio are used — require an explicit connection between the concept(s) being measured, the method, and the interpretation of the output attained. The link between the concept, the output

measure, and the context in which it can be used, relates to the construct validity of the measure or instrument. It is often referred to as the nomological network of the assessment. This network describes how the theoretical framework of the concept relates to the measures and observations (Cronbach & Meehl, 1955). Current assessment tools were developed under a psychometric conceptual framework which is based on a trait approach to learning (Schuwirth & der Vleuten, 2006). The theoretical framework behind these statistical tools limits their application to new concepts like higher order thinking.

Measuring the Process, not Solely the Outcome

The emphasis on the process of case discussion is another important aspect in case-based teaching. The importance of argumentation and justification of decisions throughout the discussion of cases are not supported by the outcome-oriented nature of many current assessment practices. The focus on grading the answers without giving much attention to the process leads learners to focus on grades (Boud & Falchikov, 2006). Many important cognitive aspects of the learning process are not well captured by current static assessment practices including students' use of strategies and self-monitoring skills (Pellegrino et al., 2001). Exclusion of reasoning process elements that lead to problem resolution in assessment challenges the construct validity and the consequential validity of assessment measures. Without acknowledging the process leading to the measured outcomes, assessment practices cannot document how well learners do at reasoning throughout the resolution of the problems, nor does it enable feedback on the process by pointing to where learners might have gone wrong. This lack of emphasis on the process does not promote the importance of looking back at one's reasoning steps to reflect on the strengths and weaknesses of elements in the reasoning sequence.

Acknowledging Multiple Perspectives

The concept of supporting reasoning and argumentation leading to different possible answers is characteristic of case-based learning in all professional fields. The importance of the ability to shift perspectives and evaluate knowledge as it evolves emphasizes the need for dynamic reasoning and the ability to develop the learners' ability to conceptualize more than one possible solution. In the positivist paradigm in which most psychometric methodologies have their origin (Herman, 1997), knowledge is obtained by objective hypothesis testing. Researchers in this paradigm consider that truth and belief are intrinsically different and they aim at revealing "reality" that can be objectively investigated. The methods used in educational assessment try to approximate methods from the "hard sciences," which explains why a lot of early work used primarily quantitative and correlation methodologies. Assumptions of replication and reliability at the core of these methodologies are challenged by the problem solving situations in case-based learning because the same case may not be solved the same way by the same person twice; moreover, variability in the answers and ways to get to these answers — which is a desirable outcome when teaching with cases — adds to the challenge of measuring outcomes given that traditional assessment approaches treat variability as an error (Moss, 1994; Schuwirth & der Vleuten, 2006).

Case Based Learning and Assessment in Medical Education

Medical education represents a unique environment to study case-based teaching and assessment as it has successfully implemented problem-based learning — a type of case-based learning — for the past 30 years. Singularly, clinical instructors in this field are required to have expertise in a specific field of research and maintain an active practice while teaching. The number of assessment tools developed and experienced, along with recent discussion on the nature of competencies, support discussions about the need to explore new modes of assessment.

Case Base Teaching Tradition in Applied Settings

Medicine as a discipline has a long-established tradition of using patient cases to teach in both classroom and clinical settings. Although medical schools may differ in terms of when and how cases are introduced to students, the use of cases is widespread as a way of transmitting knowledge (Cox, 2001). Case-based teaching in medicine is often equated to problem-based learning but it only represents one particular approach to teaching using cases. Problem-based learning (PBL) implementations vary depending on context and purposes (Savin-Baden & Howell Major, 2004); however, the original approach is characterized by small cooperative teams where tutors act as facilitators with the goal of developing learners' clinical problem solving skills, and self-directed learning abilities (Barrows, 1986). Similar to proponents of the case method used in Law schools, early proponents of the PBL method had a prescriptive approach regarding the definition of PBL, but the field has moved beyond focusing on which method is better and has begun asking questions about the nature of learning associated with specific features of PBL (Norman, 2004).

Practical and Theoretical Co-Requirements for Clinical Instructors

Another interesting characteristic of the medical education field is that clinical instructors are not only required to be expert in a specific field: they are also required to maintain an active practice. Unlike other professions like Law, Education, and Engineering where academic roles favour and promote theoretical over practical knowledge, medicine acknowledges the importance of practical knowledge for teachers. In these other fields some experience might be considered a positive asset at the hiring process, but none of these professions enable nor require professors to have an active practice while teaching in the academic setting. The requirement is related to the setting in which teaching occurs; an apprenticeship model of knowledge transmission. The apprenticeship model is based on the gradual introduction and integration of individual within a community of practice (Lave & Wenger, 1991). Under this model, novices learn the trades or professions by observing and gradually being introduced to different tasks with increasing levels of difficulty and complexity. Clinical instructors have to provide care and teach concurrently; wearing two hats, they are both attending physicians and teachers for any given case and must juggle a list of explicit and implicit tasks and responsibilities with limited time and resources. This double identity has implications for the notion of competence and for assessment practices in this field.

Revised Notions of Competence and Expertise

The notion of competence in medicine is complex. It not only involves elements of performance and skills around patient-care delivery but also involves notions of professionalism, leadership, and scholarly contribution to the community. Elstein and Hundert (2002) have defined competence as “ the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in a daily practice for the benefit of the individuals and communities being

served" (p.2). The Royal College of Surgeons in Canada has developed a framework for the *expert clinician* that requires the expert clinician to coordinate six different roles around effective patient-care delivery (Frank, 2005). Expert clinicians need to be good communicators, collaborators, managers, health advocates, scholars, and professionals. In his review of assessment in medical education, Epstein (2007) discusses specific challenges related to these roles and emphasizes that competence is not a state that one achieves or possesses but a life-long habit that needs to be fostered and developed. Adding to the developmental nature of competence, he also stresses the contextual nature of competence. The contextual factors influencing the performance of clinicians include the environment in which the physician is practicing, the presenting features of the patient's illness, as well as demographic characteristics of the patient and the physician.

The study of the structure and acquisition of expertise in various domains aims to inform decision-making as well as facilitate and improve the training of novices in these domains. The range of theoretical frameworks that have guided the study of expertise have emphasized different aspects related to individual aptitudes, extended experience, different representation and organization of knowledge or superior learning environments (Ericsson, 2006). More recent models of expertise suggest that deliberate practice, not only time and practice, is required to achieve consistent, measurable and reproducible expert performance (Bransford & Schwartz, 2009; Charness et al., 2006; Hatano & Inagaki, 1986). Under the prevalent cognitive approach in medical education, expertise has been conceptualized as a state of mastery of knowledge and techniques in specialized areas. Mylopoulos and Regehr (2007) suggest updating the current framework to incorporate the notion of adaptive expertise. Under this paradigm, knowledge is not a static notion and the status of expert is not simply a developmental stage due to years of experience. The development and maintenance of expertise is

conceived of an “approach to practice” where knowledge is a dynamic resource that is used, built and continuously adapted to new situations. This expertise framework emphasizes the notion that expert performances are dynamic and fallible. Studying the way experts use their knowledge in a dynamic context, not only what knowledge they possess, might lead to better understanding of how to foster expert-like performance.

Assessment Methods

Research on the contextual and developmental nature of competence has implications for the assessment of knowledge and skills in work-based settings. To measure and orient the development of medical competence, the field has developed and experimented with a great variety of methods and tools (Epstein, 2007; Nendaz & Tekian, 1999; van der Vleuten & Schuwirth, 2005). Research on assessment in medical education has evolved from a framework that had a rigid hierarchical perspective of the types of knowledge to be assessed (Miller, 1990) to a more complex and comprehensive programmatic view on assessment (van der Vleuten, 1996). The research focus is moving from trying to establish which method works best to how best use and combine appropriate assessment methods in comprehensive assessment programs (van der Vleuten & Schuwirth, 2005).

One of the remaining challenges identified by authors in recent reviews is the psychometric challenge related to circumstances where there is no consensus on a single correct answer or no consensus on ways to reach an answer (Epstein, 2007; Schuwirth & van der Vleuten, 2006; van der Vleuten & Schuwirth, 2005). In this context, Medicine as a field has been identified as an ill-defined domain (Pople, 1982). Diagnostic reasoning, which is an important aspect of competence in Medicine, involves resolving patient cases for which there are the presence of unknown or evolving elements, no single unique unambiguous final diagnosis, and more than one way to

reach that diagnosis. The current psychometric framework is built on the assumption that medical competence is a combination of constructs like knowledge, skills, problem solving, and attitudes (Schuwirth & van der Vleuten, 2006). These unobservable skills, which are referred to as constructs, are treated as traits and are assumed to be stable, generic and independent (Schuwirth & der Vleuten, 2006). As the concepts of stable and generic skills have been rejected, the notion of competence has taken over the idea of stable knowledge and skills (Elstein, Shukman, & Sprafka, 1978) while competencies in medicine are defined as tasks that a qualified physician should be able to do or perform successfully. The following section will explore how the use of an inductive competency based approach to case-based reasoning can encourage the development of models of competent clinical teachers by building on context specific situations.

Research Objectives

Current educational assessment practices are based on psychometric tests that focus on “the answer” without attention to how learners reach an answer. In contrast, case-based practices recognize the importance of the process in producing a solid, defensible, and acceptable answer adapted to a given context. The context-specific nature of reasoning via cases also emphasizes that there are many ways to reach an answer, and potentially multiple, appropriate answers. The absence of a “right” answer and the desirable variability in the answers of case-based learning contexts challenges the notion of reliability as defined and used by current assessment methods. We will briefly introduce the context and objectives prior to elaborating on each research question and describe how the main objective aims at aligning assessment to instructional goals and thereby addressing some aspects of the assessment challenge in case-based learning contexts.

To address this problem, we propose to use an evidence-centered design approach (ECD) (Mislevy, 1994) to lay the foundation for a case-based learning assessment system. The evidence-centered design framework is based on the use of evidentiary argument principles. This system imparts that assessment is conducted using imperfect data; therefore, the process is conceived as a programmatic and longitudinal use of complex data to sample and reveal patterns of reasoning. One of the key components of the conceptual assessment framework is the evidence model (Mislevy, Steinberg, & Almond, 2003) where this model contains evidence rules that are directives for interpreting the quality of specific answers for a given task (Mislevy et al., 2003). This ECD system however, has traditionally been used in well-defined domains like science, for which a deductive approach can be used successfully to determine what

key elements and processes are required to assess the quality of an answer to a problem.

This study aims to adapt an evidence-centered design by using an inductive approach to create case-specific proficiency models. The problem solving task is conceptualize as a performance and the dynamic decision making processes of mere experts is scrutinized to establish similarities and differences. The study is anchored in a representative task of knowledge transmission in medicine through the study of physicians' competent case resolution protocols used to create blueprints of competent problem solving. These blueprints correspond with the initial requirement of building evidence models in the context of an evidenced-centered assessment system. By reaching a better understanding of the problem space and strategies related to competent performances, these evidence models can inform the design and evaluation of student models.

Two main questions precede and guide the analysis leading to the design of these models of competent performance:

1. What is the degree of consistency/convergence between protocols of expert medical instructors on patient diagnosis?
    a. Do they show agreement in their solution outcomes? If so, in what ways?
    b. Do they show agreement in their solution processes leading to their diagnosis of patient cases? If so, in which ones?
2. Do expert medical instructors evaluate key elements of their reasoning processes similarly? In other words, does the use of differential weighting improve the level of convergence about the key elements leading to the resolution of ill-defined patient cases?

Modeling Case Specific Problem-Solving Performance

Medicine, as a discipline, has a long-standing tradition of modeling expertise through tutorials that are based on real patient cases. Case presentation is both a formal and informal practice in medicine (Greenhalgh & Hurwitz, 1999). A case presentation in medicine generally consists of a detailed analysis of a patient case, but depending on the instructor's prior experience and the facilities in which the patient is seen, the solution to these cases varies substantially. As instructors present a patient case to students, they think out-loud and thereby externalize their thoughts for the social purpose of instruction. By verbalizing their reasoning to others (justifying their diagnostic reasoning), instructors are performing a type of "think-aloud" protocol that can be a great source of information for learning and research. Analyzing the content of case presentation can inform us about the types of knowledge and skills that constitute good performance; we aim at documenting and building on key elements related to this practice.

We draw a parallel to the literature on problem solving where the seminal work of Newel and Simon (1972) leads the conceptual and methodological foundations to discuss and study how people solve problems. Generally, the act of solving a problem is defined as a situation where one is trying to attain a goal for which no simple and obvious means is known (Newell & Simon, 1972). Literature on problem solving typically makes a distinction between well-defined and ill-defined types of problems (Jonassen, 1997; Newell & Simon, 1972; Voss, 2005). Well-defined problems are characterized by a unique verifiable solution, and where the initial description usually includes all the necessary information and constraints to solve the problem and attain the goal state. The series of steps in between the initial state and the goal state is referred to as the solution path that represents the problem solving process where analysis through the problem space is typically guided by hypothesis formulation and hypothesis testing. On

the other end of the continuum, ill-defined problems involve vague or undetermined goals, an unlimited number of constraints, and many sub-goals and phases that require the storage and manipulation of a large quantity of information to reach a solution that typically is not right or wrong nor final (Voss, 2005).

Diagnostic reasoning about patient cases can be conceptualized as ill-defined since these cases usually involve the presence of unknown or evolving elements about the patient and the development of the disease, have no single correct unambiguous final diagnostic, more than one way to reach a diagnosis or final decision, include multiple ways to reach an acceptable answer, and require the ability to handle a large amount of conceptual and practical knowledge to integrate and understand the numerous elements involved in the patient problem situation.

Using the “search through the problem space” analogy, the process of how competent physicians solve ill-defined problems is studied through the use of external problem-solving representation for specific cases. The goal of using external problem representations is to synthesize the entire process thereby enabling a reflection not limited to the memory of participants. We chose to study ill-defined problems for which there are agreed upon answers but many ways to get to the answers. Exploring the convergence and variability of both process and outcomes aims at identifying case specific similarities and differences of valid performance. The goal of the visual representation is to gain insight about the problem solving process by capturing rich descriptions and explanations occurring throughout the decision-making context. These informal models are built through iterative analysis and design involving participants and using their ability to reflect and select key elements in their reasoning processes.

The method used for this inquiry is situated within the realm of cognitive task analysis (CTA) methods. CTA integrates task analysis with data about knowledge, thought processes, and goals into the analysis of the task performance (Schraagen,

Chipman, & Shalin, 2000). Typically, CTA involves three phases: knowledge elicitation, analysis, and representation of knowledge (Crandall, Klein, & Hoffman, 2006). Recent work reviewing the use of different methods recognizes the importance of context with an interest in finding innovative ways of combining methods to produce better results (Hoffman & Lintern, 2006). The present adaptation of the cognitive task analysis incorporates knowledge elicitation, analysis, and knowledge representation in an iterative design. We use a computer-based learning environment, BioWorld (Lajoie, 2009; Lajoie, Lavigne, Guerrera, & Munsie, 2001), to simulate the interactive presentation and collection of data related to the patient cases. Concurrent protocol analysis is combined with a retrospective protocol analysis (Ericsson & Simon, 1984) to collect, analyze, and model problem solving processes of participants.

Purpose of Models

The goal in building models is not to formalize the problem space of specific ill-defined problems but to inform an understanding of successful resolution process performance. A model is an abstract representation of a specific aspect(s) of a complex phenomenon or reality built for specific purposes. A roadmap is a good example of a model whereby it represents a geographical area by symbolically highlighting special relationships between physical elements. The distance between City A and City B on a map symbolically represents the actual physical distance between them in reality; depending on the ratio of the map, the details about the different roads one can take to travel from City A to City B will vary. Having a map detailing every turn one encounters when traveling the distance may not be practical, possible, nor appropriate for orientation purposes.

Models are tools for studying and understanding complex realities or phenomenon; this study aims at building empirical qualitative models, not quantitative

ones. Models in problem solving research tend to be quantitative and prescriptive; they are built from an artificial intelligence perspective that tends to focus on well-defined types of problems. From this viewpoint, models need to be prescriptive to enable objective testing of theories using a computational program. As discussed by McCarthy (1977), the epistemology of problem-solving is compatible with a realist and an empiricist perspective. In the present situation, modeling from an empiricist perspective would aim at connecting data with actions and thoughts as they unfold whereas a realist perspective would try to find facts about a world that exist independently of the model and represent the "truth" in the domain of study. In this study the empiricist perspective is favoured given the state of the understanding of the human reasoning processes, the limitation of current computational languages, and the researchers' lack of domain knowledge. The purpose is to build descriptive models of ill-defined problem solving processes prior to addressing the numerous challenges related to the design of formal models of these problem-solving phenomena. As suggested by the ECD framework, the use of probability and neural networks may later be appropriate to deal with the complexity of problem solving performance, but it is believed that the initial design of the models requires more flexibility as all of human reasoning cannot be translated into a formal representation (McCarthy, 1977).

How to Model

For these purposes and context, the empirical perspective combined with the lack of ability of formal languages to deal with the contextual and complex nature of natural language limits the use of formal languages for the phenomena of interest. Given that the aim is not to design formal models of the phenomenon, the use of natural language combined with an informal mode of visual representation is appropriate for the elaboration of the models. It is also important to stress that the intent is not to build

mental models or to try to unravel the internal knowledge representation of participants, but rather to build an external representation of the performances. The empirical perspective also applies to the discourse analysis, where pragmatics and shared meaning-making are used to guide the analysis and summation of event from the protocol. The design of the visual representations attempts to preserve the chronological stream of stories and explanations from participants to reveal some convergence on how and why they resolve cases the way they do. Since the purpose and the perspective on knowledge differ from previous problem-solving and medical reasoning research, the way the discourse is analyzed also differs from semantic or propositional analysis (Frederiksen, 1975; Patel & Groen, 1986; van Dijk, 1985).

What to Model

According to ECD, the design of the system requires a detailed understanding of the performance to be assessed (Mislevy, Steinberg, & Almond, 2003). By documenting and comparing expert teacher problem-solving performances, the goal is to build an empirically-based evidence model of what the shared parameters or elements of a good performance look like. The analyses of the reasoning performances integrate elements of quantitative and qualitative analysis to reveal the problem space and strategies of ill-defined problems. Instructional experts were chosen for this study since they are well trained at communicating their judgment strategies to an external audience. Unlike other types of experts, experienced instructors have the ability to break down the diagnostic reasoning process into comprehensible sub-units, and provide appropriate explanations for their decisions (Weiss & Shanteau, 2003).

Explicit Notion of Reliability

It is important to address the notion of replication necessary to any measure of reliability (Brennan, 2001) prior to describing the convergence of the data in this study;

however, the notion of replication in the context of problem solving is not straightforward. Asking a participant to solve the same case twice may not lead to an identical performance. In light of this and to increase the chances of comparable performance, “easy” types of problems were chosen according to the level of competence and experience of participants who were expected to provide an accurate final diagnostic using the five performances as five replications of a valid answer for each case.

Degree of Convergence in the Analysis of Protocols: Outcomes and Process Analyses

Empirical assessment performance is accomplished by detailed protocol analysis based on relatively small numbers of subjects (Embretson & Gorin, 2001). Data were comprised of the computer trace that documents the problem solving steps, verbal data from the think aloud as participants solved the problem, and video data that integrated the computer screen with the audio transcripts. Through the analysis of data and the construction of individual models for each case, the analyses focused on how solution processes and outcomes converge and diverge for each case. The analyses are designed to provide a telling representation of how experts teach and diagnose patient cases and look at the emergence of possible evaluation criteria. Analyzing the performances of multiple experts for each case aims at unfolding a more complete problem space of competent performance in which solution processes are identified and compared in terms of convergence or divergence and actions and explanations.

Comparing Participants’ Categorization of Key Elements of their Reasoning Processes

A recent model of expertise suggests that deliberate practice, not only time and practice, is required to achieve consistent, measurable, and reproducible expert performance (Bransford & Schwartz, 2009; Charness et al., 2006). In medical education, Mylopoulos and Regehr (2007) suggest that the development and maintenance of expertise is conceived of as an “approach to practice” where knowledge is a dynamic

resource that is used, built, and continuously adapted to new situations. This expertise framework emphasizes the notion that expert performances are dynamic and fallible. Studying the way experts use and reflect on their knowledge in the context of performance, rather than the amount of knowledge they possess, might lead to a better understanding of how to foster and assimilate an expert-like approach into practice. The design combines strengths of both concurrent and retrospective, think-aloud methods (Ericsson & Simon, 1984) by capturing performance and later allowing a participant to revisit and comment on it. The hypothesis is that participants' reflection on their reasoning process for each case can lead to shared identification of key elements for a case; moreover, expert instructors are expected to have a greater level of agreement on what elements are absolutely necessary and important versus what are important and useful elements.

This study does not examine expert judgment on an ideal or proposed common answer but rather the reflection of experts on their own reasoning representations. It explores how the analyses of the reasoning performance can show a degree of convergence and reliability in and among competent individuals in the field. Instructional experts have experience at determining the optimal and sub-optimal characteristics of learners' performance; thus, it is estimated that these experts will be able to identify and discriminate between the key elements that lead to their own solutions.

METHOD

The research is anchored around the specific task of case presentation, which is an authentic form of teaching in medicine (Greenhalgh & Hurwitz, 1999). A case presentation generally consists of an instructor offering a verbal, detailed analysis of the decision making process related to a patient case to an external audience. I build on this task to examine the resolution process of competent physicians as they solve a case in a computer-based learning environment. The use of a simulation enables participants to interactively explore patient cases through the ordering of diagnostic tests, requesting information about vitals, visiting the library, or asking for a consult. Using this type of open ended computer simulation can be considered as a more authentic task than the typical paper case used in most medical reasoning studies given that the problem-solver can explore and select actions instead of being constrained to respond to information in a linear fashion.

Quantitative and qualitative measures are used to get a better understanding of the reasoning performances of physicians as they solved cases using a simulation. The mixed method analyses serves to provide a clearer understanding of ill-defined problem solving, integrating elements of the problem space as well as strategies related to competent performance in these case-based contexts. In this section, I first review the variables of interest corresponding to the two main questions that look at the convergence of outcome and process, and measure the details about the categorization task. Then, I explain the methodology in a sequential manner, dividing the different steps of data recording and analyses into phases. The explanation uses samples from the coding process to show concrete examples of how the data is analyzed to lead to the design of individual and merged visual representations.

Degree of Convergence in the Analysis of Protocols: Outcomes and Process

The performance measures for the case resolution include both the outcome and process measures. The goal is not to oppose these measures but to explore how they can both inform the interpretation of valid performance in an ill-defined problem-solving context. Outcome measures in the problem-solving task of Bioworld involve the final diagnosis submitted for each case along with the list of evidence supporting this final diagnosis. Since the cases chosen are the types of ill-defined problems for which there is a well agreed upon answer but no single right way to reach this answer, I include the list of evidence to corroborate that there is agreement on how or why the diagnosis is reached. I set the agreement level at over 50 percent, which means that a minimum of three participants out of the total of five need to use the same argument or evidence to count as a convergent element. For each of the three cases I compare the amount, type, and relative importance given to each component of evidence for each of the five case resolutions. I also include time in this section — not as an indicator of performance, which is often used in medical problem solving studies, but as a descriptive measure to enable a fair comparison of performance.

To gain insight into the process, I examine the measures from the interaction with the computer learning environment as well as measures from the verbal protocol. For each case I compare the list of evidence selected, the list of tests ordered, and the list of hypotheses selected along with the range of confidence levels that are recorded throughout the resolution process. In the protocol, I look at the number of words to enable a fair comparison of the number of lines and number of episodes for each case. I also consider the number of common episodes categorized by each participant as key elements for the resolution of each of the cases.

Comparing use of Differential Weights in Categorization Task

In this section, I explore in more detail how medical instructors select and categorize the elements from their problem solving protocol. The expectation is that the reflection of participants on their reasoning process for each case can lead to shared identification of key elements for a case; moreover, the hypothesis is that expert instructors will have a greater level of agreement on what elements are absolutely necessary elements versus what elements are important and useful elements.

Analysis and Design of Models

The idea of using an external visual representation of performance to summarize the data came from Henderson, Yerushalmi, Heller, Heller & Kuo (2003). They used a multi-layered concept map to organize the set of ideas and analysis based on the discourse interviews of six participants. For the conceptual analysis of their interview data, they used a visual representation to summarize and gain perspective on a macro level. This technique enabled the researcher to deepen and share their analysis without loosing a connection to the raw interview data. Using multiple layers to enable the summary of data for our case models is a concept that is built upon here. Going back to the roadmap modeling analogy, I propose a methodology that incorporates technologies and that takes advantage of this macro level idea by enabling a zoom feature which moves in and out from one level of detail to another thereby enabling the interaction with the visual representation in a similar way to what geographic information systems (GIS) do through GPS or interactive maps on the Internet (Google Map, Mapquest, etc.). Additionally, I use the visual representation in an iterative cycle of analysis that involves participants in the analysis and thereby frames the cognitive task analysis in a transactional view, where the analysis is shared with participants.

Participants

For our purpose, an expert is someone with “prolonged or intense experience through practice and education in a particular field” (Ericsson, 2006). For this study medical expert teachers are defined as individuals who have an excellent grasp of the domain knowledge, sufficient amount of exposure to solving these types of cases, and are recognized by their peers as excellent teachers. Upon recruitment, the five participants were asked to complete the consent form and a questionnaire that documented descriptive data about their areas of expertise, recent clinical experience, and overall teaching experience. (see Appendix A for consent form and ethics and Appendix B for a copy of the questionnaire). Teachers were selected for this experiment to circumvent the problem of automatic reasoning, also known as knowledge encapsulation, that medical experts exhibit in when doing cases in their area of expertise (Rikers, Loyens, & Schmidt, 2004; Schmidt & Boshuizen, 1993). We selected experts in a specific area of medical expertise that matched the types of cases used for the study. Following their case resolution, participants’ experience with similar cases was queried to enable a better interpretation of their performance given the content specific nature of diagnostic expertise (Elstein et al., 1978). Participants were not remunerated for their participation.

The five experts are all internal medicine practitioners, three men and two women. Expert 1 is a gastroenterologist with 10 years of experience who mainly teaches graduate students. In the week prior to the experiment this expert had seen eight patients. However, none of these patients were new admits and thus there was no need for complete diagnosis investigations. Expert 2 is an internist with 26 years of experience who teaches both undergraduate and graduate students. In the week prior to the experiment this expert had seen approximately 18 patients and admitted two new ones. Expert 3 is an internist with five years of experience who teaches both

undergraduate and graduate students. In the week prior to the experiment this expert had seen 20 patients and 10 new admissions. Expert 4 is an internist with 28 years of experience. In the week prior to the experiment this expert had seen approximately 35 patients and 15 new patients. Expert 5 is an internist with 37 years of experience who mainly teaches undergraduate students. This expert had not seen any patients in the week prior to the experiment.

Study Duration, Recruitment, and Task Completion

This study was conducted in two phases. A pilot study in 2007 was the first phase with two participants. Material, design and analysis were tested and minor changes were made to the research protocol. Four minor changes were made to the second phase in 2008: one, I added the recording of the screen capture video of participants to improve the transcription phase and add the option of looking back at what participants were doing in the simulation when saying something about any given aspect of the problem; second, the task of categorization, which initially required participants to use a coloured pencil to select an element on a paper representation, was adapted to a direct computer interaction with the representation in the Cmap software (This was more efficient as participants did not show any difficulty in using the software for the validation phase.); third, the error prediction task that followed the categorization task was abandoned due to time limitations; and fourth, the inclusion criteria for recruitment were narrowed down to general internal medicine and not sub-specialties. These changes did not affect the task per se but improved the way the recording and analysis were done.

The recruitment of participants was challenging given the small number of expert physicians meeting the criteria. Recruitment was done through email and solicitation at public presentations but it took over a year to recruit three additional participants and the

delay in this study led to a technical problem for expert 5 in case 2. Case material was modified for a case leading to an identical case description but one with a different test result. After analysis, the researcher realized that the reasoning for the case had been influenced by this result and could not be used for the analysis.

Programs Used for Data Collection and Analysis

Case-Based Learning Environment - BioWorld and Case Builder

BioWorld is a computer-based learning environment (Lajoie, 2009) that provides a realistic patient case simulation where users have to diagnose patient cases by interpreting and collecting a patient's symptoms, conducting diagnostic tests, and collecting appropriate information in the library and consult sections. The companion-authoring tool, CaseBuilder (Lajoie, Faremo, & Wiseman, 2001) enables the creation and modification of cases and related content presented in BioWorld. These systems are designed and developed as part of our research activities on cognitive tools to support learning.

Computer-Assisted Qualitative Data Analysis Software (CAQDAS) - Transana

Transana (Woods & Fassnacht, 2007) supports the transcription, coding, and analysis of digital data sources. This software enables the researcher to attach time markers to both audio and video segments, providing a mechanism to coordinate information sources so that participants can be watched and heard as they interacted in specific moments of the problem solving task. This feature of providing a way to refer back to segment units enables better transparency and collaboration for the data analysis. Transana was chosen over other programs because it is open source software based at Michigan State University, has a growing community of users providing support, and was free at the start of this study.

Graphical Tool Software - IHMC CmapTools

Cmap is a graphical tool used to construct, navigate, and share knowledge representations (Novak & Cañas, 2006). The concept map representation underlies the design of the Cmap tool but I did not select the software to develop concept maps per se, but rather to explore software that could be used to create multi-layered graphical representations. I chose Cmap over other more powerful software programs because I needed a program that could be easily used by our participants; additionally, this software is free, available on both PC and Mac OS, and is supported by the Institute for Human and Machine Cognition (IHMC), a university research affiliated institute.

Screen Capture Software – Camtasia Studio

Camtasia Studio is a screen capture program that enables both sound and sections of computer screens to be recorded dynamically. This software is used to help with transcription and to enable a better collaborative use of data by providing a more detailed context of transcribed action and verbal behaviours. This product was chosen because it is supported by a university-wide license ("Camtasia Studio," 2003).

Data Collection and Analysis: Overview

The researcher individually met each participant twice: once to solve the cases and a second time to validate and reflect on the draft of their performance. During the first encounter, data collected included a questionnaire about their background and recent experience, the log of the participant's computer interaction, a screen capture of the video and audio recording of their think aloud protocol from each case resolution task. In the second meeting, data collected included validation of the visual representation of the reasoning process for each case, categorization of key elements from the visual representation, a post-questionnaire for each case about their disease specific knowledge, and observational notes from the researcher. The design and

analysis of the visual representation for each case was an iterative process where each participant reviewed the analysis of the researcher prior to selecting and categorizing elements of their reasoning path for each case. Data collection and analysis occurred in five phases. In phase one, the participant engaged in a “think-aloud” process while solving the case; in phase two, the researcher built a visual representation incorporating computer log data and verbal transcript into the layers; in phase three, the participant was asked to review her or his visual summary before selecting and applying weights to sections that were crucial for the resolution of each case; in phase four, the researcher coded individual transcripts for each case to add the categorization; and in phase five, data were analyzed for convergence and a visual representation from all participants for the same case was combined into one multi-layered representation.

Phase 1: Capturing the Case Resolution Performance

Structured Interview

The structured interview simulated the authentic activity of case presentation for participants. In this case analysis simulation task, I captured the teaching discourse for analysis, reuse, and comparison. According to Ericsson (2006), the use of a representative task has shown to improve the accuracy of the data collected in think-aloud experimentation. During the initial meeting, participants were asked to solve cases and do a think-aloud, explaining their actions in the same way they would present a case to undergraduate medical students.

Solving Cases in a Computer Learning Environment

I used the computer-based learning environment BioWorld (Lajoie, 2009), to present a set of three cases to each participant. Bioworld provides an interactive platform for participants to explore patient cases by reading the case history, ordering diagnostic tests, requesting information about vitals, visiting the library, or asking for

assistance. Solving a case in BioWorld is an ill-defined task as there is no one way to solve the case and each participant can decide to order tests and consult different information; yet, it is still a constrained problem space as the tests, hypothesis, library items, and consults available are limited. I ensured that the list of available information was extensive enough to avoid participants from guessing answers by process of elimination, but the simulation was not fully representative of the actual patient case resolution in effect on wards. While each participant solved and performed a think-aloud, interactions with the computer program were recorded by the computer log. Their speech in context was also recorded by a screen capture video. I briefly describe the possible actions that participants could engage in while going through the simulation and the computer log files were coded and used in the analysis of each participant's performance. There were six possible types of actions recorded from the participant's interaction with the computer program. The list of codes in table 3 below describes the type of log actions that are recorded in the context of participants' interaction in BioWorld.

Table 3

List of Action Log Recorded in BioWorld

Code in Verbal Transcript	Description
add evidence	Selecting text and adding it as an evidence for the problem resolution activity
add test	Ordering a test from the list of possible options
switch area	User changes screen or program section. There are four different areas: problem, chart, library and consult.
select hypothesis	Selecting an hypothesis from the list of possible options
change hypothesis conviction	Selecting or changing the percentage of confidence of the chosen hypothesis
submit hypothesis	Submitting the final answer

Solving a case in BioWorld requires diagnosing the medical condition(s) affecting each patient. In addition to selecting a final diagnosis with a corresponding level of confidence, participants have to add evidence that they collect during their case resolution, whether it be patient symptoms, relevant patient history, or diagnostic tests that they order pertaining to their hypotheses. Moreover, they adjust their differential diagnoses and confidence level dynamically as they go through the case prior to submitting a final diagnosis and confidence rating. When they decide to submit their final diagnosis, they are required to select and prioritize the evidence to support and justify this diagnosis. Figure 1 below shows a screen shot of the interface of the problem area where participants initiate the problem solving task.

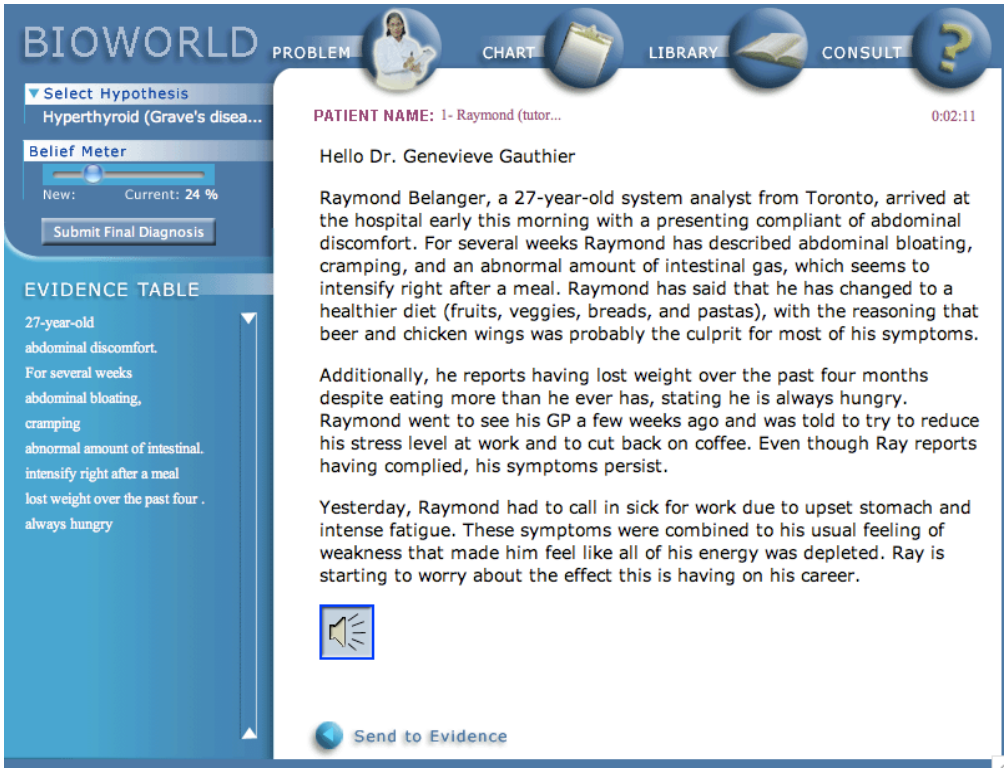


Figure 1. Screenshot of BioWorld interface

Case Description

A set of three endocrinology introductory level cases was used for this study. These three cases all had relatively clear answers but there was no unique way to get to these answers. Previous studies in medical reasoning have focused on difficult cases and used an expert/novice framework to assess answers and study the types of mistakes made be novices (Elstein & Schwarz, 2002), however this study focuses on studying the decision-making process of successful performances and therefore uses non discriminative cases. The cases represented typical instances of diabetes, hyperthyroid, and pheocromocytoma. The first case, pheocromocytoma, described a 37-year-old woman presenting with symptoms of anxiety, headaches, and episodes of palpitations, sweating and flushing. She was on hypertensive medication and had lost 10 pounds in the last 4 months. The second case, the diabetes case, described a 16-year-old active teenager who was experiencing extreme fatigue, had difficulty seeing, was

feeling thirsty, and urinated more frequently. The third case, the hyperthyroid case, described a 34-year-old woman who had been having anxiety attacks, was feeling nervous, and was experiencing episodes of excessive sweating, hand tremors, and the sensation that her heart was racing.

Case Rating

After having resolved each case participants were asked to rate the level of difficulty of each case for different audiences: 1st year medical students, 2nd year medical students, 3rd year medical students, 4th year medical students, residents, and for practitioners. For each audience, they had to classify cases as being: too easy, a good revision, challenging or a difficult one. This information is useful to control for perceived difficulty of each case at the practitioner level.

Phase 2: Coding and Representing the Case Resolution Performance

Following the participants' resolution of each case, the researcher transcribed the video and combined the verbal transcript, video, and computer log into one protocol prior to coding it. The transcription and coding was done using Transana. As shown in figure 2, Transana's interface has four quadrants: the top left is for the audio data, the top right for the video data, the lower left for the transcript, and the lower right for the coding and analysis. Figure 2 shows that when a section of the transcript is selected on the bottom left on the screen, the corresponding audio and video segment become available.

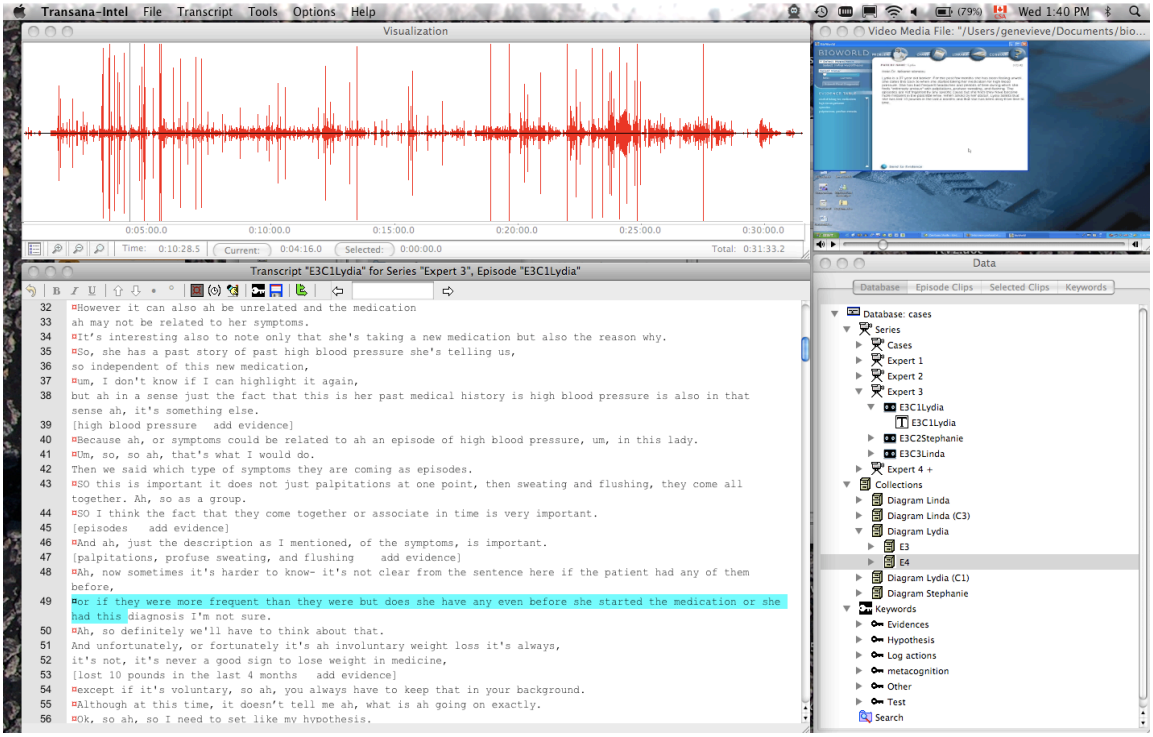


Figure 2. Screenshot of Transana interface

The first step was to transcribe the data from the think-aloud using an abridged and adapted convention notation from Atkinson and Heritage (1984). In the transcript, participants were identified as “E” and the researcher as “R” when instructions were given or when questions were answered. Comments were put in parenthesis () and double question marks indicate inaudible sections. The computer log actions were integrated into the transcript in brackets [action log]. The second step was to insert time codes into the transcript. Time codes are links, or bookmarks, corresponding to exact moments on both the audio and video file; they are instrumental for selectively calling up any portion of the video or text by simply clicking on any point in the text. In the third step, the transcript was segmented into lines using the “unit idea” (Bransford & Franks, 1971). Using this concept of idea, each segment corresponds to a single complete idea, or a single block of information whether it corresponds to a word, a clause, or a phrase.

Additionally, each action from the participant's interaction with the computer program was considered as a separate segment.

In the example presented in table 4 below, time codes, which are invisible in the coding program, are converted into millisecond values in parentheses. The example also demonstrates how the segmentation of the transcript into lines evolved using the unit idea. Aligned with this concept I also decided to segment each action log as a separate entity. The "clicks" or log action from the participant's interaction with the computer program were considered a separate segment as shown by the example of expert 3 solving case 3.

Table 4

Example of Segmented Transcript From the Verbal Protocol of Expert 3 on Case 3

1	R. (Researcher mumbling instructions ??)
2	E. (reading not audible ??)
3	(08.8) E: Ok. So again, ah just ah if I start from the case.
4	So here they mention that's it's a healthy 34 year old woman.
5	[healthy 34-year-old woman add evidence]
6	So again, I always put it in my evidence because it's my baseline—where do I start from.
7	(32.3) And then here we have the duration,
8	for two months she's been having those anxiety attacks.
9	["anxiety attacks" add evidence]
10	(42.1) So this is what the patient is telling me,
11	let's see what she really describes.
12	So she feels nervous for no reason,

In the fourth coding step, these actions from the computer log were used to anchor episodes that were summaries of a set of ideas representing meaningful steps that contributed to the resolution of the case. These episodes corresponded to evidence, hypothesis, test, plan, summary, explanation, request, and questions in the problem-solving representation. These summaries were usually, but not always, anchored on an action recorded on the computer log (e.g.: in table 4, lines 1 to 5 are anchored to the

action of add evidence). Log actions correspond to possible actions that participants do while going through the simulation, they correspond to: add evidence, add test, switch area, select hypothesis, change hypothesis conviction, submit hypothesis (see table 3 for more details).

Using the same section of verbal protocol in table 5 below, I show how the episodes were created in Transana: lines 1 and 2 are grouped together as an episode (and a future node on the visual representation) as instructions. No utterance or lines are discarded even if deemed non-relevant to the problem-solving task, thus the entire transcript is retained in the analysis. Lines 3 to 6 are grouped together as an episode — healthy 37-year-old woman. In the example below, “collection” refers to the structure of the analysis. Data were analyzed per case and for each case I added data from each expert. “Clip” is the identification of each item (I0_, I1_, I2_, I3_, etc.) that corresponds to the sequence of elements. The time indicates where the episodes are situated in the entire sequence of the problem resolution enabling researchers and participants to watch and listen to only this sequence when querying the data. “Episode node” is a text label provided by the coder to summarize each section of the transcript. In the following examples, the 12 lines of code were divided into four episode nodes.

Table 5

Example of Episodes From the Verbal Protocol of Expert 3 on Case 3

Collection: Case 3 > E3
Clip: I0_instructions
Time: 0:00:00.0 - 0:00:08.8
Episode Node: instructions
1 R. (Researcher mumbling instructions ??)
2 E. (reading not audible ??)

Collection: Case 3 > E3
Clip: 11_age
Time: 0:00:(08.8) - 0:00:32.3
Episode Node: healthy 37 yrs old woman
3 Ok. So again, ah just ah if I start from the case.
4 So here they mention that's it's a healthy 34 year old woman.
5 [healthy 34-year-old woman add evidence]"
6 So again, I always put it in my evidence because it's my baseline- where do I start from.

Collection: Case 3 > E3
Clip: I2_anxietyAttacks
Time: 0:00:32.3 - 0:00:55.5
Episode Node: anxiety attacks for two months
7 And then here we have the duration,
8 for two months she's been having those anxiety attacks.
9 ["anxiety attacks" during add evidence]

Collection: Case 3 > E3
Clip: I3_nervous
Time: 0:00:37.5 - 0:00:55.5
Episode Node: feels nervous
10 So this is what the patient is telling me,
11 let's see what she really describes.
12 So she feels nervous for no reason,

Episode nodes two and three are linked to an action but the fourth is simply linked to a piece of evidence that was mentioned orally but not selected through action. These episodes that are not anchored in actions were created based on two different criteria. The first criterion related to episodes that corresponds to similar elements or topics that are later on in the protocol anchored as action. The second criterion corresponds to episodes that relate to independent elements of the problem solving that cannot be linked to previous or following episodes. For example, in episode 3 above, the

nervousness is an independent element that is not clearly linked to the anxiety attacks by the participant. The following section in the protocol, which is not in the example above, is about the excessive sweating experienced by the patient which does not relate to the segment of the nervousness previously mentioned. The text summarizing these segments that I refer to as episodes corresponds to nodes in the visual representation. As shown in figure 5 below, only the text of the node is visible, but the entire episode transcript is included within the episode nodes, keeping a direct link to the raw data accessible through a simple mouse over.

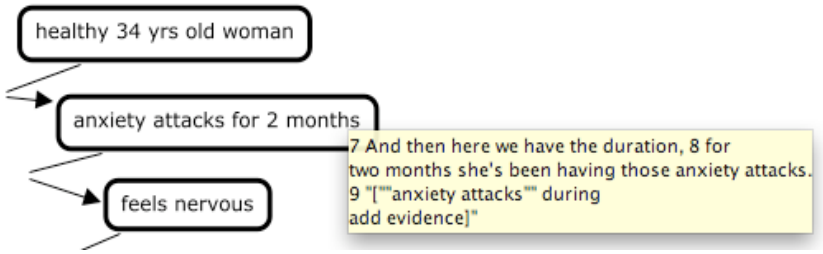


Figure 3. Extract of visual representation of an episode and its related verbal transcript

Other sections of the transcript were more challenging to summarize. The researcher had to make decisions about which section of the transcript corresponded to a meaningful episode for the problem solving process, as in table 6 below illustrating episode 24, expert 3, case 3.

Table 6

Example of a Challenging Episode From the Verbal Protocol of Expert 3 on Case 3

Collection: Case 3 > E3
Clip: I24_controversy
Time: 0:06:20.4 - 0:06:59.0
Episode Node: Controversial whether you should do nuclear scan and ultrasound

113 Now there is controversy should you do or not a nuclear scan of her thyroid gland,
114 and should you do even an ultrasound,
115 I don't know if we have an ultrasound of her neck.
116 Ah, radio- we only have abdominal but should we do an ultrasound of her thyroid, some people would say yes, some people will say no.

In the example above, the episode precedes the action of ordering a diagnostic test of a Radionuclide Scan of Thyroid Gland, and could have been integrated into the subsequent episode 25 but was added as a distinct element because it represented a questioning of the procedure; this questioning might have been relevant for the meaning and interpretation of the diagnostic test ordered. Each line of the protocol was included in one of the episodes for each case. Even when lines referred to technical problems or irrelevant questions, they were included in one of the co-occurring episodes. The inclusion of each utterance and action aimed at assuring the transparency of the coding of these episodes by enabling participants and other researchers to revisit the original lines of transcript included in each episode. This direct link to the raw data was kept in the subsequent graphical transposition of data.

The last step in this phase required the researcher to transfer these episodes and their corresponding lines of transcript into a visual representation using a graphic tool. Cmap (Novak & Cañas, 2006), a user-friendly tool, was selected to construct a multi-layered visual representation for the resolution of each case. Each node in the visual representation as shown in figure 4 below corresponds to an episode; inside each node, the original lines of transcript are inserted and afterwards easily accessible by mouse-over (as shown in the screen capture of figure 3 above). The square box reveals the transcript associated to that node. Version 1 of the visual representation of the case resolution was designed according to the sequence in which the episodes occurred. As shown in figure 4 below, the visual representation was structured sequentially and the links ending with arrows indicated the sequence in which the episode occurred.

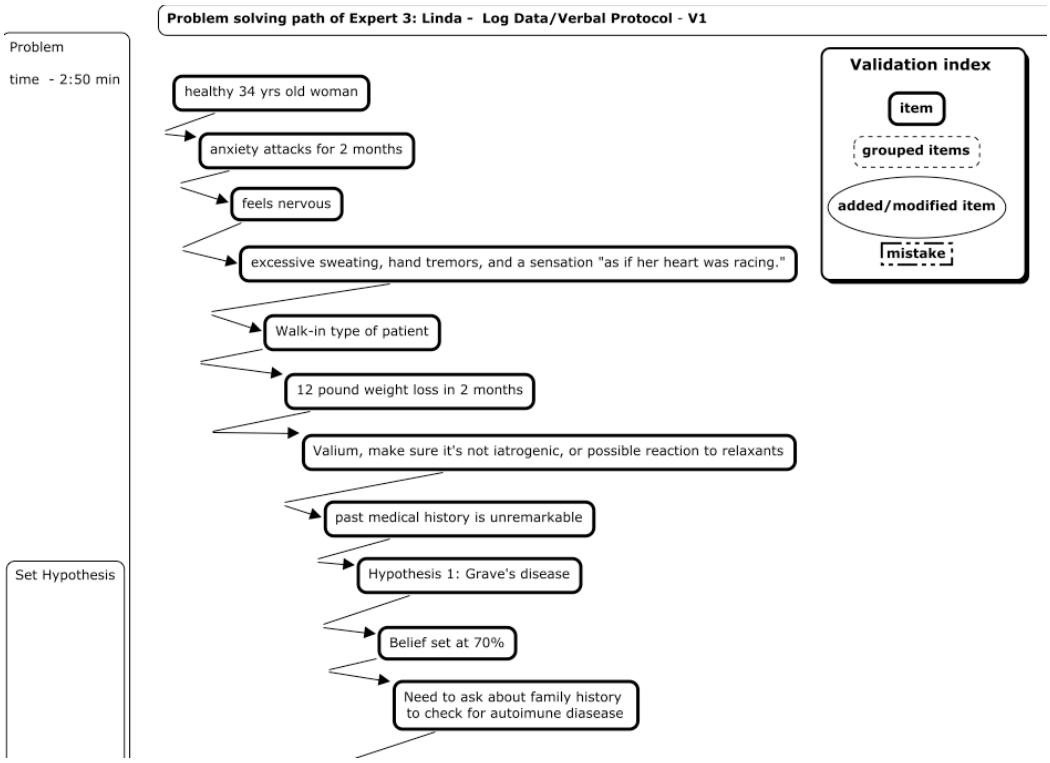


Figure 4. Extract of a section of the visual representation

Phase 3: Validation and Categorization Tasks

In a second meeting, participants were instructed to inspect the visual representation of each case and to verify that the description was congruent with their thinking. At this stage they were told that they could add, correct, delete, regroup, and comment on the content in the summary of the episodes and change the disposition of the episode if needed. The visual representation was a tool that helped communicate data to participants for validation purposes.

After the validation task, participants were asked to reflect on their solutions by categorizing the key steps in their resolution process for each case. Participants were asked to use color codes and tag key elements relating to the level of importance of the case resolution: red for “absolutely necessary,” yellow for “necessary,” and blue for “adds useful information.” Corresponding patterns were associated with the colours to enable a meaningful black and white printing of this second version. Red was associated

with a crosshatched background, yellow with lines and blue with a dotted background. Each colour category was associated with a pattern and a numerical weight (5, 3 and 1) as shown in table 7 below. The use of these specific weights of 5, 3 and 1 as opposed to 3, 2 and 1 were chosen to emphasize the difference of importance between the categories and improve the rate of agreement by giving more weight to absolutely important elements. This weight assignment presupposes that participants would show a better agreement regarding absolutely important elements of the problem as opposed to useful information adding to the problem resolution.

Table 7

Grid of Weights for Categorization of Elements

	Red (+5)	Yellow (+3)	Blue (+1)
Key elements	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)

Phase 4: Categorized Case Representations of Participants

A second version of the visual representation was constructed, incorporating both validation changes and categorizations made by each participant. The validation and categorization by each participant was also integrated into the coding of the protocols in Transana. Figure 5 below shows a section of the categorized individual visual representation.

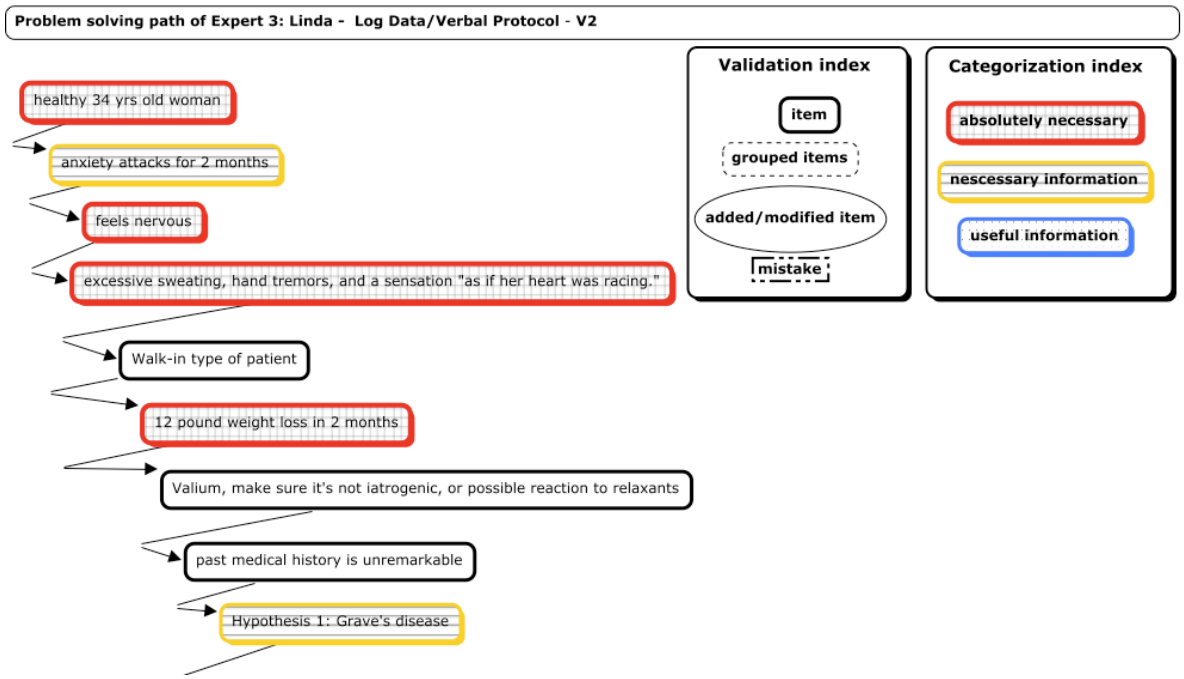


Figure 5. Extract of an individual visual representation after categorization

Phase 5: Analysis and Design of Merged Case's Representations

In this phase, the two research questions guided the analysis of the data protocol. First, the degree of consistency and convergence of the outcome and process data is examined. I analyzed both outcome and process measures to explore how they could both inform the interpretation of valid performance in an ill-defined problem solving context. The comparison of outcome measures in the problem solving task involved the final answer for each case along with the prioritized list of evidence supporting these answers. In contrast, process measures compare the list of evidence selected, the list of tests ordered, the list of hypotheses selected, and the range of confidence levels recorded throughout the resolution process. For the second question, the analysis focused on the comparison of key elements selected as well as the use of the differential weights by experts.

Following these analyses, the researcher combined the individual visual representation of all participants for each case. Prior to merging individual case representation, repetitions and technical moments were deleted to simplify the visual

representation without affecting the overall validity. Technical moment generally referred to segments of the protocol where participants commented or questioned procedures related to BioWorld. Common evidence comprised a third level of summary, created when an episode was categorized as being important (regardless of the category) by more than 50 percent of the participants. Common evidence corresponded to the episodes that were categorized as key element by experts in their respective individual representations. This merged representation attempted to present the paths of multiple participants in order to show similarities and differences in the sequence of decision-making leading to acceptable answer(s) for a specific case. Thus, in these final merged representations there are three levels of data: original protocol transcript, individual participant episodes and evidence, and common evidence nodes. This visual representation was built using categorization that highlighted the similarities and contrasts of the steps of each participant. Data from five experts shown in Figure 6 illustrate each expert's sequence along with groupings of elements that experts had in common.

To create the final combined representation for each case resolution, the five visual representations for each case were copied into one file. To each of the five resolution paths, a different colour was applied, but the background of weighted episodes was kept intact. Expert 1 was associated with grey, expert 2 with lilac, expert 3 with dark red, expert 4 with green, and expert 5 with blue. Each episode was then compared to the ones having similar content and combined as one common evidence when a minimum of three experts had categorized similar episodes as evidence. For example, age of the patient was categorized as important episode by four of the five experts; therefore, these pieces of evidence were merged as a common evidence node. In the visual representations, common evidence nodes, which show where participants agree, correspond to level 1. The episode nodes from each participant correspond to

level 2 of the representation, whereas the transcript inside each of these node corresponds to level 3.

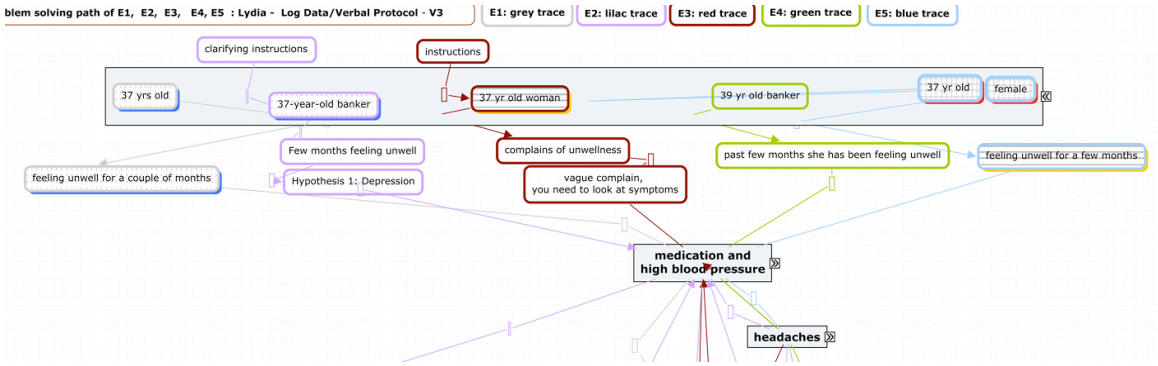


Figure 6. Example of a section of a merged representation

Figure 6 shows a section of a merged representation. This section of the case resolution shows elements at two of the three levels of representation. Level 1 corresponds to the common node shown at the bottom, which is labeled: “medication and high blood pressure”. When common nodes are open by clicking on them, content at the level 2 shows up. An example of level 2 is shown in the blue square in the upper section of the figure, it corresponds to the common node of “37 year old woman”. When this node is opened it shows which episodes were selected and combined from each participant for this common evidence. Level 3, which is not visible in figure 6, would correspond to the original transcript that can be explored in each of the expert’s episodes through a mouse-over.

RESULTS

Medical problem solving is considered ill-structured in that there is not one best way to solve a patient case. The results section presents evidence models of competent performance for five experts on three specific patient cases. The analysis was guided by two main research questions.

1. What is the degree of consistency/convergence between expert medical instructors' protocols on patient diagnosis?
    a. Do they show agreement in their solution outcomes? If so, in what ways?
    b. Do they show agreement in their solution processes leading to their diagnosis of patient cases? If so, in which ones?
2. Do expert medical instructors evaluate key elements of their reasoning processes similarly? In other words, does the use of weights on key elements of their reasoning processes show convergence for key elements leading to the resolution of ill-defined patient cases?

We first present data from both outcome and process measures related to the problem-solving task of three cases. We then examine how experts used the categorization task to select common evidence from the visual representation of their reasoning process leading to the final diagnosis. Finally, we merge the individual expert data into one summary representation that illustrates both convergence and divergence between performances of the experts.

The results section is divided into three sections with each section presenting results for one individual case. Each of these case specific sections begins with a brief case description, followed by the participants' recent clinical experience with the study's type of case, as well as their perception of the case's difficulty from the post-questionnaire. This description situates the problem and the problem solver in the

context of the problem-solving task. We then answer the first research question regarding the convergence of the protocols by looking at the solution outcome and the solution processes to establish the similarities and differences across performances. The following section answers question two by looking at how experts used the categorization task and the last section describes the similarities and differences between the five experts as they diagnose patient case 1 (Lydia), providing a narrative description as well as a merged visual representation for the case. The final merged representation enables the interpretation of the previous measures by providing the context of the performances, demonstrating what was done during the performances as well as why and in what sequence.

Case 1

Brief Background Information: Problem and Problem Solver

Case Description

"Lydia" is a case about a 37-year-old woman who presents symptoms of anxiety, headaches, and episodes of palpitations, sweating, and flushing. She is on hypertensive medication and has lost 10 pounds in the last four months. This is a typical case of pheocromocytoma[1]. However, this disease is very rare and diagnosing this case involves the identification of important evidence and the exclusion of more common diagnoses like hyperthyroid disease, essential hypertension, or panic attacks before confirming and investigating this diagnosis. There are several important diagnostic tests: TSH, T4, and T3 to rule out hyperthyroid disease, followed by Urinary Catecholamines and imaging tests to confirm and investigate the disease.

[1] Pheochromocytoma is an extremely rare endocrine disorder, which usually consists of benign tumors of the adrenal glands. It often leads to over-production of certain hormones thereby raising the blood pressure and heart rate. This condition can be life-threatening if untreated.

Experience of Participant and Case Difficulty Rating

Each of the five participants completed a post-questionnaire after completing the case that revealed their level of familiarity with the case in question. Expert 1 wrote that he had never seen any real patient case of Pheocromocytoma but that he remembered studying about this type of case 15 years ago. Experts 2 and 4 had seen three or four cases in the last 10 years, expert 3 had seen 1 or 2 cases a couple of years ago, and expert 5 had seen five cases of pheocromocytoma in the last 30 years. In the case rating section, participants were asked to rate the level of case difficulty between 1 and 4 (4 being most difficult). Only experts 1 and 2 rated this case as being too easy (1 = too easy). Expert 3 rated the case at 1.5, which falls in between too easy (1) and good revision (2), and experts 4 and 5 rated the case as being a difficult case for physicians (4 =difficult).

Convergence on Protocol Measures

Solution Outcome Measures

Data shown in Table 8 presents the summary of the outcome measures related to the final diagnosis submitted by experts at the end of the case resolution task. From this table we can see that all but one expert submitted a final hypothesis of pheocromocytoma. The confidence level about final hypothesis varied between 50 percent for expert 5 and 60 percent for expert 4, to 100 percent for expert 1 and expert 2. Only expert 5 did not submit the appropriate hypothesis: for this case, he submitted a final hypothesis of panic attack.

Table 8

Overview of Final Hypothesis, Confidence Level and Final Evidence

Case 1	E1	E2	E3	E4	E5
Final hypothesis	Pheocro-mocytoma	Pheocromo-cytoma	Pheocromo-cytoma	Pheocromo-cytoma	Panic Attacks
Confidence level (%)	100	100	65	60	50
List of prioritized evidence in final argument	1. Urinary Catecholamines / Norepinephrine - 89-591 nmol/day 2. Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr 3. Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr 4. palpitations, profuse sweating, and flushing 5. high blood pressure.	1. high blood pressure 2. frequent headaches 3. periods of time during which she feels "extremely anxious" with palpitations, profuse sweating, and flushing. 4. Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr 5. Urinary Catecholamines / Free - 1560 mmol / 24 hr 6. Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr	1. palpitations, profuse sweating, and flushing 2. episodes 3. high blood pressure 4. Urinary Catecholamines / Free - 1560 mmol / 24 hr	1."extremely anxious" with palpitations, profuse sweating, and flushing. The 2. lost 10 pounds 3. started taking her medication for high blood pressure 4. Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr 5. more frequent 6. Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L	1. "extremely anxious" with palpitations, profuse sweating, and flushing 2. few months 3. episodes 4. has lost 10 pounds 5. headaches

All 5 participants had a somewhat similar number of items (range of 4 to 6) listed as final evidence supporting their diagnosis; however, there was low agreement on the nature of prioritized evidence. Only one item, heart palpitations, was top priority on three of the five lists. Comparing which evidence had been selected using a level of agreement of three out of five, two others were added to the consensus: high blood pressure and the Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol /

24 hr. As shown in Table 9 below, using an agreement level of three out of five experts' protocols improved the consensus rate from 19 percent to 42 percent.

Table 9

Consensus Rate of Final Evidence Submitted

	E1	E2	E3	E4	E5	Total	Consensus Rate
Common evidence (5/5)	1	1	1	1	1	5	19%
Common evidence (3/5)	3	3	2	2	1	11	42%
Number of final evidence	5	6	4	6	5	26	

Solution Process Measures

To gain insight into common elements of the diagnostic reasoning process, we first observed the measures based on the actions of the experts while they resolved the case. In Table 10, we show the number of tests ordered and evidence selected throughout the problem solving process. Included are the number and the list of hypotheses selected throughout the resolution process and the final confidence level along with the variation of this confidence throughout the task. The list of hypotheses recorded for expert 5 shows that even though he did not submit the pheocromocytoma diagnosis, it was on his list of differential diagnoses. Additionally, the sequence of hyperthyroid and pheocromocytoma was common to three of the five experts who selected more than one hypothesis. The number of diagnostic tests requested compared to the number of hypotheses considered was of interest here. If testing for more than one hypothesis, the expectation would be to call for more tests, yet this was not the situation. From the actions recorded, the number of tests that experts ordered in common varies from two to three tests.

Table 10

Process Measures from Recorded Actions

	E1	E2	E3	E4	E5
Number of hypotheses	1	2	2	1	6
List of hypotheses	Pheocro mocy-toma	Hyperthy-roid, Pheocro-mocytoma	Hyperthy-roid, Pheocromo cytoma	Pheocro mocy-toma	Hyperthyroid, Pheocromocytoma, Depression,Panic Attack,Hyperthy-roid, Panic Attack
Final confidence level	100	100	65	60	50
Range of confidence level	100	60 – 100	55 - 70	9 - 60	50
Number of test ordered	20	12	14	17	17
Common tests	2	2	3	2	3

Additional process data from the transcription and analysis of the protocol are presented in Table 11. The time taken by each participant to submit final diagnosis for the case is presented along with the total number of words in the protocols, the estimation of the number of words per minute, the number of lines in each idea unit that was segmented, and the number of episodes created for each resolution. Episodes are summaries of a set of ideas representing meaningful steps that contribute to the resolution of the case and that were validated prior to the categorization task by the experts. I also added the number of episodes that each expert considered important enough to categorize as it provided further information about important episodes in relationship to the entire protocol. This table shows how descriptive data — time to completion, number of words, and rate of speech — were segmented into lines and reduced into episodes that were later categorized.

Table 11

Process Measures from Protocols

	E1	E2	E3	E4	E5
Time (min)	12:18	25:19	26:06	17:49	17:54
Number of words	2010	4308	4117	3012	1514
Rate of speech (word per min)	165	171	158	172	86
Number of lines	233	501	371	221	145
Number of episodes	44	65	48	38	35
Categorized episodes	21	24	15	24	30

On average, participants took 20 minutes to complete case 1. Time to complete the task may reflect of the amount of variability in the quantity of verbalization by experts. In this case, expert 1 took considerably less time than the others to complete the task. The time measure on its own might suggests that expert 1 was either very quick in solving the case or not as thorough in his reasoning performance but that perspective changes when we look at the number of words in the protocol. The total number of words provided us with an overview of the length of the transcript with significant differences between expert 5 and expert 1 who kept their verbiage to around 2000 words, to expert 4 at around 3000 words to expert 2 and expert 3 at over 4000 words. More interesting was when we divided the number of words per time it took to solve the case. Even if expert 1 and expert 5 were considered to be both slower talkers or affected by the think aloud process, only expert 5 showed a significant difference in terms of his rate of speech. His rate of speech is slower across cases, which is probably a personal characteristic and not reflecting any particular challenge for this case.

The difference between experts decreased when the transcripts were segmented into distinct expressed ideas. The number of lines varied from 145 for expert 5 to 501 for expert 3. Expert 2 showed a considerably higher number of distinct ideas expressed. It was of interest to compare expert 2 and expert 3 who used a similar number of words,

yet 130 lines of ideas in difference. The concept of idea in the line segmentation corresponds to a single complete idea or a single block of information whether it corresponds to a word, a clause, or a phrase. The difference is further narrowed down when lines are combined into episodes. The number of episodes varied from 65 for expert 2 to 35 for expert 5. Of these episodes, the one deemed relevant by each expert is comparable. The final number of categorized episodes varied from 15 for expert 3, who had the most words in the transcript, to expert 5 who had selected 30 episodes. The amount of data reduction through the analysis was the lowest for expert 5's performance, which resulted in the highest number of categorized evidence, even though he had the lowest number of episodes in the analysis.

Comparing Participants' Categorization of Key Elements

In this section I look at the analysis pertaining to how experts categorized the episodes in their reasoning process. Episodes are summaries of events corresponding to a section of the verbal protocol and together they represent the sequence of experts' performance as represented by the visual representation. For the categorization task, participants had to select elements from their visual representation and categorize them as absolutely necessary, necessary, or useful. The goals for this activity were to focus the analysis on selective key decisions and to test whether or not I could improve the convergence between experts' performances by focusing on the important elements leading to successful reasoning performance. I first looked at whether the task was meaningful for experts by looking at how much evidence was categorized and whether they were selective in their choice of evidence. Then the amount of agreement between experts in how they assigned differential weights to episodes was analyzed. Finally, I examined the agreement in the type of evidence selected.

The categorization task captured the most significant episodes of the reasoning process. Experts categorized from 15 to 30 items as evidence from all the episodes presented in their problem solving representation (see table 12). This number of categorized episodes, which I refer to as evidence, corresponds to slightly more than 50 percent of the entire list of episodes on average. Most experts were selective in their categorization but there is variation between expert 3 who categorized 15 of the 48 episodes and expert 5 who selected 30 of the 35 episodes. I had expected a rate of item categorization of over 60 percent considering that the problem solving process performed by experts would be straightforward and include only important elements. However, most participants were very discriminatory in their selection of meaningful episodes pertaining to their diagnoses. This rate of selection is similar in the two other cases and contributes to the validity of the task given that experts selected items carefully. Expert 5 does not show a similar pattern but this might be due to the fewer ideas expressed.

Table 12

Number and Percentage of Categorized Evidence

	E1	E2	E3	E4	E5
Total evidence categorized	20	18	15	24	30
Total number of episodes	44	65	48	38	35
Percentage of categorization	48%	37%	31%	63%	86%

As discussed in the methodology section, weights were associated to each of the three categories: absolutely necessary (+5), necessary (+3), and useful information (+1). I predicted that experts would show stronger agreement on the selection of absolutely necessary episodes. Therefore, the use of differential weightings — giving more weights to most important elements — could improve the agreement among experts. However,

table 13 indicates that experts did not use the weights in a similar way. Experts 1 and 5 categorized only three episodes as being absolutely necessary while considering 8 and 10 episodes as useful. However, the three other experts selected a core number of items as absolutely necessary, and a few as useful information.

An important factor, upon detailed examination, was to understand which evidence was considered absolutely necessary, and to examine why there was not one episode that more than two experts agreed upon. As shown in the Appendix C, there was no convergence on any absolutely necessary episodes selected by the five experts. Contrary to our hypothesis, the differential categorization of episodes emphasizes rather than reduces the variability between experts. The pattern shown in table 13 combined with the list of evidence from the appendix C demonstrate the lack of consensus on how the weights were used for case 1 and given that there is no consensus on the categorization of evidence either for the other cases, I do not discuss this aspect of the result in the two following cases.

Table 13

Pattern of Differential Categorization of Evidence

	E1	E2	E3	E4	E5
Absolutely necessary (+5)	3	12	8	13	3
Necessary (+3)	7	3	4	6	19
Useful information (+1)	10	3	2	5	8

Given the lack of consensus between experts on how they assigned differential weights to the importance of the elements they categorized, I decided to focus our attention on the number of similar elements selected by our five experts. Evidence was identified as common if three of the experts selected it, regardless of the categories. Table 14 summarizes the number of common evidence items between experts (see

details in Appendix D) where the level of agreement between our five experts increased to almost 60 percent. The range varied: expert 3's list had the highest number of evidence in common with all the other experts (87 percent), whereas expert 5 had the least in common (43 percent) with other experts' categorized evidence.

Table 14

Common Categorized Evidence

	E1	E2	E3	E4	E5
Common evidence	9	12	14	13	13
Total categorized evidence	20	18	16	24	30
Percentage of common evidence	45%	67%	87%	54%	43%

Synthesis

Overall, experts showed a moderate level of similarity for case 1: all but one expert submitted the expected final diagnosis. They submitted a similar amount of evidence supporting their diagnosis, however the level of agreement on the nature of the evidence was lower than 50 percent (42 percent). Process measures show a similar variance in the range of confidence for experts who did register a variation of confidence. It also shows similarities in the sequence of hypotheses noted when more than one hypothesis was considered. The descriptive process measures from the verbal protocol show that experts do not take the same amount of time and do not produce a similar amount of utterances for solving this case. The analysis and synthesis of the protocol into episodes enables experts to select a similar number and types of episodes, which are the important processes leading to a successful diagnosis for this case. Episodes are summaries of a set of ideas representing meaningful steps contributing to the resolution. The categorization of evidence for this case results in a consensus level

around 60 percent, which is higher than the consensus on the outcome measures at 42 percent.

Visual Representations of the Problem Solving Process

The previous section on outcome and process measures contributed to the design of the individual and merged visual representations. As a result, these visual representations showed the contextual nature of the performances and enabled an interpretation of both convergence and divergence for this case. To better understand how competent physicians explore the problem space of an ill-defined problem, I used the visual representations to document the sequential episodes of their problem resolutions.

I first present an example of the individual summaries in some detail before moving to the presentation of the merged representation that highlights the similarities and differences across the individual performances. A paper copy of each individual representation is available in Appendix E and two copies of the merged representation are shown with two levels of details in the Appendix F. Merged representations have three levels where the first level shows the common evidence in bold. Level 2 illustrates the details of each individual expert pertaining to the common evidence. Level three corresponds to the original transcript included in each episode and unfortunately cannot be included using paper format since it would be illegible. However, the description of some sections along with the overview in appendices gives sufficient detail to understand the nature and purpose of these blueprints.

Individual Visual Representations

I begin with a detailed narration of case one with expert 3's performance as it represents the "most typical path" given the other performances. It is considered typical because this expert has the most evidence in common with other experts, as seen in

table 14. This narrative should be read alongside the visual representation in figure 7, since it highlights what is seen in the image as well as the content corresponding with each episode of the representation. The flat representation presents enough information to show how the analysis contributes to the synthesis of the problem solving path, but one needs to see the multiple layers that can be viewed by querying inside the episode to see the details of the action, time, and other aspects embedded inside the episodes. Following this detailed narration of a section of expert 3's performance, I show how the common evidence selected leads to the successful resolution of diagnosing the patient's problem. Following the narrative below I provide a review of expert convergence along with summary section of the merged representation of all experts for this case

Detailed Narrative of the Problem Section of Expert 3

It is important to note here that the initial seven episodes did not contain any actions. The expert read the case, and commented on what she was reading, and interpreted the patient case prior to taking any actions. This phenomenon was common to all experts for case 1. It appears that repetition in the representation and discourse is related to the experts developing an understanding of the task.

Looking at the initial problem section of expert 3's visual representation in figure 7, I see that the session started with instructions. Then, in the first two episodes, she talked about the woman's age and that she complained of feeling unwell. In the third episode, she commented on the vagueness of the complaint, but also about the need to pay close attention to symptoms and the importance of describing them succinctly. In episode four, expert 3 stated the importance of the following symptoms: frequent headaches, feeling anxious, palpitations, and profuse sweating and flushing. In the next episode, she linked the symptoms and comments by stating that these symptoms were episodic and had recently increased in frequency. In the subsequent episode, she

highlighted other important details of weight loss and dizziness as evidence into her symptomatology; in episode eight, she related the collected symptoms to the new blood medication that Lydia had begun taking, questioning if that the medication was responsible for the symptoms.

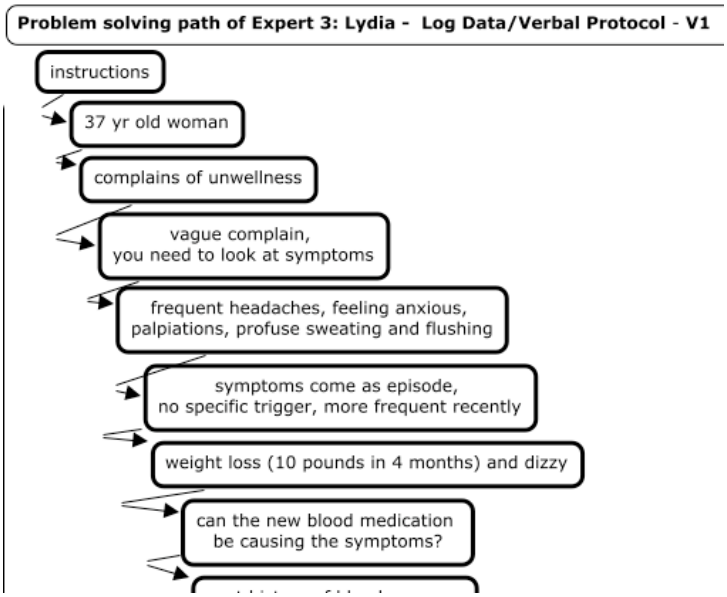


Figure 7. Section of Expert 3 Visual Representation of Case 1

Common Evidence in the Narrative of Expert 3's Visual Representation

In the previous section I examined the underlying depth of details contained in the sequence of episodes and how they pertained to the narrative sequence of the participant's case presentation performance. This detailed example demonstrated how these visual representations were used to capture key aspects of the experts' problem solving process along with the evidence they considered relevant during problem solving. Experts could then reflect on representations of their own knowledge and performance. The following overview provides an understanding of sequence and context of the evidence analyzed by expert 3 on case one. Evidence presented corresponds to episodes that experts have categorized regardless of the degree of importance they associated to the episode. The description of expert 3's key elements

relates to figure 8 below, but a full representation of each individual expert representation is available in Appendix E.

The first evidence selected by expert 3 was the age of the patient. The next evidence identified was the association of the new blood pressure medication as a potential explanation of the symptoms followed by symptoms experienced in episodes; the linking of the different symptoms of palpitations, profuse sweating and flushing together; and finally, the 10 pound weight-loss in four months. Expert 3 then set her hypothesis of hyperthyroid and reviewed an alternative endocrinology diagnosis prior to progressing to the patient chart where she could order diagnostic tests. She commented that the patient is hypertensive and then commenced testing, first ordering a Random Blood Glucose, then a TSH, a T4, and a T3. The negative results for all three led her to change her main hypothesis to pheocromocytoma. She tested for pheocromocytoma by ordering the Urinary Catecholamines / Total (Epinephrine + Norepinephrine) test, and then stated she would do a CBC and basic tests to ensure that the patient is not anemic. Expert 3 then ordered a Urinary Catecholamines / Free before going into a review of the evidence supporting the hypothesis and planning a CT of the abdomen. Finally, she submitted the hypothesis of pheocromocytoma with 65 percent level of confidence.

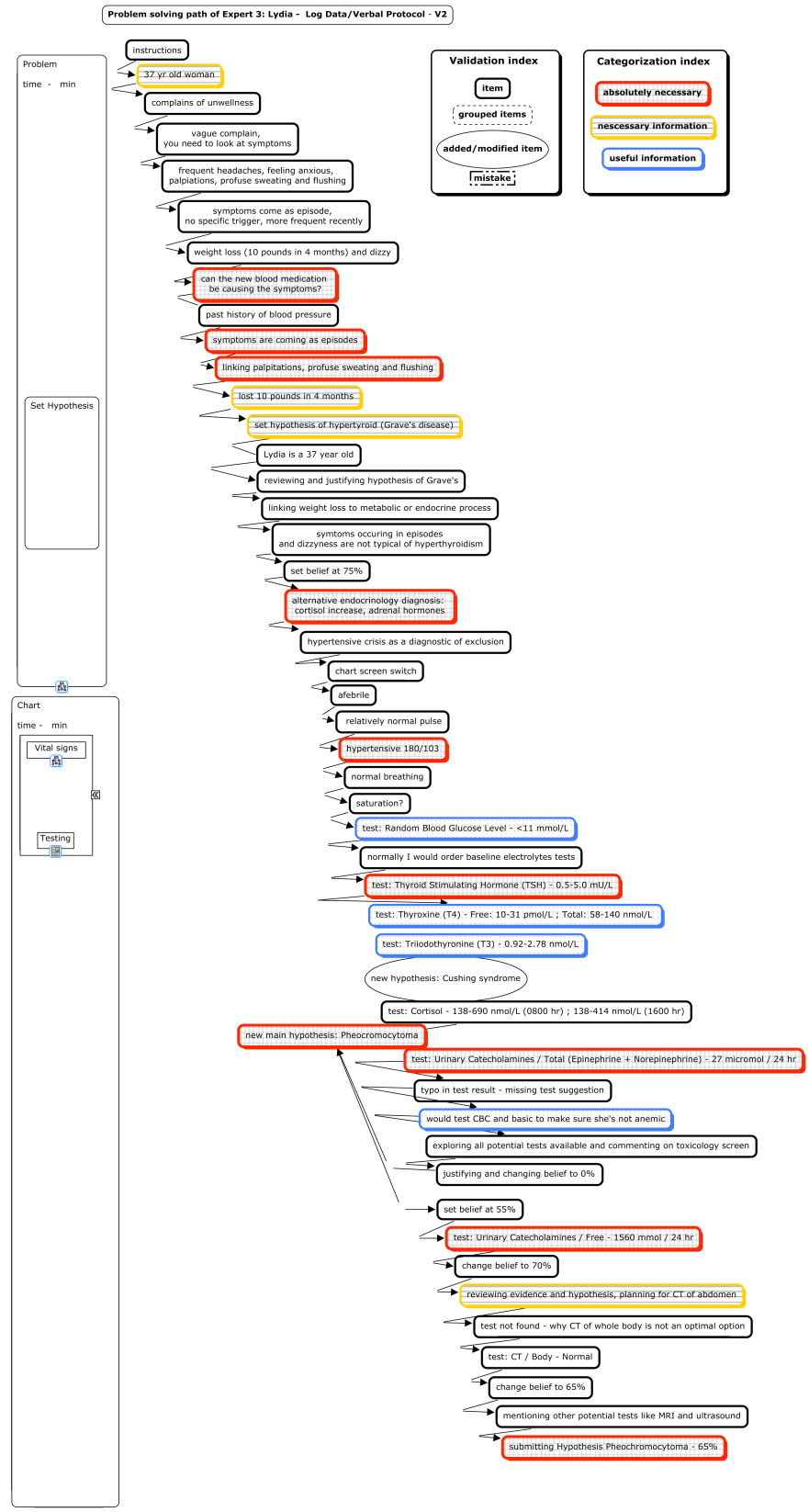


Figure 8. Overview of individual categorized visual representations of expert 3

Merged Representation

The previous description includes most of the common evidence extracted from the comparison of the five individual problem solving representations. Given that Expert 3 identified 11 of the 12 common evidence represented in the merged representation, only one was missing from the narrative. She missed the common evidence related to test baseline, blood glucose, electrolytes and biochem tests (see merged case representation for more details). Figure 9 illustrates the first section of the merged representation with the first four common pieces of evidence showing level one of the representation. Below, it shows the same section at level 2 where the common evidence statements are shown in open position in figure 10 and where the common evidence contains the original individual experts' evidence with their transcript. Figure 9 displays information at level one which is the macro structure of the problem solving process. I only discuss this brief section in detail here, but the entire model is available in both level one and two versions in Appendix F.

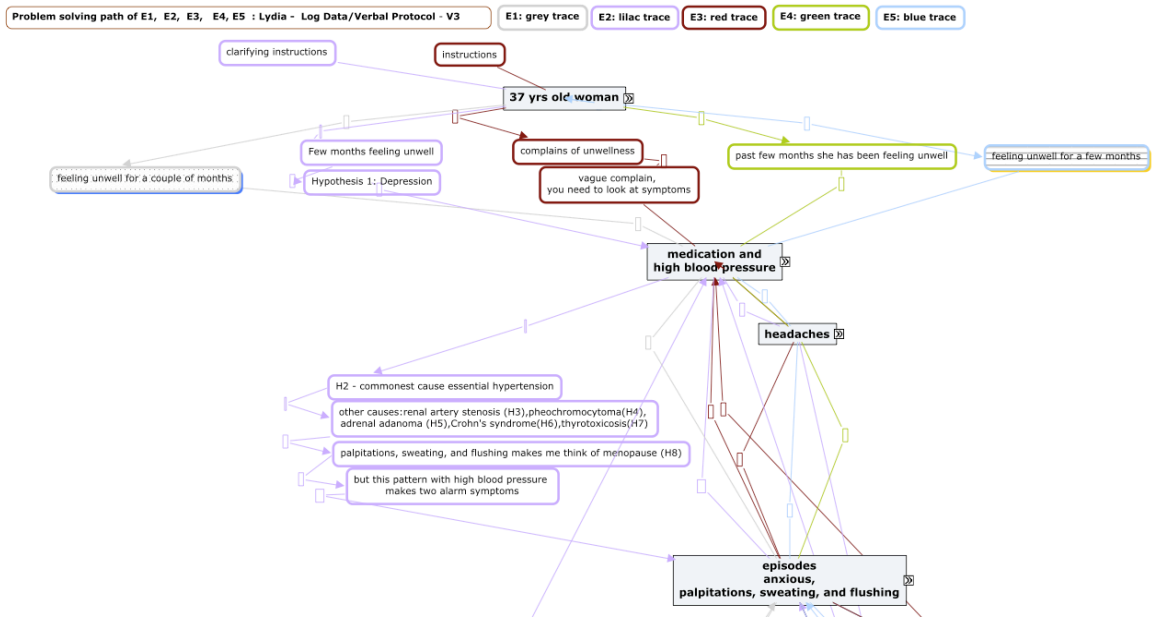


Figure 9. Section of merged visual representation for case 1 at level 1

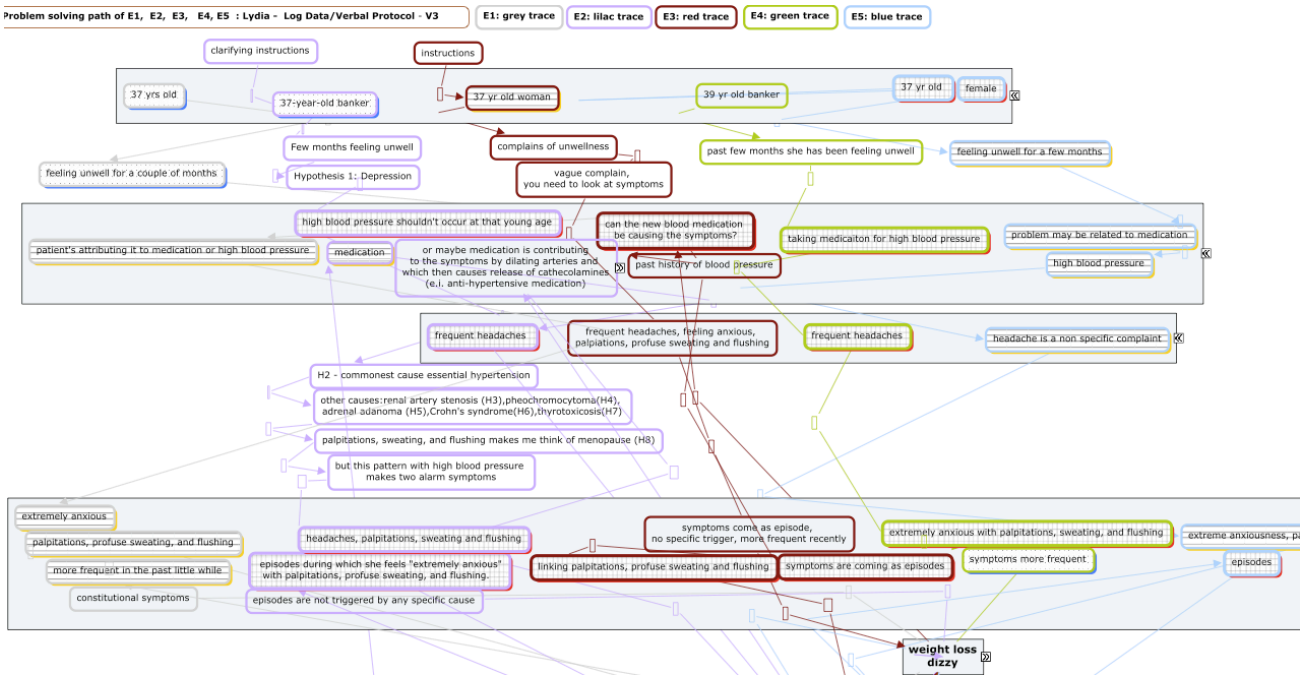


Figure 10. Section of merged visual representation for case 1 at level 2

In this section all experts mentioned the age of the patient but only four of them categorized this episode as evidence. Within these episodes, expert 2 and expert 5 commented on the impact of the age for the evaluation of the current problem. Then, even if all of the experts mentioned the complaint of feeling unwell, only 2 experts categorized this episode as evidence, so it was not identified as common evidence. The next common evidence is the one related to high blood pressure and the medication the patient is taking. All experts mentioned these two elements in one or two separate episodes and categorized them as evidence. High blood pressure and medication count as an item of evidence since the high blood pressure and medication were not clearly separated by three of the five experts. In the representation, I also added non-categorized episodes that relate to the topic. While a non-categorized episode does not count in the merging of nodes, it generally adds to the details of the narrative around the reasoning process.

The next important symptom mentioned by four experts was the frequent headaches, but only three categorized it as evidence. Expert 2 had 500 lines

representing different ideas and listed all the possible diagnoses that could be related to these symptoms mentioned up to this point. His verbalization about the problem then reconnected with the three other experts and the next common evidence about the episodic nature of the symptoms. Four experts mentioned the fact that palpitations, sweating, flushing and anxiety qualified as evidence; three of them linked anxiety and the symptoms together in the same episode; and one considered them as separate entities. The episodic nature of these symptoms was combined to the symptoms themselves by expert 2 but considered separate evidence by experts 3 and 5. However, in an episode which was not categorized as evidence, expert 3 did link the symptoms and their episodic nature together, but I listed them as two separate lines in this common evidence to indicate that they are not linked together by all experts.

This brief section aims to describe where experts converge on the common evidence related to the pheocromocytoma case, as illustrated in the merged representation. Convergence in the reasoning performance becomes clear when I consider the context in which the evidence is selected and categorized. Looking at the entire merged representation, I can narrow down the problem solving process to twelve common pieces of evidence, as shown in the level 1 version of the representation in Appendix F. Only 30 of the 114 categorized episodes are not included as common evidence, and a third of these 30 categorized episodes are associated with expert 5 who did not submit the same diagnosis -- an exception worth noting. Given that expert 5 had a different final diagnosis of panic attack, I was expecting a greater difference in the reasoning process, but this was not the case as his reasoning had a lot in common with the four other experts. In the case of Lydia, this merged representation, together with the individual representations, were effective tools for illustrating the similarities and differences among the performance of experts.

Case 2

Brief Background Information: Problem and Problem Solver

Case Description

Stephanie is a case about an active 16-year-old female who had been experiencing extreme fatigue, vision problems, excessive thirst and frequent urination. Her mother brought her to the emergency room because she was experiencing nausea, vomiting, and abdominal pain. This was a typical case of diabetes Mellitus type I that showed several typical symptoms of diabetes such as polyuria, polydipsia, weight loss, and blurred vision; however, she also had diabetic ketoacidosis that can sometimes be mistaken for pancreatitis or an acute abdomen. The critical diagnostic tests for this condition are blood glucose level and serum ketones; additionally, it was indicated to test her electrolytes as there are many metabolic complications that can accompany diabetic ketoacidosis (DKA).

Experience of Participant and Case Difficulty Rating

On the post-questionnaire following the completion of this diabetes case, expert 1 wrote that he remembered having seen one or two cases about three or four years ago. Expert 2, 3 and 4 indicated that they see these diabetes cases on a regular basis and as mentioned in the analysis, due to the technical problems, explained in the method section, data from expert 5 were not included in this case. With respect to the case rating section, participants were asked to give a rating on a scale of difficulty of 1 to 4: experts 2, 3, and 4 rated this case as too easy (1) while expert 1 rated this case as being a good revision (2).

Convergence on Protocol Measures

Solution Outcome Measures

Table 15 represents the summary of the outcome measures related to the final diagnosis submitted by experts at the end of the case resolution task. Table 15 demonstrates that all experts submitted a final hypothesis of Diabetes Mellitus (type I). The confidence level about their final hypothesis is around 100 percent for everyone.

Table 15

Overview of Final Hypothesis, Confidence Level and Final Evidence

	E1	E2	E3	E4
Final hypothesis	Diabetes Mellitus (type I)	Diabetes Mellitus (type I)	Diabetes Mellitus (type I)	Diabetes Mellitus (type I)
Confidence level (%)	100	100	100	99
List of prioritized evidence in final argument	1 Random Blood Glucose Level - 18.2 mmol/L 2. pH - 7 3. Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L 4. Serum Ketones - Present 5. Serum Electrolytes / Osmolilty - 320 mmol/Kg H20 6. urinate more frequently, 7. nausea, vomiting, and abdominal pain 8. feeling excessively thirsty 9. nauseated 10. difficulty seeing 11. 6 pound weight loss 12. extreme fatigue 13. fatigue has even progressed to 14. HbA1C - 12.5% 15. Serum Electrolytes / Sodium (Na) - 130 mEq/L 16. Serum Electrolytes / Potassium (K) - 5.8 mEq/L 17. Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21	1. Random Bood Glucose Level - 18.2 mmol/L 2. Serum Ketones - Present 3. having to urinate more frequently 4. excessively thirsty 5. difficulty seeing 6. 6 pound weight loss in the past month 7. Today Stephanie is experiencing nausea, vomiting, and abdominal pain 8. Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21 fatigue	1. Random Blood Glucose Level - 18.2 mmol/L 2. nausea, vomiting, and abdominal pain 3. urinate more frequently 4. excessively thirsty 5. difficulty seeing 6. nauseated	1. Random Blood Glucose Level - 18.2 mmol/L 2. pH - 7 3. Serum Ketones - Present 4. thirsty 5. urinate more frequently 6. difficulty seeing 7. 6 pound weight loss 8. pCO2 - 24 mmHg 9. Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L 10. Serum Electrolytes / Potassium (K) - 5.8 mEq/L 11. Serum Electrolytes / Phosphate - 1.8 mg/dL 12. nauseated 13. HbA1C - 12.5% 14. Today Stephanie is experiencing nausea, vomiting, and abdominal pain 15. WBC (total) - 12 x 10E9 / L

The total number of final evidence participants submitted to support their final diagnosis varied from six for expert 3 to seventeen for expert 1. The list of prioritized

evidence showed some level of agreement and the actual prioritization of evidence showed better agreement than in the previous case. One evidence selection was at the top of the priority on each list and a second one was within the top four of each participant's priority list. Looking at the substance of the final evidence, as shown in Table 16, all four experts shared five items of supporting evidence. If we take the same criteria of three experts out of four, which corresponds to the above fifty percent criteria set for the previous case, we have up to seven shared items of evidence for a consensus rate of 55 percent.

Table 16

Consensus Rate of Final Evidence Submitted

	E1	E2	E3	E4	Total	Consensus Rate
Common evidence (4/4)	5	5	5	5	20	43%
Common evidence (3/4)	7	7	5	7	26	55%
Final evidence	17	9	6	15	47	

Solution Process Measures

To gain insight into common elements of the process, I first identified the measures based on the actions experts took while resolving the case. Table 17 shows the number of tests ordered throughout the resolution. It also includes the number and the list of hypotheses selected throughout the resolution process and the final confidence level along with the variation of this confidence through the task.

One hypothesis was recorded for all experts with a final confidence of 99 or 100 percent. The range of confidence varied between 79-100 percent. Expert 4 began with 79 percent, expert 2 and 3 began at around 80 percent and expert 1 at 100 percent certain throughout the case. The number of diagnostic tests contrasted with the confidence level where expert 1, who was 100 percent certain, ordered twice as many tests as all the other experts. For this case there was a high level of agreement regarding which diagnostic tests needed to be ordered; using the criteria of above 50

percent agreement, resulted in a 63 percent agreement level, which was twice as high as the consensus regarding the diagnostic tests in case 1.

Table 17

Process Measures from Recorded Actions

	E1	E2	E3	E4
Number of hypotheses	1	1	1	1
List of hypotheses	Diabetes Mellitus (type I)	Diabetes Mellitus (type I)	Diabetes Mellitus (type I)	Diabetes Mellitus (type I)
Final confidence level (%)	100	100	100	99
Range of confidence level	100	89 - 100	90 – 100	79 - 99
Number of test ordered	24	10	12	12
Common tests	10	7	9	8

Additional process data from the transcription and analysis of the protocol are presented in table 18. As seen in the previous case, the table presents the time taken by each participant to submit their final diagnosis for the case along with the total number of words in the protocols, the estimation of the number of words per minute, the number of lines in each idea unit that was segmented, and the number of episodes created for each resolution. The number of episodes that each expert categorized as evidence is provided to further information about important episodes in relationship to the entire protocol. Descriptive data pertaining to time to completion, number of words, and rate of speech were segmented into lines and reduced into episodes that were later categorized.

Table 18

Process Measures from Protocols

	E1	E2	E3	E4
Time (min)	10:43	27:05	16:03	15:07
Number of words	1638	5112	2163	1679
Rate of speech (word per min)	157	189	135	111
Number of lines	214	464	176	85
Number of episodes	56	67	46	38
Categorized episodes	25	27	25	30

On average, participants took 17 minutes to complete case 2. Time to complete the task may reflect the amount of variability in the quantity of verbalization by experts. Again, for this case, expert 1 took significantly less time than the others to complete the task; expert 2, on the other end of the spectrum, took 27 minutes to complete the task — almost twice as much time as the others. The total number of words used by the participants gave us an overview of the length of the transcript citing a substantial difference between expert 2 and the other three experts. The transcript of expert 2 had over 5000 words where the other three transcripts had between 1600 and 2000 words. The number of lines showed again that expert 2 expressed more ideas for this case than did his colleagues: he had 464 lines, whereas expert 4 had only 85 lines. The concept of “idea” in the line segmentation corresponds to a single complete idea, or a single block of information whether it corresponds to a word, a clause, or a phrase. The difference is further narrowed down when lines are combined into episodes. The number of episodes varied between 38 for expert 4 to 67 for expert 2 and interestingly enough, the number of episodes categorized as evidence were very similar, varying between 25 for experts 1 and 3 to 27 for expert 2, and 30 for expert 4.

Comparing Categorization of Key Elements

The categorization task aimed at capturing the important episodes of the reasoning process. As shown in table 19, experts categorized between 25 to 30 items of evidence from their problem solving representations. The number of categorized episodes corresponded to slightly more than 54 percent of the entire list of episodes on average. Most experts were selective in their categorization, but expert 4 categorized more episodes as evidence, selecting 30 of the 38 episodes.

Table 19

Number and Percentage of Categorized Evidence

	E1	E2	E3	E4
Total evidence categorized	25	27	25	30
Total number of episodes	56	67	46	38
Percentage of categorization	45%	40%	54%	79%

Given the lack of consensus on the differential categorization of elements overall, I focused the analysis on the number of similar elements selected by our five experts. When looking at all the evidence selected, regardless of the categories, I was able to find strong convergence among all the categorized evidence. Evidence was identified as common if three of the experts had selected it. Table 20 summarizes the common evidence found for each expert (see details in Appendix G) and the level of agreement between our five experts, at 87 percent. The range varied with expert 3's list, which again had the most evidence in common with all the other experts to expert 4 having only 80 percent of all the selected evidence agreeing at some level with other experts' categorized evidence.

Table 20

Common Categorized Evidence

	E1	E2	E3	E4
Common evidence	22	23	24	24
Total categorized evidence	25	27	25	30
Percentage of common evidence	88%	85%	96%	80%

Synthesis

Overall, experts showed a relatively high level of similarity in solving case 2. All four experts submitted the same diagnosis of diabetes mellitus type I with almost a 100 percent confidence level. While there is a greater amount of evidence submitted to support the diagnosis for this case, experts all agreed that the most important evidence was the random blood glucose diagnostic test. Overall the agreement of evidence to support the answer was higher than for the previous case at 55 percent agreement when applying a more strict three out of four criteria. The degree of agreement for evidence supporting the final diagnosis of 55 percent for case 2 was higher than 42 percent with a three out of five criteria for case 1.

In terms of process measures, all experts considered only one hypothesis and their level of confidence varied from 79 to 100 percent. The number of diagnostic tests was similar — around 12, except for expert 1 who ordered 24 tests. The diagnostic tests ordered were fairly consistent across subjects: three of the four experts had seven to ten tests in common. Despite these similarities the descriptive process measures from the verbal protocol show that experts did not take the same amount of time to solve the problem. This is especially true for expert 2, who took almost twice as long as others to complete case 2. More time to completion resulted in a different amount of utterances for this case; however, the analysis and synthesis of the protocol into episodes enabled

comparisons between experts on the nature of the important processes leading to a successful diagnosis for this case.

Interestingly enough expert 2 did not end up with the highest number of categorized episodes. Out of the 67 episodes, he only selected 27 as being of some importance for the resolution of the case. The number of categorized episodes varied from 25 for expert 1 to 30 for expert 4. The categorization of evidence for this case resulted in a consensus level at approximately 87 percent which is higher than the consensus on the outcome measures of 55 percent of the supporting evidence.

Visual Representations of the Problem Solving Process

The previous section on outcome and process measures contributed to the design of the individual and merged visual representations. As a result, these visual representations show the contextual nature of the performances and enabled an interpretation of both convergence and divergence for this case. For case 1, a section of the individual representation of the case is presented to facilitate an in-depth examination of details in each representation but I found it unnecessary to repeat the exercise for case 2 as the result of the analysis are not about particular narrative in transcript but about how these narratives converge. I use the common evidence embedded in the performance of expert 4 to situate a summary of the common evidence identified for this diabetes case. Both expert 3 and expert 4 identified the same amount of common evidence corresponding to common topics. Here expert 4's performance and representation is used to illustrate the convergence given that expert 3's data were used for the previous case. A paper copy of each individual representation is available in Appendix H and two copies of the merged representation showing details at two levels are available in the Appendix I.

Individual Visual Representations

Common Evidence as Described by Expert 4's Performance

The following section provides a description of the sequence and context in which expert 4 demonstrated consensus with other experts on the evidence selected for this diabetes mellitus I case. The description of expert 4's common elements relates to figure 11 below, but a full representation of each individual expert's representation is available in Appendix H.

The series of common episodes selected by expert 4 for this case started with the fact that the patient was a 16-year-old teenager with symptoms of fatigue, frequent urination, and excessive thirst. These facts led expert 4 to her first hypothesis: the onset of juvenile diabetes. She then reviewed the evidence and stated that anemia could be an alternative hypothesis. The next evidence related were the expert's concerns about the patient's nausea and difficulty seeing. The cause was explained as the glucose changes affecting the sorbitol level in the lens, causing the lens to swell and shrink at a strange rate affecting vision. She then linked the nausea, vomiting, and abdominal pain to type I diabetes with ketoacidosis and selected diabetes mellitus type I as the hypothesis before repeating and selecting the key evidence of extreme fatigue, polyurea, polydipsia, nausea, difficulty seeing, and the six pound weight loss. Today's symptoms of nausea, vomiting and abdominal pain were also selected and repeated prior to progressing to the chart section. Expert 4 repeated why the tachycardic, hypotensive and tachypeic situation of the patient did not surprise her and then ordered a diagnostic test of random blood glucose. The next diagnostic tests she ordered were serum electrolyte / bicarbonate and serum ketones followed by a series of diagnostic tests in the same category: potassium, creatinine and blood urea nitrogen. She then ordered a HbA1C, a ph 7, a PCO2 and mentioned that she was looking for infections that could be precipitating the diabetes. Expert 4 then performed another round of diagnostic tests —

WBC, phosphate, and magnesium — before summarizing and reviewing the clinical picture that supported the final diagnosis that she submitted with 99 percent confidence.

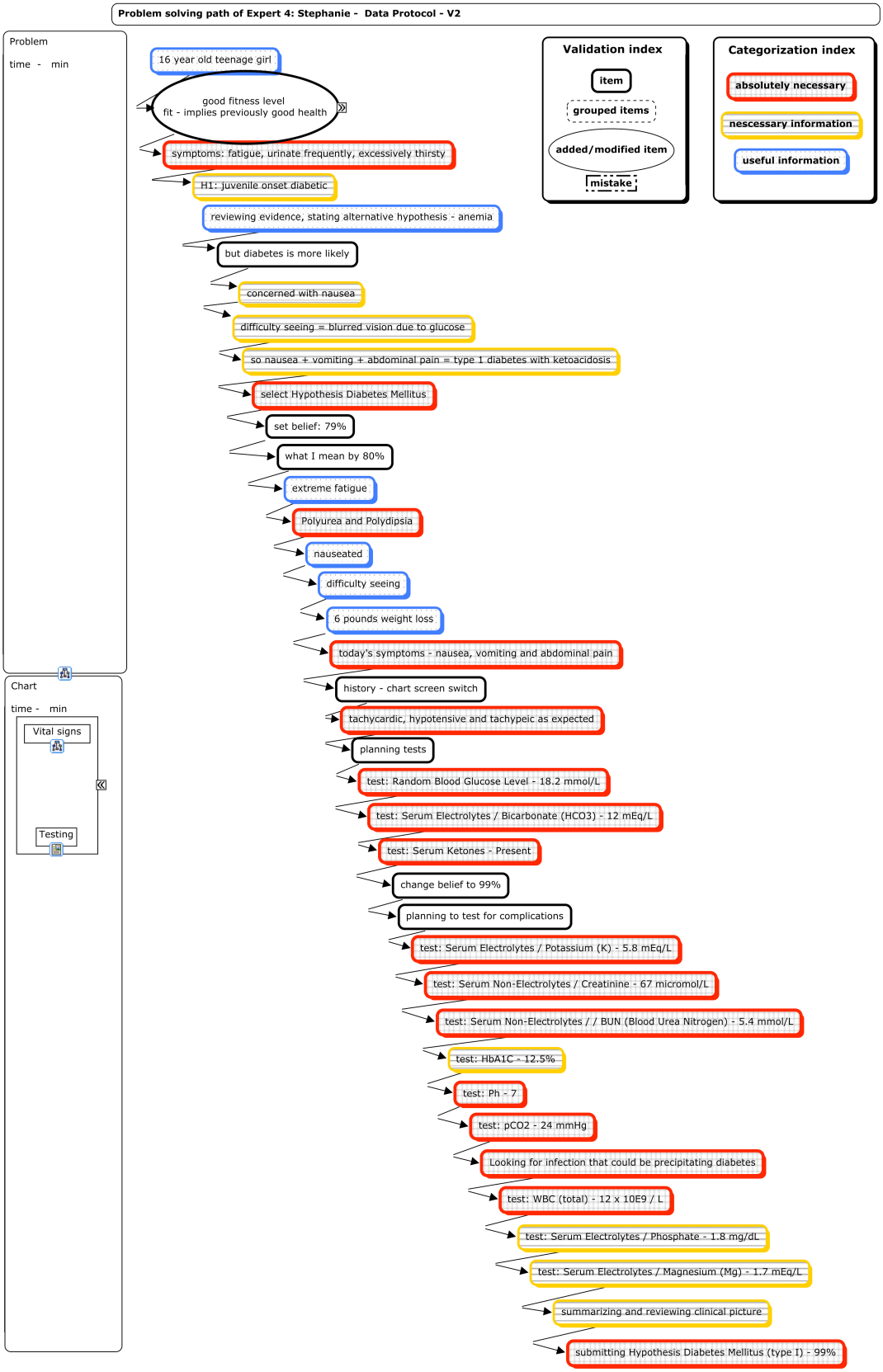

Problem solving path of Expert 4: Stephanie - Data Protocol - V2
Problem
time - min
Chart
time - min
Vital signs
Testing
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
16 year old teenage girl
good fitness level
fit - implies previously good health
symptoms: fatigue, urinate frequently, excessively thirsty
H1: juvenile onset diabetic
reviewing evidence, stating alternative hypothesis - anemia
but diabetes is more likely
concerned with nausea
difficulty seeing = blurred vision due to glucose
so nausea + vomiting + abdominal pain = type 1 diabetes with ketoacidosis
select Hypothesis Diabetes Mellitus
set belief: 79%
what I mean by 80%
extreme fatigue
Polyurea and Polydipsia
nauseated
difficulty seeing
6 pounds weight loss
today's symptoms - nausea, vomiting and abdominal pain
history - chart screen switch
tachycardic, hypotensive and tachypeic as expected
planning tests
test: Random Blood Glucose Level - 18.2 mmol/L
test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L
test: Serum Ketones - Present
change belief to 99%
planning to test for complications
test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L
test: Serum Non-Electrolytes / Creatinine - 67 micromol/L
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 5.4 mmol/L
test: HbA1C - 12.5%
test: Ph - 7
test: pCO2 - 24 mmHg
Looking for infection that could be precipitating diabetes
test: WBC (total) - 12 x 10E9 / L
test: Serum Electrolytes / Phosphate - 1.8 mg/dL
test: Serum Electrolytes / Magnesium (Mg) - 1.7 mEq/L
summarizing and reviewing clinical picture
submitting Hypothesis Diabetes Mellitus (type I) - 99%

Figure 11. Overview of individual categorized visual representations of expert 4

Merged Representation

The previous description represents common evidence for this diabetes case in the context of expert 4's performance. Expert 4 identified 24 items of evidence that are contained in 13 of the 14 common evidence represented in the merged representation. Common evidence often contained more than one individual evidence; they were often grouped when experts discussed them together. Figure 12 illustrates the section prior to the hypothesis submission where expert 4 misses the common evidence regarding the x-ray diagnostic testing. Figure 13 illustrates the same section when the common evidence is shown at level two that corresponds to an open position of the common evidence of level 1. Only a small section is discussed here in detail, but the entire model is available at level 1 and 2 versions in Appendix I.

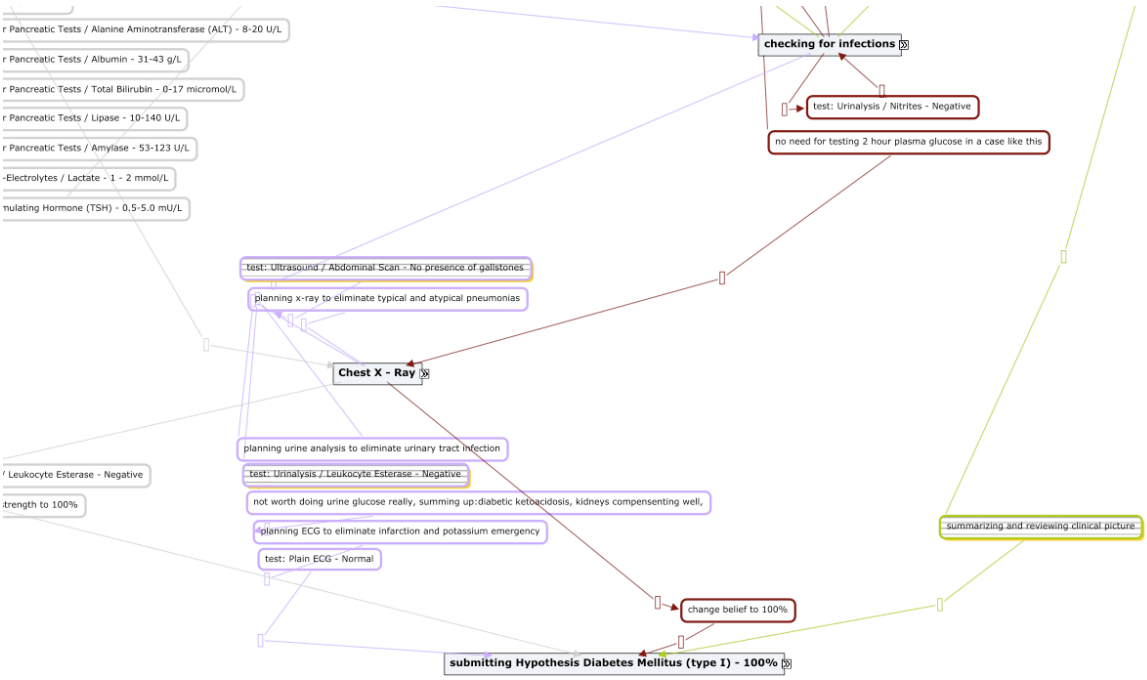


Figure 12. Section of merged visual representation for case 2 at level 1

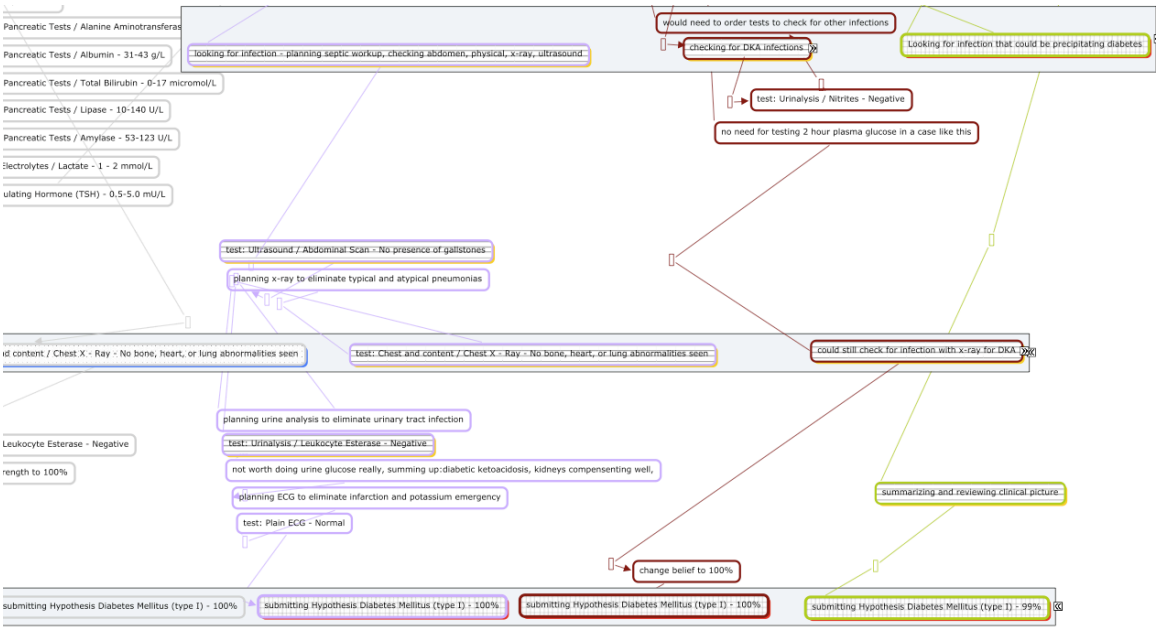


Figure 13. Section of merged visual representation for case 2 at level 2

This section illustrated by figure 12 and 13 represents the last diagnostic tests and actions for this case of diabetes where experts were trying to pinpoint possible infections potentially causing the onset of the diabetes. Expert 1's path is partially visible on the left side of the representation and expert 4's is the last one on the right. At the top right section of figure13 there is one common evidence selected by expert 4 that refers to the summarizing and reviewing of the clinical picture. Expert 4 does not include chest X-ray as part of her evidence while other experts do have this item in common (see middle of page). In the level 2 version of this section in figure 13 expert 1 and 2 ordered the diagnostic test of chest x-ray whereas expert 3 stated that one could check for infection with a diagnostic test of x-ray but did not order it before submitting her final diagnosis. Expert 1 and 2 do order other tests like urinalysis, or ECG before submitting their final diagnoses.

This brief section above aimed at giving some account of a marginally convergent section of the merged representation for the diabetes case. The context in which this variability occurred can easily be interpreted when reading the content of

previous episodes where experts explained that they could not identify the source of the deterioration of this patient even though they considered most possibilities. In this case, 14 common items of evidence were contained in the 107 categorized episodes, and only 12 of these categorized episodes were outside the common evidence.

Case 3

Brief Background Information: Problem and Problem Solver

Case Description

Linda is a case about a 34-year old woman who was having anxiety attacks, felt nervous and experienced episodes of excessive sweating, hand tremors, and sensations that her heart was racing. This is typical of Grave's disease, but her constellation of symptoms — weight loss, anxiety, tremor, palpitations, and sweating – could have been explained by a psychological diagnostic. In such a case it is recommended to rule out physiologic causes before making a psychological diagnostic. The critical tests are TSH, T4, and T3 to confirm Hyperthyroid (Grave's disease).

Experience of Participant and Case Difficulty Rating

On the post-questionnaire following the completion of the case, expert 1 wrote that he had seen more than four similar cases in the last five years, experts 2 and 4 reported seeing these types of cases on a regular basis, expert 3 had seen three or four cases in the last three years, and expert 5 had seen more than four cases in the last ten years. In the case rating section, participants were asked to give a rating on a scale of difficulty of 1 to 4; experts 1, 2, 3 and 4 all rated the case as being too easy (1) for physicians while expert 5 rated this case as a good revision (2).

Convergence on Protocol Measures

Solution Outcome Measures

Table 21 presents the summary of the outcome measures related to the final diagnosis submitted by experts at the end of this case resolution. This table shows that all experts submitted a final hypothesis of Hyperthyroid (Grave's disease). The

confidence level about their final hypothesis varied between 90 percent for expert 5, 94 percent for expert 4 to 100 percent for experts 1, 2 and 3.

Table 21

Overview of Final Hypothesis, Confidence Level and Final Evidence

	E1	E2	E3	E4	E5
Final hypothesis	Hyperthyroid (Grave's disease)	Hyperthyroid (Grave's disease)	Hyperthyroid (Grave's disease)	Hyperthyroid (Grave's disease)	Hyperthyroid (Grave's disease)
Confidence level (%)	100	100	100	94	90
List of prioritized evidences in final argument	1. excessive sweating, hand tremors, and a sensation "as if her heart was racing." 2. "anxiety attacks" 3. Thyroid Stimulating Hormone (TSH) - 0.2 mU/L 4. 12 pound weight loss in two months; 5. Thyroxine (T4) - 89 pmol/L (free) 6. Triiodothyronine (T3) - 4.2 nmol/L 7. Radionuclide Scan of Thyroid Gland - diffuse high uptake 8. Valium	1. Thyroxine (T4) - 89 pmol/L (free) 2. Thyroid Stimulating Hormone (TSH) - 0.2 mU/L 3. Radionuclide Scan of Thyroid Gland - diffuse high uptake 4. episodes of excessive sweating, hand tremors, and a sensation "as if her heart was racing."	1. excessive sweating, hand tremors, and a sensation "as if her heart was racing." 2. healthy 34-year-old woman 3. Thyroid Stimulating Hormone (TSH) - 0.2 mU/L 4. Triiodothyronine (T3) - 4.2 nmol/L 5. Thyroxine (T4) - 89 pmol/L (free) 6. Thyroid Stimulating Immunoglobulin Assay - Present 7. "anxiety attacks" during 8. Radionuclide Scan of Thyroid Gland - diffuse high uptake 9. 12 pound weight loss in two months;	1. Thyroxine (T4) - 89 pmol/L (free) 2. Thyroid Stimulating Immunoglobulin Assay - Present 3. heart was racing 4. sweating 5. tremors 6. "anxiety 7. 12-pound weight loss in two months; 8. Triiodothyronine (T3) - 4.2 nmol/L 9. Thyroid Stimulating Hormone (TSH) - 0.2 mU/L	Thyroxine (T4) - 89 pmol/L (free) 2. 34-year-old woman 3. Triiodothyronine (T3) - 4.2 nmol/L 4. Thyroid Stimulating Hormone (TSH) - 0.2 mU/L 5. Plain ECG - Sinus tachycardia 6. very nervous for no particular reason; 7. 12-pound weight loss 8. episodes of excessive sweating, hand tremors, and a sensation "as if her heart was racing." 9. "anxiety attacks" during

The total number of final evidence experts submitted to support their final diagnosis varied from four to nine: this list of prioritized evidence conveyed a sense of agreement. Even though the agreement on the prioritization of evidence is low with only one evidence at the top of the priority, when looking at all the evidence, three of them are the same for all experts for a consensus rate of 40 percent. This rate improves to 78 percent when using the three out of five expert criteria.

Table 22

Consensus Rate of Final Evidence Submitted

	E1	E2	E3	E4	E5	Total	Consensus Rate
Common evidence (5/5)	3	3	3	3	3	15	40%
Common evidence (3/5)	7	4	7	5	6	29	78%
Number of final evidence	8	4	9	7	9	37	

Solution Process Measures

To gain insight into common elements of the process, the measures based on the actions of the expert while they were solving the case are reported. Table 23 shows the number of diagnostic tests ordered throughout the problem solving process as well as the number and the list of hypotheses selected. Additionally, it reports the final confidence level along with the variation of this confidence throughout the problem resolution. With four hypotheses recorded, only expert 5 recorded more than one hypothesis. The list also provides an indication of the sequence in which this expert progressed through these hypotheses, beginning with panic attacks as his initial hypothesis, then selecting arrhythmia and Grave's disease before going back to panic attacks. For his final submission, he re-visited the Grave's disease hypothesis. Given this range of hypotheses, it is surprising that there was no variation in the confidence level for this expert confirming that he was not indicating change in confidence in the hypothesis selected while going through each resolution. The same pattern is observed for the other cases he completed. Expert 5 kept a confidence level of 90 percent which

was the lowest confidence level among the five experts for this case. Yet, that level of confidence is comparable to expert 4 who submitted her final diagnosis with 94 percent level of confidence. The three others submitted their final hypothesis with 100 percent confidence. The number of diagnostic tests ordered is comparable: experts 2 and 3 ordered seven and six tests respectively and the others three ordered four tests. Overall, there seemed to be very little disagreement over the diagnostic tests needed for this case; the consensus rate when using the criterion of three out of five experts was 75 percent.

Table 23

Process Measures from Recorded Actions

	E1	E2	E3	E4	E5
Number of hypotheses	1	1	1	1	4
List of hypotheses	Hyperthy-roid (Grave's disease)	Hyperthy-roid (Grave's disease)	Hyperthy-roid (Grave's disease)	Hyperthy-roid (Grave's disease)	Panic Attack, Arrhythmia, Hyperthyroid (Grave's disease), Pheochromocyto-ma, Panic Attack, Hyperthyroid (Grave's disease)
Final confidence level (%)	100	100	100	94	90
Range of confidence level	95 - 100	52 - 100	70 - 100	80 - 94	90
Number of test ordered	4	7	6	4	4
Common tests	4	4	4	3	3

Additional process data from the transcription and analysis of the protocol are presented in table 24. The table presents the time taken by each participant to submit final diagnosis for the case along with the total of words in the protocols, the estimation of the number of words per minute, the number of lines in each idea unit that was segmented, and the number of episodes created for each resolution. The number of

episodes that each expert categorized as evidence is provided to further information about important episodes in relationship to the entire protocol. Descriptive data pertaining to time to completion, number of words, and rate of speech were segmented into lines and reduced into episodes that were later categorized.

Table 24

Process Measures from Protocols

	E1	E2	E3	E4	E5
Time (min)	3:57	20:25	13:31	4:54	10:59
Number of words	449	3353	1477	672	965
Rate of speech (words per min)	126	166	111	148	91
Number of lines	75	454	176	85	110
Number of episodes	25	46	21	27	24
Categorized episodes	10	21	14	20	21

The mean time taken by participants to complete the case is slightly more than 10 minutes; however, this mean was only representative of expert 3 and 5 given that experts 1 and 4 took approximately 4 minutes and expert 2 took 20 minutes. The number of words was representative of the time taken by each expert where expert 2 had 3353 words and expert 1 had only 449 words. The difference between experts decreased when the transcripts were segmented into distinct expressed ideas, yet the discrepancy was still present in the number of lines for each transcript where expert 2 had 454 lines and expert 1 had 75 lines. The difference was reduced when lines were combined into episodes with the number of episodes varying from 21 for expert 3 to 46 for expert 2. In terms of categorized evidence, a comparable number of episodes were categorized with the result of expert 1 who categorized 10 episodes and expert 2 and 5 who categorized 21.

Comparing Participants' Categorization of Key Elements

The categorization task aimed at capturing the important episodes of the reasoning process. Table 25 demonstrates that experts categorized between 10 and 21 items of evidence from their problem solving representations. On average, experts categorized items was 63 percent of the episodes presented on their visual representation. Experts 4 and 5 show a higher percentage of categorization, with their respective selection of 74 percent and 88 percent.

Table 25

Number and Percentage of Categorized Evidence

	E1	E2	E3	E4	E5
Total evidence categorized	10	21	14	20	21
Total number of episodes	25	46	21	27	24
Percentage of categorization	40%	46%	67%	74%	88%

Table 26 shows that the convergence rate reached 80 percent when all the evidence selected was taken into consideration, regardless of the categories. The range varied with expert 3's list, where categorization led to the identification of 13 items of common evidence from the 14 she identified. With a total of 9, expert 1 had the least common evidence, but almost all of the ten categorized episodes he selected are in common with the other experts. The content and number of common episodes found for each expert is detailed in Appendix J.

Table 26

Common Categorized Evidence

	E1	E2	E3	E4	E5
Common evidence	9	16	13	15	14
Total categorized evidence	10	21	14	20	21
Percentage of common evidence	90%	76%	93%	75%	67%

Synthesis

Overall, experts showed an important level of similarity for case 3: all five experts submitted the same final diagnosis of Hyperthyroid (Grave's disease) with a level of confidence between 90 and 100 percent and evidence supporting the final diagnosis at under ten for all experts. The prioritization showed little agreement but the content of the evidence was similar. The highest agreement rate was reached with 78 percent when using the three out of five criteria and in terms of process measures, all but one expert recorded one hypothesis. Experts showed great variation in their confidence ratings in their hypothesis, ranging from 5 percent confidence for Expert 1 to 40 percent for Expert 2. The number of diagnostic tests ordered is similar with experts 2 and 3ordering seven and six tests and experts 1, 4, and 5 ordering four tests.

The descriptive process measures from the verbal protocol show that experts did not take the same amount of time to solve case 3. Expert 2 took the most time — 20:25 minutes — and expert 1, the least at 3:57 minutes with the average for all five experts at ten minutes. This resulted in different amount of utterances for this case but analysis and synthesis of the protocol into episodes enabled comparisons between experts on the nature of the important processes leading to a successful diagnosis for this case. The number of episodes was still higher for expert 2 who had twice as many as the others at 46; however, the categorization task narrows down the difference. Expert 1 selected ten

episodes whereas expert 2 and 5 selected 21. The categorization of evidence for this case results in a consensus level of 80 percent, which is slightly higher than the consensus on the outcome measures.

Visual Representations of the Problem Solving Process

The previous section on outcome and process measures contributed to the design of the individual and merged visual representations. As a result, these visual representations show the contextual nature of the performances and enabled an interpretation of both convergence and divergence for this case. I use the individual common evidence as embedded in the performance of expert 2 to situate a summary of the common evidence identified for this hyperthyroid case. For case 3, the individual evidence of expert 2 is used to situate the summary of the common evidence for the hyperthyroid case. I chose this verbose expert because he had 21 items of common evidence covering all common topics for this third case and provided an expert voice different from the previous two descriptions. A paper copy of each individual representation is available in Appendix K and two copies of the merged representation showing details at two levels are available in the Appendix L.

Individual Visual Representations

Common Evidence as Described by Expert 2's Performance

The following section provides an understanding of the sequence and context of the common evidence analyzed in the context of expert 2's performance on this hyperthyroid case. The description of expert 2's common elements corresponds with figure 14, but a full representation of this individual's performance is available in Appendix K.

The age of the patient was again the first common evidence: he stated that a 34-year-old woman had been having anxiety attacks for two months. Expert 2 then

considered panic attacks as a diagnosis since it is a common diagnosis but considered it as a diagnosis of exclusion prior to moving on to other diagnoses. He stressed that weight loss is an element of the problem that is contrary to the anxiety or panic disorder hypotheses. He then summed up and linked the evidence of anxiety attacks, episodes, and weight loss to the differentials of pheocromocytoma, drug abuse, alcohol withdrawal, cardiac arrhythmia, and panic disorder. Expert 2 then selected the evidence of the episodes of excessive sweating, hand tremors, and a sensation "as if her heart was racing" as key evidence. This action was followed by a repetition in the representation due to the manual selection of previously discussed items for this case. Expert 2 selected the evidence of 12-pound weight loss and the borrowing of Valium from her mother prior to selecting Grave's disease as his main hypothesis. Once he was in the chart section, he commented on the pulse of 120 and a widening blood pressure before he ordered a plain ECG followed by the TSH and T4. The following selected evidence was the B-HCG diagnostic test, the T3 and radionuclide scan diagnostic tests. He then conducted a diagnostic test of WBC prior to submitting his final hypothesis of Grave's disease with 100 percent confidence.

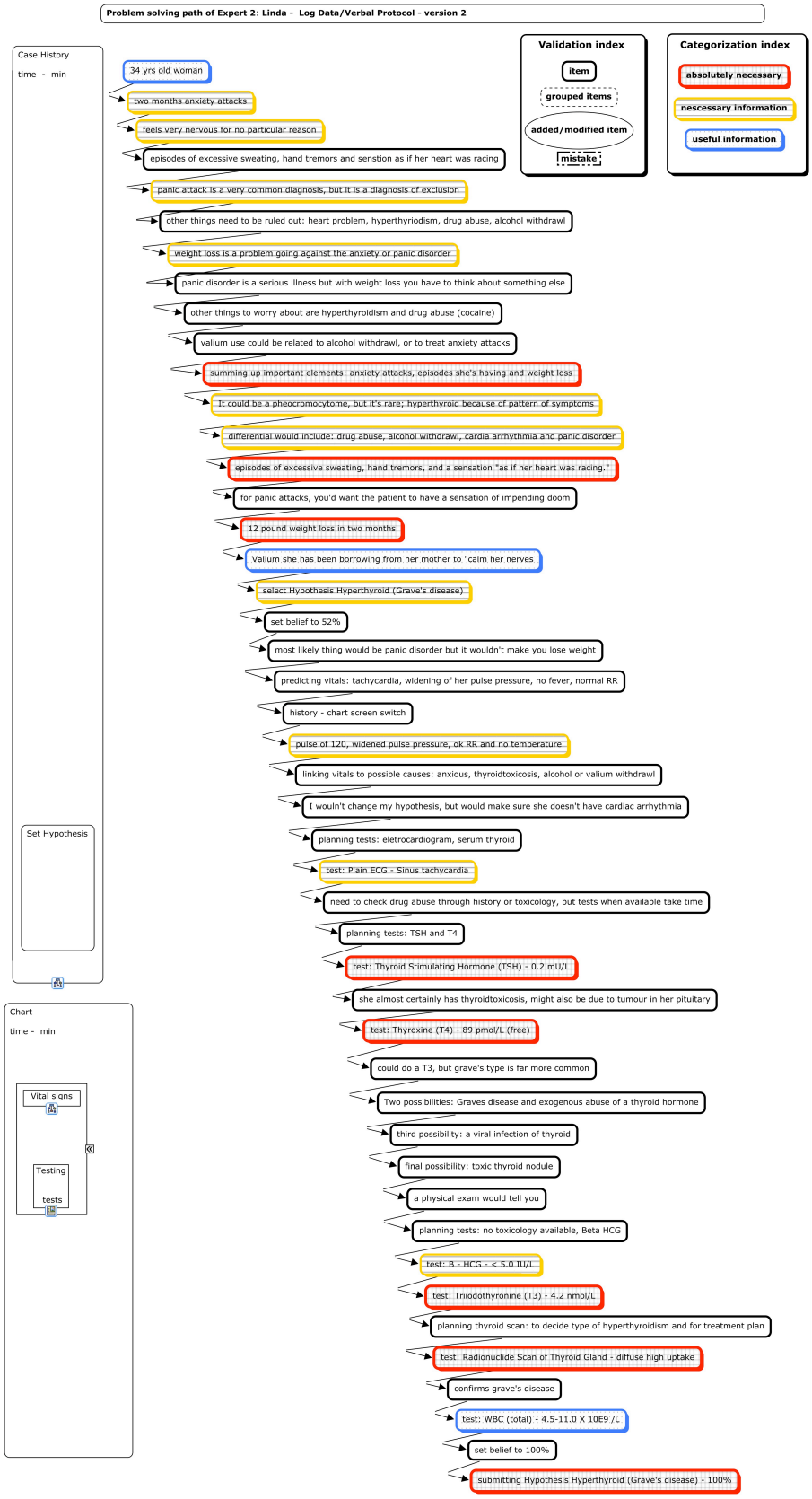


Figure 14. Overview of individual categorized visual representation of expert 2

Merged Representation

The previous description represented convergent common evidence as expressed and selected by expert 2. He identified five extra items of evidence that are not part of the 14 items in common with other experts. Figure 15 illustrates what experts have as common evidence prior to a hypothesis selection or conducting diagnostic tests. This section illustrates the different number of hypotheses formulated by experts, which ones were common, how they varied, and how they were justified.

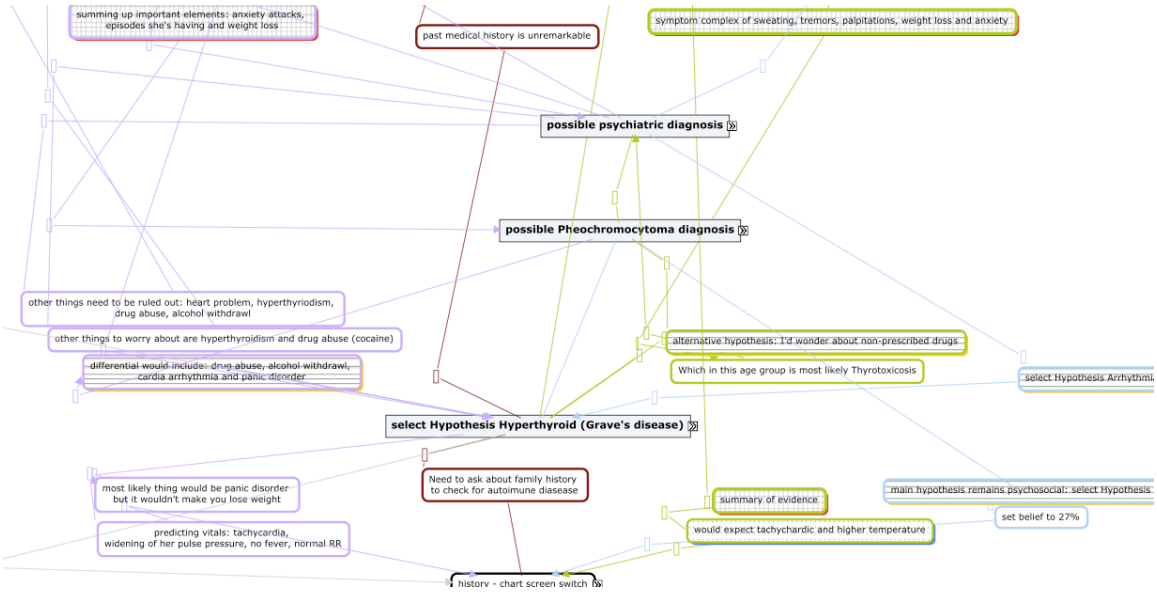


Figure 15. Section of merged visual representation for case 3 at level 1

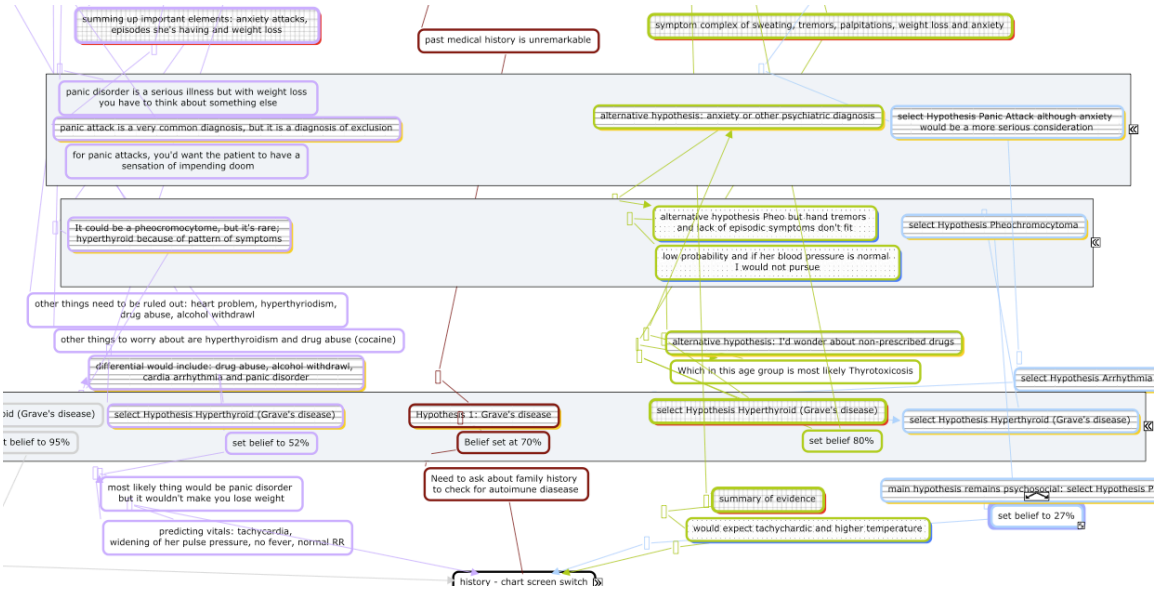


Figure 16. Section of merged visual representation for case 3 at level 2

Note that expert 1 was almost entirely absent from this section since he did not mention any hypothesis prior to selecting the Grave's disease hypothesis as shown at the bottom left of the figure 16. Similarly, expert 3, in the middle position, did not offer a hypothesis prior to her selection of the hypothesis at the bottom. Expert 2, 4 and 5 all considered panic attack or other psychiatric diagnoses prior to mentioning the unlikely hypothesis of pheocromocytoma, then expert 2 and 3 mentioned options of drug abuse while expert 2 went into detail about which drugs could induce these types of symptoms. Expert 2 and 5 considered the possibility of arrhythmia, then all five experts selected Grave's disease as their working hypothesis except expert 5 who continued and changed his hypothesis to panic attack, which he considered a more likely cause.

This brief section illustrates that even for a case that has a high level of agreement and a high level of confidence with the final hypothesis, there is a wide variety of options in the potential hypotheses formulated. It is not clear whether expert 1 and 3 simply did not express these other alternatives or if they went on "auto-pilot" mode for a case that did not show any potential challenge. The high level of agreement is

shown through the merged representation, with only 18 of the 143 categorized episodes that were not included in the common evidence.

Synthesis

Degree of Convergence in the Analysis of Protocols: Outcomes and Process

Expert medical instructors solved a set of three patient cases in a computer-based learning environment by doing a “think aloud” as if they were doing a case presentation. Outcome and process measures are compared to explore how they could inform the interpretation of valid performance in an ill-defined problem-solving context. The purpose of comparing both types of measures was to estimate if both solution processes and outcomes converge or diverge in a similar way for each case. The analyses aimed at providing an effective representation of how experts teach and diagnose patient cases to identify possible outcome and process evaluation criteria. For the case resolution task in the computer-based learning environment Bioworld, outcome measures involved the final answer for each case, the level of confidence in this final answer, along with the prioritized list of evidence supporting the answer. BioWorld recorded process measures dynamically, as participants interacted with the case in an effort to find the solution. Measures included the type and number of diagnoses or hypotheses selected prior to their final diagnosis and the evidence collected from the patient scenarios such as symptoms and diagnostic tests they conducted as they went through the patient chart; these interactions were examined in the context of the verbal protocol data. For each case, experts were compared against one another with respect to the list of hypotheses they considered, their confidence with respect to their hypotheses, as well as the diagnostic tests they ordered throughout the resolution process. The verbal protocols were examined as well in terms of the time participants took to solve the problem, the number of words uttered, and the number of ideas and

episodes summarizing each case. The portions of the episodes categorized as key elements for the problem solving of each case were also considered.

Experts agreed on most final diagnosis, and to some level — approximately 60 percent — on outcomes leading to the final diagnosis for each case. The exception on the final diagnosis agreement relates to Case 1, which is discussed in more detail in the discussion section. For the other cases all experts submitted the same answer for each case. Experts also had a high level of confidence in their diagnoses with two exceptions for Case 1. These exceptions of lower confidence may have been due to the low prevalence of the disease rather than the lack of confidence of experts in their answer. Given this result, I looked at the first case in more detail to insure that the assumptions for considering the performances as comparable were not challenged. To ensure that the cases were easy for our experts I asked them about the difficulty level of these cases as well as their experience with similar types of cases. “Easy” types of problems were chosen based on the level of competence and experience of participants to facilitate a comparison of successful and accurate performances. One of the three cases posed a threat to this assumption: the first case, Lydia, was rated as difficult by two of the five experts due to its rarity. The odds of seeing a case of pheocromocytoma are quite low, therefore making this a more challenging case; however, because both these experts had experience with this type of case enabled us to consider their performance as valid and correct.

The list of prioritized evidence submitted to support the final diagnosis shows very little convergence in terms of prioritization for all cases; however, when ignoring the rank order of the evidence, the content of the list shows a given amount of convergence. The consensus rate on the evidence for Case 1 is 42 percent, for Case 2, it is 55 percent, and for Case 3, the rate is 78 percent. Evidence is identified as similar when at least three of the five experts incorporated it in their list to support their final diagnosis.

When examining the convergence of the evidence concerning the process of the case resolution, it heads in a similar direction: there was 60 percent for case 1, 87 percent for case 2, and 80 percent for case 3. The process measures regarding evidence for each case yield a better agreement rate for the three cases. This result suggests that both outcomes and process measures could be used to support and assess performance of case resolution.

As experts worked through the problem and collected evidence they also recorded their level of confidence in their current hypothesis. The confidence range rather than the final confidence level demonstrated interesting differences between experts. In Case 1, confidence levels fluctuated from 60 to 100 (40) percent for expert 2, from 55 to 70 (15) for expert 3, and from 9 to 60 (51) percent for expert 4. Case 1 appeared the most problematic, demonstrating lowest overall confidence by experts and the highest variation between experts in confidence levels and less variance in confidence was found on the other 2 cases. These fluctuations that experts have in their hypothesis also gains at being interpreted by expert rather than by case. As expert 4 clearly stated during Case 2, “So 80 percent for me is a fairly high thing, I guess ‘cause I'm a skeptic and I would never say 100 sort of thing.” As a result, the 99 percent confidence level she submits for this case can be interpreted as being similar to the 100 percent confidence level submitted by the three other experts for the same case.

The time taken for by each expert for each case varies. On average experts took more time for Case 1, a bit less for Case 2 and the least time for Case 3. This decline could be due to case familiarity, ease of case solving over time, fatigue, or spending less time on later problems; however, not all experts showed a decrease in time and time is not an indicator of the quality of the reasoning or thinking skills involved in their performance. Even though time is often used as a performance variable in many

examination scenarios, our data suggest that time taken is not a valid indication of better or worse reasoning skills.

Comparing Participants' Categorization of Key Elements of their Reasoning Processes

I expected that the categorization of experts with regards to their reasoning process for each case would lead to shared identification of key elements for a case; moreover, I expected that these expert instructors would have a greater level of agreement on what elements were absolutely necessary elements versus what elements were the important and useful elements. As mentioned in the Results section on case 1, the differential use of weights did not help in narrowing down the most important elements; experts did not use the weights in any comparable way. This lack of agreement on the degree of evidence importance did not show any similarities. The result is similar to the lack of agreement in the prioritized lists submitted by experts and possibly attributed to lack of training and instruction related to this categorization task, yet the task is not a coding task, nor a self-assessment task since the experts are not evaluating their strength or weaknesses but selecting the important steps required for successful resolution of the case. It could be compared to a performance assessment but without the inherent biases or misinterpretation related to rating and categorizing another's performance. In Medicine, performance assessment raters also fail to distinguish between more than one or two dimensional concepts (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007).

Analysis showed a comparable use of the categorization in terms of the number of episodes selected from the entire representation. As presented in the section on case 1, experts categorized, on average, 53 percent of all their episodes. For the other cases, results were similar with averages of 54 and 63 percent for case 2 and 3. Only expert 5 systematically categorized over 85 percent of the episodes. Considering that this expert

did not produce as many words as others, it could be linked to a less verbose style where every word counts; yet, this systematic categorization of every episode created by the researcher, combined with no change in the validation done by this expert, raises some concerns over the appropriateness of the task with this expert.

Overall analyses of the categorization task were key in demonstrating a high degree of convergence in the reasoning process among experts. For all three cases, the common evidence identified varied from 16 to 24 per case. As mentioned in the previous section on process measures, the convergence for the common evidence is 60 percent for Case 1, 87 percent for Case 2, and 80 percent for Case 3. The main purpose of this activity is conclusive as it enabled a focus on the analysis of key decisions related to the reasoning performance of experts.

Reliability and Validity

Reliability of the Coding Process

The reliability of the segmentation task was completed on all three cases with a Cohen's kappa of 80.94 percent. The creation of episodes and corresponding summaries was also coded by two researchers and compared afterwards for one case. The number of episodes differed by 3 nodes; the content of the nodes was not identical but none of the categorized episodes was missed and the meaning remained similar.

Validity of the Visual Representation

The validation task with participants primarily aimed at improving the validity of the summary and content of these nodes. I asked participants to validate the accuracy of the summary and they were able to interact with their own representations easily. The observations, comments, and the number of changes made by participants in the validation phase confirmed that the summary done by the researcher was suitable; for example, in case 1, all five participants used the verbal transcript inside the nodes to

verify what they had said and to refresh their memories. Expert 1 made no modification, expert 2 made two modifications, expert 3 made one, expert 4 made three modifications, and expert 5 did not change anything for case 1. The amount and types of modifications for the two other cases were similar.

DISCUSSION

The literature on case-based instruction demonstrates that cases help situate learning in authentic contexts where students learn as they solve problems. Yet, the assessment of case-based approaches poses a critical challenge to current educational assessment practices. Problem-solving abilities at the heart of case-based teaching cannot be measured by standard measurement practices (Pellegrino et al., 2001). Another shortcoming of assessment practice is related to the lack of attention given to the decision-making process leading to acceptable solution(s). Moreover, the goal of fostering multiple perspectives and no unique right answer for a case challenges the concept of reliability, as conceived by current psychometric paradigm. Current assessment approaches use standards and principles that originate from a psychometric measurement where learning is considered under a trait-like approach. As the notion of learning is moving towards a competency based approach and where learning is conceived as developmental and contextual in nature, assessment practices need to reflect this perspective. In this research I proposed to treat problem solving as a performance and to use a bottom-up approach to collect evidence of experts' reasoning. Aligned with key instructional goals of case based learning, these case specific models focusing on the reasoning processes of a number of experts are built to inform assessment practices.

I explored this problem within the context of medical education, where I studied competent case resolution protocols of expert physicians to create blueprints of competent problem solving. The merged representations aim at gaining insight about the problem solving process by capturing rich descriptions and explanations occurring throughout the decision-making context. These blueprints correspond to the initial requirement of the evidence models in the context of an evidenced-centered

assessment system (Mislevy et al., 2003). By reaching a better understanding of the processes and outcomes related to competent performances, these evidence models and measures can inform the design of assessment practices aligned with case-based learning instructional goals and purposes. In this section I begin by discussing how these blueprints can be used to support assessment practices by showing how higher order thinking occurs in specific contexts, by emphasizing the problem solving process and by promoting multiple perspectives on a given situation. Then, I review different aspects of validity embedded in this performance approach to reasoning and problem solving. Finally, I discuss the limitations of the study and its implications for future research before validating its original contribution to knowledge.

Supporting Instructional Goals of Case Based Learning

Representing Thinking in Context

The visual representations can inform the contextual nature of reasoning that occurs as experts are collecting data, generating hypotheses, testing hypotheses before ruling out or ruling in a final hypothesis. Even if the analyses are primarily anchored to what participants think and do, they also depict how this content is used in context by including justifications in which experts evaluate elements of the problem and make their decisions. For example, while participants selected and commented on weight loss occurring in all three cases, they linked it and justified it differently depending on the context as well as their understanding of the problem at that point. Weight loss in Case 1 is evaluated differently in case 1 by expert 2 than in Case 2; in Case 1 he combined weight loss with other symptoms to complete the clinical picture related to a potential endocrinal problem while in case 2 he interpreted the weight loss within the context of a fit young adult who should not be losing that much weight in a month. He questioned the objectivity of the weight loss only in the first case where he mentions that it would need

to be documented. Comparable to Faremo's studies of clinical reasoning in the context of BioWorld (Faremo, 2004), experts' transcripts all revealed numerous occurrences of planning, inferences, reviewing and summarizing statements for each case. However, given the limited number of cases performed by experts it was not possible to reach any meaningful comparison about specific contextual elements triggering these statements.

Emphasizing the Reasoning Process not Only the Final Answer

The representations not only portray the problem solving processes of more than one competent problem solver from a community of practice but it compares these decision making processes to show similarities and differences throughout the resolution. As emphasized by Voss and Post (1988), the display of argumentation is a way to judge the quality of solutions for ill-defined problems since there are no universal criteria or absolute truth for any given problem. Measures of process and outcome from expert problem solving performance complement each other and they could be used to interpret future novice performances. The representations for each case provide a comprehensive repertoire of shared and divergent decisions throughout the problem solving task for each case.

Fostering Multiple Perspectives

The focus of this study is to analyze and represent similarities and differences among experts' reasoning processes to emphasize the dynamic and contextual nature of expert knowledge. Instead of trying to find out which expert is "right" or "wrong", it is important to find ways to acknowledge the inherent variability of human performance. The visual representation for each case suggests that even though there is some variability in the problem solving processes, there is agreement about key elements of the problem. These models have the potential to inform assessment practices where

instructors and students learn to differentiate acceptable versus non acceptable variability in the reasoning process of students.

Validity Issues

Both similarities and differences in performances can inform the rating procedures of future performance and help improve the construct-irrelevant variance. Construct-irrelevant variance and construct under-representation are two major threats to validity (Messick, 1995). The term construct under-representation refers to the inability to incorporate important dimensions or facets of the construct it pretends to assess (Cook & Campbell, 1979). Often the assessment does not reflect the complexity of the construct that is being measured. In a problem solving context, this term is used to discuss the challenge of defining or incorporating essential aspects of the construct related to the cognitive requirements of a task and the ways in which it can be solved. The use of performance models can inform the design and development of assessment procedures by enabling raters to base their inferences on the collection of evidence, both procedural and empirical, to evaluate not just an outcome but the performance in context based on samples of real performances instead of idealized ones. Furthermore, anchoring assessment design in empirical sampling of more than one expert performance might prevent the construct irrelevant variance which relates to variance in the data that is not relevant to the interpretation of the construct of interests (Messick, 1989). Showing the variance that exists at the level of “competent,” practitioners can challenge the notion of what is and is not appropriate to evaluate. For example, when assessing learners’ performance on their reasoning process, grading of elements corresponding to common evidence will be different than on aspects that do not lead to convergence among experts. In other words, when learners repeat similar sub-optimal

reasoning that experts might also have exhibited, the grading of the performance might lead to more nuanced feedback.

Overall, the use of the case specific models could improve what is referred to as systemic validity (Frederiksen & Collins, 1989) as it has the potential to inform both the learners and the instructor about the nature of competent performance. If learners can understand where their relative weaknesses are, it improves the transparency of the assessment procedure and enables a better evaluation of the validity claim and the corresponding inference of proficiency related to its scoring in small-scale educational settings (Kane, 1992). To formulate clear and transparent arguments concerning the data collected one must understand the meaning of the data in relationship to both the global and contextual nature of the performance used as a standard.

Limitations of this Study

Sample Size, Number of Cases and Design

This study has a number of limitations and findings should be interpreted in appreciation of these limitations. The number of participants is small because recruitment of experts with the necessary qualifications was a challenge. Similarly, the small number of cases presented in this study is related to the scarcity of medical expertise to develop the cases, as each case requires a significant investment of time and resource. The cases were not designed to be representative of the field of internal medicine, nor do they significantly cover the topic of endocrinology which limit the implication of the findings. Finally, it is important to note that the researcher is not a medical content expert and thus the coding and analysis was based on her own judgments with the support of one medical content expert who gave regular advice. A second coder with content knowledge would have improved the robustness of the analysis process.

Modeling Diagnostic Reasoning or Teaching Performance about Diagnostic Reasoning?

The specific choice of instructional experts and the framing of the task as a "teach-aloud" case presentation aimed at improving the authenticity of the think-aloud procedure by linking it to the more authentic task of case presentation for participants. It also aimed at avoiding the problems of knowledge encapsulation reported in experts' protocols when they solved cases that were not challenging to them (Boshuizen & Schmidt, 1992). In their study of expert reasoning they found that experts did not link or refer to biomedical knowledge when solving cases that were not challenging. In the present study we chose relatively easy cases but we asked instructional experts to present and explain their reasoning for a novice audience. Their transcripts do reveal clear links of how biomedical knowledge informs and connects to the presentation and development of the clinical cases. For example, in the diabetes case two experts explained how the changes of blood glucose concentration were affecting the vision.

A valid question remains regarding the exact object of analysis of this study; one could argue that the analysis and models either reflect the reasoning performance of experts or that they pertain to the teaching performance of these experts. The nature of the task for the study incorporates both the performance and the reasoning about the performance. It is difficult to know whether experts did what they would have done normally in a "pure" simulation task: experts may not perform exactly the way they teach, yet it is interesting to note that three of the five experts mentioned in their transcripts that if they were actually teaching they would not instruct students the way they are actually performing. For example in the case of expert 5, he insisted on submitting the diagnosis of panic attack because if he had been teaching it would have been the appropriate diagnosis to submit. In the case of experts 3 and 4, the quotes below show that they did one thing but would insist on the importance of doing it a different way if they were teaching. Expert 4 mentions that even though she is now planning to go directly to the

diagnosis test for pheocromocytoma, in a teaching situation she would show that it is necessary to begin by first ruling out more common secondary causes:

> *If I were teaching a student about this, I would say well, you're going — you're going to want to do some of the tests for hypertension, both to rule out a secondary cause and also to rule out ah, and to assess whether the ah blood pressure has had an effect on things like kidneys and heart.* Quote from expert 4 on case 1

On the same case, Expert 3 says that if she was teaching and in a real situation she would repeat the diagnostic tests to confirm the result:

> *This test doesn't exclude it [the hypothesis] and what I would teach is that you need to ah, the standard is to repeat it on three different occasions. Cause it's a cyclic pattern if it is due to increased production of epinephrine and norepinephrine.* Quote from expert 3 on case 1

There may be variation between experts in the way they teach and the way they perform. Investigating this issue could be an interesting research direction. However, for the purpose of this study we were intent on building an evidence model that would contribute to the assessment of learning and thus our focus was on modeling what instructors see as important, even if they teach differently then their practice on the wards.

Educational Implications and Future Directions

Modeling Novice Level Performances

The modeling of the performance of experts is only one half of the evidence model in the evidence design framework. Further work would be required to model the performance of novices. The method would probably need to be adapted as the task of case presentation or teach aloud would not be a nature or learned skill for them. The sampling of different levels of novices would enable setting criteria and probabilities to the specific aspects of the problem solving processes and outcomes. By building a more comprehensive or complete representation of the different ways in which a patient case can be solved, better assessment and feedback routines can be adapted to individual

differences in diagnostic reasoning; furthermore, developing a more fine-grained analysis of learner differences along the proficiency dimension would enable gaining understanding of how reasoning skills develop. Once such differences are understood, these representations can be used for instructional purposes, providing appropriate levels of scaffolding based on these complex problem-solving models.

Use of Models as Worked Examples to Support Learning

One avenue that we have briefly explored is the use of these models as worked examples to provide novice learners with feedback about possible ways to reason and solve the case. Worked examples are common instructional tools used in mathematics and physics to teach problem solving skills, but their application to less well-defined domains remain a challenge (Moreno, 2006). We have started to use the visual representations as a form of feedback for novice learners where they review them after they attempt to solve a case and reflect on how their solutions may have differed from that of an expert. Results suggest the use of these representations improves students' awareness and critical appraisal of their own reasoning processes (Gauthier, Lajoie, Naismith, & Wiseman, 2008). Our next step is to use the merged representation of experts in a systematic manner with medical students.

Building on the Reflection Task

Recent literature on expertise suggest that the ability of experts to continuously monitor and improve their performance is one of the distinguishable and vital characteristics that allows them to achieve consistent, measurable and reproducible above-average performance. Thus, studying how experts initiate and identify the limits of their knowledge can yield insight into our understanding of the self-assessment ability. In this study, the verbal protocol of experts shows occurrence of self-assessment where experts often challenge the state of their knowledge and ability. A future study could

build on a similar task to explore the nature and context in which experts exhibit moments of awareness of having reached their limits. Such studies could provide insight into our understanding of how experts monitor their weaknesses and inform the design of assessment and instruction to promote the ability of reflection and self-assessment in learners.

Original Contribution to Knowledge

A great deal of assessment research is about high-stakes assessment (Dochy & Moerkerke, 1997), even though most assessment occurs at the classroom level. In medicine, small-scale assessment corresponds to 80 percent of student assessment (Govaerts et al., 2007), but little scientific attention is dedicated to the understanding of competent performance for typical cases presented in these settings. This work proposes an empirical method and design that explores the process and decision making that can improve human judgment involve in the assessment process occurring in these small-scale instructional settings. The models produced show convergence in the way case-based knowledge is applied in case-specific context. These, in turn, can be used to inform assessment that is valid and that reinforces goals promoted by case based instruction.

BIBLIOGRAPHY

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communication, 7*(1), 39-59.

Airasian, P. W. (1997). Empiricism and Values: Two Faces of Educational Change. *International Journal of Educational Research, 27*(5), 433-445.

Anastasi, A. (1996). *Psychological testing* (7th ed.). New York: Macmillian.

Atkinson, J. M., & Heritage, J. (1984). Transcript Notation. In J. M. Atkinson & J. Heritage (Eds.), *Structures of Social Action: Studies in Conversation Analysis* (pp. ix-xvi). Cambridge: Cambridge University Press.

Baker, E. L. (2007). 2007 Presidential Address The End(s) of Testing. *Educational Researcher, 36*(6), 309-317.

Barnes, L. B., Christensen, C. R., & Hansen, A. J. (1994). *Teaching and the Case Method : Text, Cases, and Readings* (3rd ed.). Boston, MA: Harvard Business School Press.

Barrow, H. S. (1988). *The tutorial process*. Springfield, IL: Southern Illinois University Press.

Barrows, H. S. (1986). A taxonomy of problem-based learning methods. *Medical Education, 20*, 481-486.

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-Based Learning: An Approach to Medical Education*. New York, NY: Springer Publishing Company.

Beckwith, J. B. (1991). Approaches to learning, their context and relationship to assessment performance. *Higher Education, 22*, 17-30.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364.

Birenbaum, M. (2003). New insight into learning and teaching and their implications for assessment. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (Vol. 1, pp. 13-36). Dordrecht: Kluwer Academic Publisher.

Birenbaum, M., & Dochy, F. (1996). *Alternatives in assessment of achievement, learning processes and prior knowledge*. Boston: Kluwer Academic Publishers.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals, by a committee of college and university examiners*. New York: Longman, Green.

Boekaerts, M. (1996). Self-regulated learning at the junction of cognition and motivation. *European Psychologist, 1*(2), 100–112.

Boshuizen, H. P., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science, 16*, 153-184.

Boud, D. (1985). *Problem-based Learning in Education for the Professions*. Sydney: Higher Education Research and Development Society of Australasia.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399 - 413.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2*(4), 331-350.

Bransford, J. D., & Schwartz, D. L. (2009). It takes expertise to make expertise. In K. A. Ericsson (Ed.), *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*. Cambridge, UK: Cambridge University Press.

Brennan, R. L. (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Educational Measurement, 38*(4), 295-317.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Camtasia Studio. (2003). Okemos, Michigan: TechSmith Corporation.

Carter, K. (1993). The place of story in the study of teaching and teacher education. *Educational Researcher, 22*(1), 5-12.

Carter, K. (1999). What is a case? What is not a case? In B. B. L. M. A. Lundeberg, & H. L. Harrington (Ed.), *Who learns what from cases and how: The research base for teaching and learning with cases*. Mahwah, NJ: Lawrence Erlbaum Associates.

Charness, N., Hoffman, R. R., Feltovich, P. J., & Ericsson, K. A. (Eds.). (2006). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge, England: Cambridge University Press.

Christensen, C. R., Hansen, A. J., & Moore, J. F. (1987). *Teaching and the Case Method*. Boston: Harvard Business School.

Cizek, G. J., & Gary, D. P. (1996). Learning, achievement, and assessment: Constructs at a crossroads. In *Handbook of Classroom Assessment* (pp. 1-32). San Diego: Academic Press.

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. New York: Cambridge University Press.

Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold & M. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 75-87). Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings*. Chicago, Illinois: Rand McNally.

Cox, K. (2001). Stories as case knowledge: case knowledge as stories. *Med Educ, 35*(9), 862-866.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Cambridge, MA: MIT Press.

Cronbach, L. J. (1971). Test validation. In R.L.Thorndike (Ed.), *Educational measurement* (Second ed.). Washington, D.C: American Council on Education. .

Cronbach, L. J., & Meehl, P. C. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Dewey, J. (1916). *Democracy and Education*. New York: Free Press: Free Press.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston: DC Heath and Company.

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279-298.

Dochy, F. J. R. C., & McDowell, L. (1997). Introduction: Assessment as a Tool for Learning. *Studies in Educational Evaluation, 23*(4), 279-298.

Dochy, F. J. R. C., & Moerkerke, G. (1997). Assessment as a major influence on learning and instruction. *International Journal of Educational Research, 27*(5), 415-432.

DuBois, P. H. (1964). A test-dominated society: China, 1115 B.C.-1905 A.D. *Proceeding of Invitational Conference on Testing Problems*, (pp. 3-11), New York.

Elstein, A., Shukman, L., & Sprafka, S. (1978). *Medical Problem Solving*. Cambridge, MA: Harvard University Press.

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal, 324*(7339), 729-732.

Embretson, S., & Gorin, J. (2001). Improving Construct Validity with Cognitive Psychology Principles. *Journal of Educational Measurement, 38*(4), 343-368.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice, 32*(3), 179-186.

Epstein, R. (2007). Assessment in Medical Education. *N Engl J Med, 356*, 387 - 396.

Epstein, R. M., & Hundert, E. M. (2002). Defining and Assessing Professional Competence. *JAMA, 287*(2), 226-235.

Ericsson, K. A. (2006). An Introduction to The Cambridge Handbook of Expertise and Expert Performance: Its Development, Organization, and Content. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 3-20). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: verbal reports as data*. Cambridge, Mass.: MIT Press.

Etymology Online. (2008). Definition of case.In Retrieved 02/12/2008, from http://www.etymonline.com/index.php?search=case&searchmode=none

Faremo, S. (2004). Examining Medical Problem Solving in a Computer-Based Learning Environment. Unpublished doctoral dissertation. McGill University.

Fenstermacher, G. D. (1994). The knower and the known: The nature of knowledge in research on teaching. *Review of Research in Education, 20*, 3-56.

Frank, J. (Ed.). (2005). *The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care*. Ottawa: The Royal College of Physicians and Surgeons of Canada.

Frederiksen, C. H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology*(7), 371-458.

Frederiksen, J. R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher, 18*(9), 27-32.

Gauthier, G., Lajoie, S. P., Naismith, L., & Wiseman, J. (2008). *Effectiveness of visual expert teachers' solution process on undergraduate students solving clinical cases*. Paper presented at the American Association of Medical Colleges (AAMC). San Antonio, Tx.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355-392.

Glaser, R., & Silver, E. (1994). Chapter 9: Assessment, Testing, and Instruction: Retrospect and Prospect. *Review of Research in Education, 20*(1), 393-419.

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Advances in Health Sciences Education, 12*, 239-260.

Greenhalgh, T., & Hurwitz, B. (1999). Narrative based medicine: Why study narrative? *BMJ, 318*(7175), 48-50.

Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*(1).

Haertel, E. h., & Herman, J. l. (2005). A Historical Perspective on Validity Arguments for Accountability Testing. *Yearbook of the National Society for the Study of Education, 104*(2), 1-34.

Haig, B. D., & Borsboom, D. (2008). On the Conceptual Foundations of Psychological Measurement. *Measurement: Interdisciplinary Research & Perspective, 6*(1), 1 - 6.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. *Child development and education in Japan*, 262-272.

Henderson, C., Yerushalmi, E., Heller, K., Heller, P., & Kuo, V. H. (2003). Multi-Layered Concept Maps for the Analysis of Complex Interview Data*. Proceeding of Physics Education Research Conference*Madisson, WY.

Herman, J. L. (1997). Large-Scale Assessment in Support of School Reform: Lessons in the Search for Alternative Measures. *International Journal of Educational Research, 27*(5), 395-413.

Herreid, C. F. (Ed.). (2007). *Start With a Story*. Arlington, VA: National Science Teachers Association Press.

Hoffman, R. R., & Lintern, G. (2006). Eliciting and representing the knowledge of experts. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *Handbook on expertise and expert performance* (pp. 203-222). Cambridge: Cambridge University Press.

Jonassen, D. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research & Development, 45*(1), 65-94.

Joughin, G., & Macdonald, R. (2004). A model of assessment in higher educational institutions: The Education Academy.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*(112), 527-535.

Kane, M. T. (2008). Terminology, Emphasis, and Utility in Validation. *Educational Researcher, 37*(2), 76-82.

Knight, P. (2006). The local practices of assessment. *Assessment & Evaluation in Higher Education, 31*(4), 435 - 452.

Knorr, C. K. (1981). *The Manufacture of Knowledge - An Essay on the Constructivist and Contextual Nature of Science*. Oxford: Pergamon Press.

Kolodner, J. L. (1993). *Case-based Reasoning*. San Francisco, CA: Morgan Kaufmann.

Kolodner, J. L. (2006). Case-based reasoning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 225-242). Cambridge: Cambridge University Press.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting Learning by Design (TM) into practice. *Journal of the Learning Sciences, 12*(4), 495-547.

Lajoie, P. S. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine. In K. A. Ericsson (Ed.), *The Development of Professional Performance: Approaches to Objective Measurement and Designed Learning Environments* (pp. 61-83). Cambridge: Cambridge University Press.

Lajoie, S. P., Faremo, S., & Wiseman, J. (2001). A knowledge-based approach to designing authoring tools: from tutor to author. In J. D. Moore, C. Redfield & L. W. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 77-86). Amsterdam: IOS Press.

Lajoie, S. P., Lavigne, N., Guerrera, C. P., & Munsie, S. D. (2001). Constructing knowledge in the context of BioWorld. *Instructional Science, 29*(2), 155-186.

Lave, J., & Wenger, E. (1991). *Situated Learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.

Linn, R. L. (1994). Performance Assessment: Policy Promises and Technical Measurement Standards. *Educational Researcher, 23*(9), 4-14.

Lissitz, R. W., & Samuelsen, K. (2007). Further Clarification Regarding Validity and Education. *Educational Researcher, 36*(8), 482-484.

Llewellyn, K. N. (1948). The current crisis in legal education. *Journal of Legal Education, 1*, 211-220.

Lundeberg, M. A., Levin, B. B., & Harrington, H. L. (1999). *Who learns what from cases and how: The research base for teaching and learning with cases*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lundeberg, M. A., & Yadav, A. (2006). Assessment of Case Study Teaching: Where Do We Go From Here? Part I. *Journal of College Science Teaching, 35*(5), 10-13.

McCarthy, J. (1977). Epistemological problems in artificial intelligence. *Proceeding of Fifth International Joint Conference on Artificial Intelligence (IJCAI-77)* (pp. 1038-1044), Cambridge, Massachusetts.

Merseth, K. K. (1994). *Cases, case methods, and the professional development of educators*. Washington, DC: ERIC Clearinghouse on Teaching and Teacher Education (BBB30990).

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63-67.

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439-483.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centered Design. CSE Report 632*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3-62.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Rejoinder to Commentaries for "On the Structure of Educational Assessments". *Measurement: Interdisciplinary Research & Perspective, 1*(1), 92 - 101.

Moreno, R. (2006). When worked examples don't work: Is cognitive load theory at an Impasse? *Learning and Instruction, 16*(2), 170-181.

Moskovitz, M. (1992). Beyond the Case Method: It's Time to Teach with Problems. *Journal of Legal Education, 241*(2), 241-270.

Moss, P. A. (1994). Can There Be Validity Without Reliability? *Educational Researcher, 23*(2), 5-12.

Mylopoulos, M., & Regehr, G. (2007). Cognitive metaphors of expertise and knowledge: prospects and limitations for medical education. *Medical Education, 41*(12), 1159-1165.

Nendaz, M. R., & Tekian, A. (1999). Assessment in Problem-Based Learning Medical Schools: A Literature Review. *Teaching and Learning in Medicine, 11*(4), 232 - 243.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norman, G. (2004). Editorial – Beyond PBL. *Advances in Health Sciences Education, 9*(4), 257-260.

Novak, J. D., & Cañas, A. J. (2006). *The Theory Underlying Concept Maps and How to Construct Them. Technical Report IHMC CmapTools 2006-01*: Florida Institute for Human and Machine Cognition.

Oxford English Dictionary. (2004) (11 ed.). New York: Oxford University Press Inc.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10*(1), 91-116.

Patterson, E. W. (1951). The case method in American legal education: Its origins and objectives. *Journal of Legal Education, 4*(1), 1-24.

Pellegrino, J., & Chudowsky, N. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 12*(1), 75-83.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Advances in the sciences of thinking and learning. In J. W. Pellegrino, N. Chudowsky & R. Glaser (Eds.), *Knowing What Students Know: The Science and Design of Educational Assessment* (pp. 59-110). Washington, D.C: The National Academic Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The Science and Design of Educational Assessment*. Washington, D.C: The National Academic Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Rethinking the foundations of assessment. In J. W. Pellegrino, N. Chudowsky & R. Glaser (Eds.), *Knowing What Students Know: The Science and Design of Educational Assessment* (pp. 17-37). Washington, D.C: The National Academic Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pople, H. E., Jr. (1982). Heuristic Methods for Imposing Structure on Ill-Structured Problems: The Structuring of Medical Diagnostics. In P. Szolovits (Ed.), *Artificial Intelligence in Medicine*. Boulder, Colorado: Westview Press.

Posner, G. J. (1985). *Field experience: A guide to reflective teaching*. New York: Longman Publishing Co.

Rikers, R. M. J. P., Loyens, S. M. M., & Schmidt, H. G. (2004). The role of encapsulated knowledge in clinical case representations of medical students and family doctors. *Medical Education, 38*(10), 1035-1043.

Sadler, D. R. (1998). Formative assessment: revisiting the territory. *Assessment in Education, 5*(1), 77-84.

Savin-Baden, M. (2004). Understanding the impact of assessment on students in problem-based learning. *Innovations in Education and Teaching International, 41*(2), 221 - 233.

Savin-Baden, M., & Howell Major, C. (2004). *Foundations of problem-based learning*. Maidenhead: SRHE/Open University Press.

Schank, R. (1998). *Inside Multi-Media Case-Based Instruction*. Mahwah, NJ: Lawrence Erlbaum Associatees, Inc.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. Erlbaum: Hillsdale, NJ.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory and Cognition, 21*(3), 338-351.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schön, D. A. (1987). *Educating the reflective practitioner*. San Fransico: Jossey-Bass.

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (Eds.). (2000). *Cognitive Task Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schuwirth, L. W. T., & der Vleuten, C. v. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*(4), 296-300.

Schwab, J. J. (1971). The Practical: Arts of Eclectic. *The School Review, 79*(4), 493-542.

Scriven, M. (1967). The methodology of Evaluation. In R. Tyler, R. Gagné & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation (AERA Monograph Series on Curriculum Evaluation)* (Vol. 1, pp. 39-83). Chicago: Rand McNally.

Segers, M., Dochy, F., & Cascallar, E. (Eds.). (2003). *Optimising New Modes of Assessment : In Search of Qualities and Standards* (Vol. 1). Dordrecht: Kluwer Academic Publishers.

Shepard, L., Hammerness, K., L., D.-H., & Rust, F. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275-326). San Francisco: Jossey-Bass.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7).

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Washington, DC:: National Council on Measurement in Education and American Council on Education/Praeger.

Shulman, J. H., & Colbert, J. A. (1987). *Cases as Catalysts for Cases*. Paper presented at the American Educational Research Association. Washington, DC.

Shulman, L. (1992). Toward a pedagogy of cases. In J. H. Shulman (Ed.), *Case methods in teacher education* (pp. 1-30). New York, NY: Teachers College Press.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4-14.

Sireci, S. G., & Hambleton, R. K. (1997). Future Directions for Norm-Referenced and Criterion-Referenced Achievement Testing. *International Journal of Educational Research, 27*(5), 379-393.

Sleeman, D., & Brown, J. S. (1982). Introduction: Intelligent Tutoring Systems. In D. S. J. S. Brown (Ed.), *Intelligent Tutoring Systems* (pp. 1-11). New York: Academic Press.

Smith, M. L. (1991). Put to the Test: The Effects of External Testing on Teachers. *Educational Researcher, 20*(5), 8-11.

Smith, M. L., & Fey, P. (2000). Validity and Accountability in High-Stakes Testing. *Journal of Teacher Education, 51*(5), 334-344.

Sperle, D. H. (1933). *The case method technique in professional training: A survey of the use of case studies as a method of instruction in selected fields and a study of its application in a teachers college* (Vol. 571). New York: Bureau of Publications, Teachers College, Columbia University.

Sudzina, M. R. (Ed.). (1999). *Case Study Applications for Teacher Education: Cases of Teaching and Learning in the Content Areas*. Boston: Allyn and Bacon.

Sykes, G., & Bird, T. (1992). Teacher education and the case idea. In G. Grant (Ed.), *Review of research in education* (Vol. 18, pp. 457-521). Washington, DC: American Educational Research Association.

Taras, M. (2005). Assessment - Summative and formative - some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Teich, P. F. (1986). Research on American law teaching: Is there a case against the case system *Journal of Legal Education, 36*, 167-188.

Trochim, W. (2000). *The Research Methods Knowledge Base* (2nd ed.). Cincinnati, OH: Atomic Dog Publishing.

Van Den Heuvel-Panhuizen, M. (1995). Student-Generated Problems: Easy and Difficult Problems on Percentage. *For the Learning of Mathematics, 15*(3), 21-27.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*(1), 41-67.

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes. *Medical Education, 39*(3), 309-317.

van Dijk, T. A. (1985). Semantic discourse analysis. In T. A. v. Dijk (Ed.), *Handbook of Discourse Analysis* (Vol. 2, pp. 103-135). London: Academic Press.

Voss, J. F. (2005). Toulmin's Model and the solving of ill-structured problems. *Argumentation, 19*(3), 321-329.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.), *The nature of expertise* (pp. 261-285). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. .

Walter, M. I., & Brown, S. I. (1977). Problem Posing and Problem Solving: An Illustration of Their Interdependence. *Mathematics Teacher, 70*(1), 4-13.

Weiss, D. J., & Shanteau, J. (2003). Empirical Assessment of Expertise. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(1), 104-116.

Williams, S. M. (1992). Putting case-based instruction into context: examples from legal and medical education. *The Journal of the Learning Sciences, 2*(4), 367-427.

Woods, D., & Fassnacht, C. (2007). Transana - http://www.transana.org (Version 2.30). Madison, WI: The Board of Regents of the University of Wisconsin System.

APPENDIX A

Consent Form and Ethics

Short description and purpose of the study:

Learning, as suggested from research in cognitive sciences, “is an active process of mental construction and sense making “(Shepard, 2000). The cognitivist view emphasizes the need for students to play an active role in the learning process and the importance of the metacognitive abilities. This research examines the role of case modification and case construction as a problem posing task. Problem posing and problem solving are interdependent and distinctive tasks (Walter & Brown, 1977) but they both can provide a rich source of information about learner’s level of understanding (Van Den Heuvel-Panhuizen, 1995). Creating or modifying problems can be considered an ill-defined problem in that there is not one way to solve the problem and there is not necessarily one correct answer. Case modification as a problem posing activity should improve student's learning by emphasizing the active production of content as opposed to simple recall of information. In developing a case creation activity in medicine we want to draw from the “learning by teaching” approach and focus on the impact that this activity can have on the learning, reasoning, problem-solving and self-monitoring abilities of participants. This research project expands on already existing research from Susanne Lajoie on cognitive tools for enhancing self-regulation in medical students. The main objectives are to study cognitive processes used by medical personnel as they create, solve, share and discuss medical cases for an interactive computer-based learning environment. This activity also aims at developing and validating a database of cases with explanation and assessment criteria which can lead researchers to analyse knowledge structures at different stages of medical training (second year to expert).

Potential risks and Benefits:

Results of the study will be anonymous; we will not use any comments or specific information that could identify you in the write up of the research. Your participation will not affect your academic standing or performance ratings in any way. Participants will benefit from the study by learning about their own diagnostic process.

Please leave us your contact information if you wish to be sent a follow-up report about this research.

_____

_____

_____

Please put any comments you have on the back of this sheet. Thank you!

To satisfy McGill's requirement that there be proof of informed consent for all data collected you are asked to read the following, indicate your response to the statement in the box, fill in your name and sign.
To ensure your anonymity this page will then be stored separately and will be used (a) to establish that informed consent for the data was given (if required at a later date) and (b) to connect these data with other data collected for the same participant (if any).
If you have any comments, please feel free to indicate them on the back of this sheet.

I understand that I may withdraw from this study at my own discretion and for any reason at any time without any penalty. I understand that my identity will be protected and that all records will be coded to guarantee anonymity.
____ Yes, I want to participate (if yes, please answer following 5 questions)
____ No, I do not want to participate
1) **Voice recording**: I agree to be recorded while doing the case resolution activity. I understand that my identity will be protected and that only the main researcher and her two assistants will have access to the data.
YES_____ NO_____
2) **Screen capture:** I agree that the researchers may use a screen capture software to record my computer interaction doing the case resolution or modification task. I understand that my identity will be protected and that all records will be coded to guarantee anonymity.
YES_____ NO_____
3) **BioWorld and Cmap Log:** I agree that the researchers may have access to my log created from my case resolution/ modification in BioWorld or Cmap (concept map tool). I understand that my identity will be protected and that all records will be coded to guarantee anonymity.
YES_____ NO_____
4) **Follow-up Interviews**: I agree to be interviewed at a time that is convenient to me. I understand that the interview will be recorded on video. I understand that my identity will be protected and that only the main researcher and her two assistants will have access to the data.
YES_____ NO_____
5) **Use of data for presentation**: I agree that my data from video or audio recording be used for presentation or publication purposes.
YES_____ NO_____
Name (PLEASE PRINT):_____
Signature:_____ Date:_____
Contact information (e-mail or telephone number):

This research is being carried out by Geneviève Gauthier and Solange Richard, under my supervision. If at any time during the research you have any questions or concerns do not hesitate to contact me:
Dr. Susanne P. Lajoie,
Professor, Department of Educational and Counselling Psychology,
McGill University, Phone: 398-4242
Susanne.lajoie@mcgill.ca

Please put any comments you have on the back of this sheet. Thank you!

Please put any comments you have on the back of this sheet. Thank you!

APPENDIX B

Questionnaire

Section 1: General practice and specialization

Name:

1. What is your current status:

First year student		Third year student		Resident	
Second year student		Fourth year student		Practitioner	

2. How many years have you been practicing for (post FRCPC/ CSPQ)?

_____

3. If you are a resident or a practitioner, please identify your specialty:

Anesthesiology		Neurology/neurosurgery		Psychiatry	
Cardiovascular surgery		Obstetrics/gynecology		Radiology	
Emergency medicine		Otorhinolaryngology		Thoracic surgery	
Family practice		Opthalmology		Urology	
General surgery		Orthopedic/surger			
Hematology/oncology		Pathology		Other specialization	
Internal medicine		Pediatrics			
Gastroenterology		Plastic surgery			
Neonatology					

Section 2: Clinical teaching experience

4. Are you a medical teacher at the medical school? _____yes _____no
    a. If so how long have you been teaching for and in which specialization?
    _____
5. How many students can you have under your supervision per semester (officially or unofficially – either as an senior students or resident)?

_____

Section 3: Recent clinical experience

6. Please give us an approximation of the number of new and returning patients you have seen last week.

Total number of patients: _____
Number of new patient: _____
List the types or specialization of disease you see regularly: _____
_____

Post-Questionnaire: Case information

7. Have you ever encountered patients with similar diseases and if so how many and how long ago?

Case 1: Lydia - Pheochromocytoma

	I have never seen a patient with this disease (or I do not remember)
	I have never seen a patient with this disease but I remember studying about it …… ago
	1-2 cases …….. ago
	3-4 cases ……. ago
	more than 4 cases …… ago
	I see these types of cases on a regular basis

Case 2: Stephanie – Diabetes Mellitus Type 1 (complicated by diabetic Ketoacidosis)

	I have never seen a patient with this disease (or I do not remember)
	I have never seen a patient with this disease but I remember studying about it …… ago
	1-2 cases …….. ago
	3-4 cases ……. ago
	more than 4 cases …… ago
	I see these types of cases on a regular basis

Case 3: Linda - Hyperthyroid

	I have never seen a patient with this disease (or I do not remember)
	I have never seen a patient with this disease but I remember studying about it …… ago
	1-2 cases …….. ago
	3-4 cases ……. ago
	more than 4 cases …… ago
	I see these types of cases on a regular basis

Case Rating

8. According to you what is the level of difficulty of each case for each of the audience:

Levels of difficulties:

1 = too easy	2 = good revision	3 = challenging	4= difficult

Case 1 – Lydia – Pheochromocytoma:

First year student		Third year student		Resident	
Second year student		Fourth year student		Practitioner	

What are the concepts that could be taught with this case:

Suggestions to improve this case:

Case 2 - Stephanie – Diabetes Mellitus Type 1 (with Diabetic Ketoacidosis)

First year student		Third year student		Resident	
Second year student		Fourth year student		Practitioner	

What are the concepts that could be taught with this case:

Suggestions to improve this case:

Case 3 - Linda – Hyperthyroid

First year student		Third year student		Resident	
Second year student		Fourth year student		Practitioner	

What are the concepts that could be taught with this case:

Suggestions to improve this case:

Please write any other comments or suggestions at the back.

Interview Protocols

Interview Protocol - 1st Meeting

Welcome to our study on medical reasoning.
Before we go on with this session I would like you to read and sign the consent form for this research.
Feel free to ask questions if you have any.

The next step is to complete this one page questionnaire.

In this session you will be asked to solve patient cases and do it in the style of a case presentation as you would in a hospital setting with medical students. While going through the case presentation and trying to diagnose the main disease affecting a patient we need you think aloud as much as you can.

Before we go through a real case, let's do an introduction to the computer-based learning environment BioWorld.

(Going through the instructions with participant – showing a case in BioWorld)

Reminder before the participant is doing case presentation activity

It is important to remember to talk and read aloud as you go through the cases. The goal of this exercise is to record not only your answers, but your explanation for this cases.

You can submit your final diagnostic as soon as you feel confident that your answer is good and that you have selected and discussed relevant evidences related to the case.

After the 3 cases are done.
Before I let you go I would need you to complete the case index and give us feedback about the cases you have done (10min).

Interview Protocol - 2nd Meeting

Prior to meeting
- ✓ each case's map is loaded in Cmap - with styles for V2
- ✓ 2 paper versions of each map is available for participant to get overview, take notes
- ✓ transcript are printed and on the table for each case

Welcome back to the second phase of our study.
Here is the agenda for today.

1. I will present you with our visual summary of your case resolution for each case. We have a visual representation in which every summarized item is linked to what you said and did when solving the case in BioWorld in the first meeting. If you use the printed transcript and visual representation to help you revise the summary.

2. I need you to verify and validate this visual summary.
   you can add missing elements - added item style
   you can identify mistake done
   you can delete or modify items in the summary
   you can also make cluster or regroup elements that you think go together -

3. Then I will ask you to go over each case and tell me what are the items that you think are absolutely necessary for the resolution of the case. I mean what are the key elements that if you miss you would not be able to solve the case. Then I also need you to select elements that are necessary for the case resolution (but maybe not as crucial). Last you need to select any useful information for the case resolution.

To make it more comprehensible we have a little table with weights associated to each category.

Key elements	absolutely necessary (+5)	necessary (+3)	useful information (+1)

The next step is to complete this post questionnaire and case index

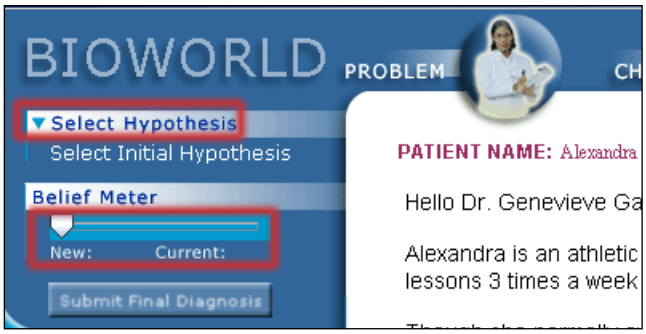
BioWorld Instruction

Instructions for Solving Cases in Bioworld

You will have the opportunity to solve patient cases as you would in a hospital setting. Your task is to diagnose the main disease affecting each patient. In addition to posing a diagnosis you need to select and organise the evidence to support and justify your decision.

Phase 1 Solving the Case -

In the first section of BioWorld there are a number of different activities that you can conduct to support your diagnostic reasoning.
You can select and change a diagnosis at any time but you are requested to register your confidence level for each diagnosis and if your confidence changes you can register that as well.

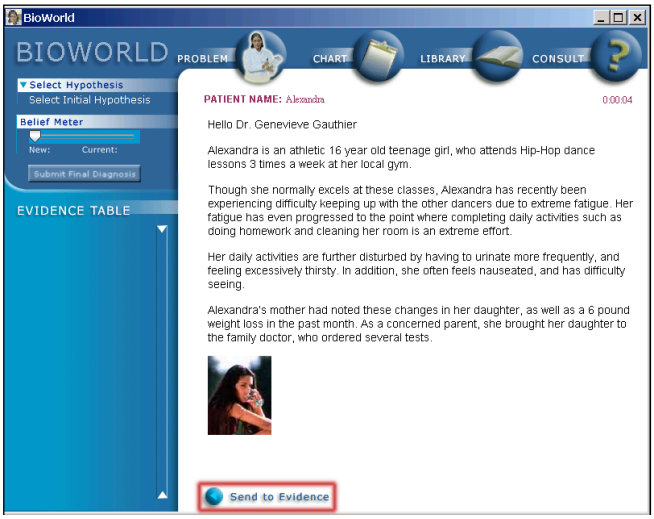


In addition you need to select supporting evidence as you browse the four different spaces in the BioWorld environment. Please note that you need to select an initial hypothesis prior to gain free access to navigate between any of the following spaces

Problem	
Chart	
Library	
Consult	

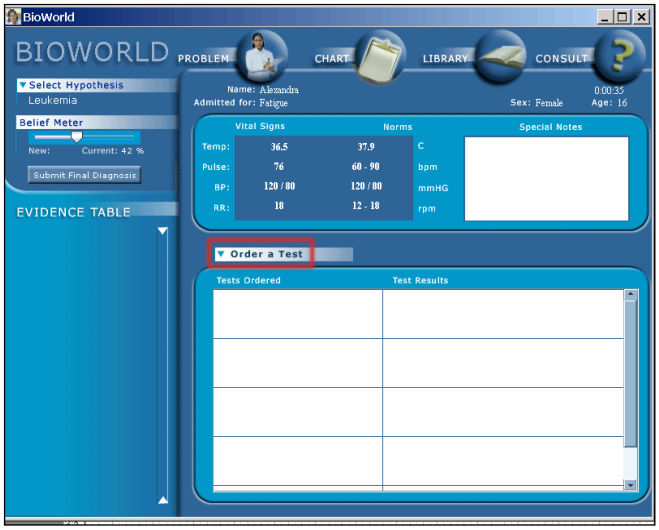
Problem

As you formulate your hypothesis please indicate which evidence supports your hypothesis by highlighting the text and clicking on "*Send to Evidence*" button at the bottom left of the screen. (note that if you select an entire sentence it will only count as one piece of evidence)



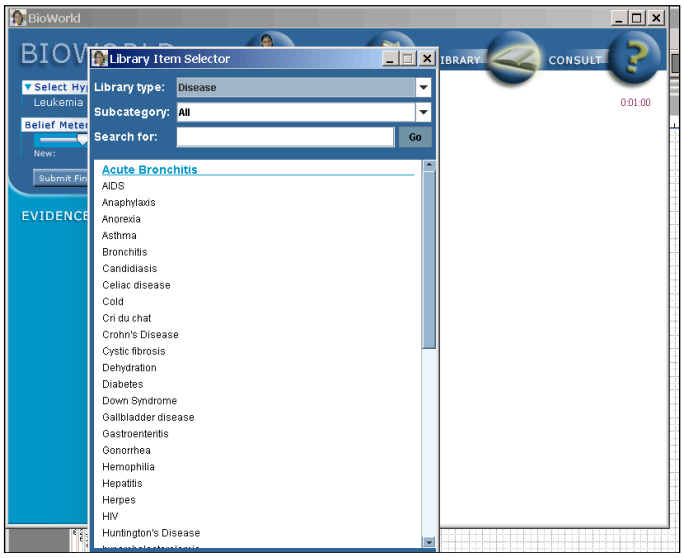
Chart

In this space you can look at the vitals signs (but cannot send it to the evidence table) and order diagnostic test by clicking on the "*Order a Test*" button in the middle of the screen. Every test ordered will be automatically added to the Evidence Table.



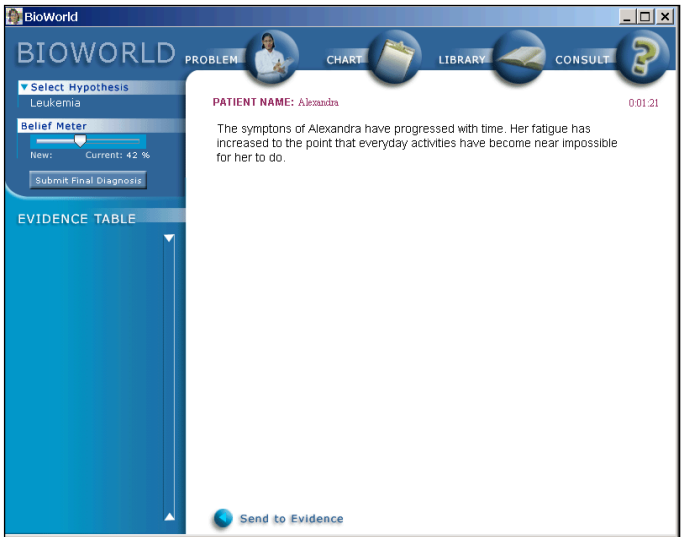
Library

You can search the library for information by *disease*, *diagnostic test* or by *glossary term*. Note that for disease and diagnostic test you can use the subcategories to narrow your search.



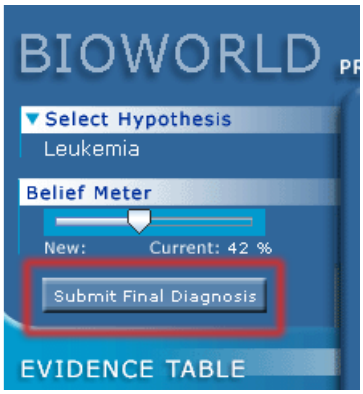
Consult

If you need a hint or some help because you are lost and the problem is too difficult for you click on consult. Consult is contextualized for each space. This means that if you are in Chart and click on consult you will get a hint that is specific to your exploration of the chart for this case. Note that there can be up to three levels of hints for each space (problem statement, chart and library) so if the hint is not helpful, click again and another one will appear.



Submit Final Diagnosis

Once you have enough evidence and are confident enough about your diagnostic, revise your confidence meter and click on submit final diagnostic

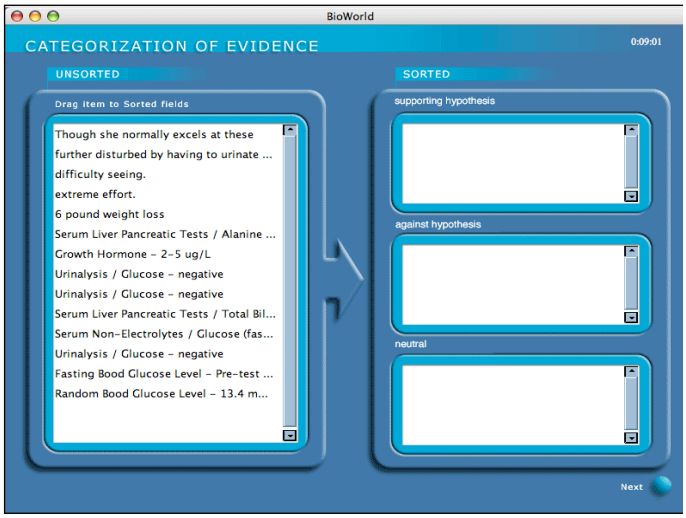


Phase 2 - Categorization and prioritization of your evidence

Once you have submitted your final diagnosis you are asked to categorize and then prioritize your evidence.

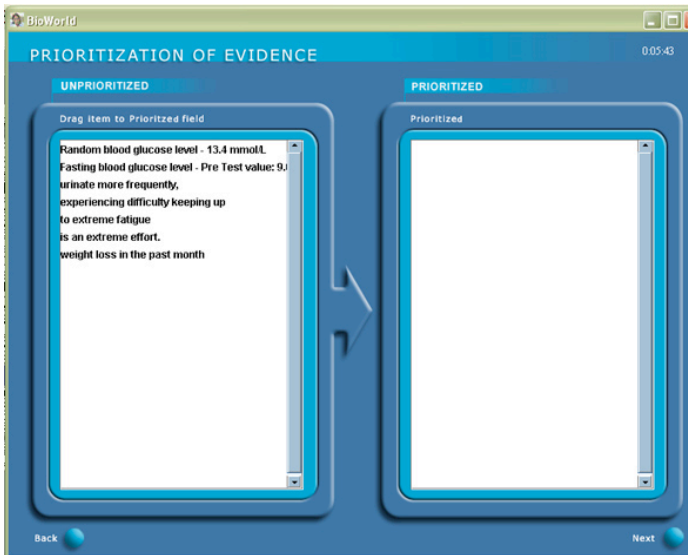
Categorization of Evidence

Drag and drop the evidence according to whether they support your main hypothesis, go against your main hypothesis or are neutral (they do not support this main hypothesis but do not go against it either)



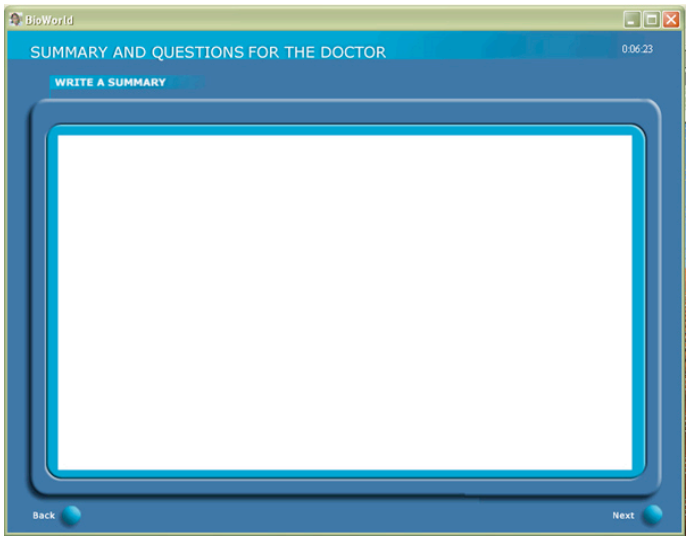
Prioritization of Evidence

Order the evidence according to their relative importance related to your final diagnostic. You can drag and drop from the left to the right but you can also move them around once they are on the right side of the screen.



Phase 3 - Articulating and explaining your solution

You are asked to write a brief summary explaining how the evidence you collected supported your diagnosis (maximum of 3 or 4 sentences).



APPENDIX C

List of Absolutely Necessary Categorized Evidence for Case 1

E1	E2	E3	E4	E5
Urinary Catecholamines / Norepinephrine	headaches, palpitations, sweating and flushing	can the new blood medication be causing the symptoms?	taking medicaiton for high blood pressure	37 yr old
Urinary Catecholamines/Total (Epinephrine + Norepinephrine)	makes me think of secondary causes of hypertension	symptoms are coming as episodes	frequent headaches	female
Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr	but pheochromocytoma is rare so you need to keep other causes in mind	linking palpitations, profuse sweating and flushing	extremely anxious with palpitations, sweating, and flushing	episodes
	high blood pressure	new main hypothesis: Pheocromocytoma	hypertensive	
	frequent headaches	test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L	planning about checking blood pressure and doing physical exam	
	periods of time during which she feels "extremely anxious" with palpitations, profuse sweating, and flushing	test: Urinary Catecholamines / Free - 1560 mmol / 24 hr	If teaching, need to tell them to rule out secondary cause of hypertension	
	Hypotheses 1,2 - Grave's disease and pheocromocytoma	submitting Hypothesis Pheochromocytoma - 65%	test: Urinary Catecholamines / Dopamine - 340-3134 nmol/day	
	Hypothesis 3,4 - Depression and essential hypertension with reaction to medication	alternative endocrinology diagnosis to keep in mind and screen for: cortisol increase, adrenal hormones	test: Urinary Catecholamines / Epinephrine - 11-131 nmol/day	
	Hypothesis 5 - Drug abuse but pheo is up on the list	test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr	test: Urinary Catecholamines / Free - 1560 mmol / 24 hr	
	vital signs: pulse of 98 per min, high blood pressure and slightly elevated temperature 37.9		test: Urinary Catecholamines / Norepinephrine - 89-591 nmol/day	
	select Hypothesis Pheochromocytoma		test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr	
	test: Thyroid Stimulating Hormone (TSH) - normal		add imaging test: CT abdomen	
	test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr test: Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr test: Urinary Catecholamines / Free - 1560 mmol / 24 hr		reviewing results - cathecolmine total high !!!	
	set Hypothesis strength to 100%			
	submitting Hypothesis Pheochromocytoma - 100%			

APPENDIX D

Case 1 Categorized Episodes for Expert 1

E1	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	Urinary Catecholamines / Norepinephrine	high blood pressure	37 yrs old
2	Urinary Catecholamines/Total (Epinephrine + Norepinephrine)	extremely anxious	not a new problem
3	Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr	palpitation, profuse sweating, and flushing	Fasting Blood Glucose Level - normal
4		more frequent in the past little while	Serum Electrolytes / Anion Gap (Na-(Cl+HCO3))
5		weight loss	Serum Electrolytes / Magnesium (Mg)
6		Ultrasound / Abdominal Scan	Serum Liver Pancreatic Tests / Alanine Aminotransferase (ALT)
7		CT / Body	Aldosterone
8			Adrenocorticotropin hormone (ACTH)
9			Cortisol
10			Dehydroepiandosterone Sulfate (DHEA-S)

Case 1 Categorized Episodes for Expert 2

E2	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	headache, palpitations, sweating and flushing; makes me think of secondary causes of hypertension	medication	37 yrs old
2	pheochromocytoma is rare so you need to keep other causes in mind	10 pounds in the last 4 months and (evidence)	Dizzy
3	high blood pressure	checking toxicology tests	eye exam test
4	frequent headaches		
5	periods of time during which she feels "extremely anxious" with palpitation, profuse sweating, and flushing.		
6	hypothesis 1: Grave's disease		
7	hypothesis 2: pheochrocytoma		
8	hypothesis 3: essential hypertension with reaction to medication		
9	hypothesis 4: drug abuse		
10	pulse of 98 a minute		
11	one of the 3 followint tests: a)Urinary catecholamines, b) Urinary Metabolites VMA, c) Urinary catecholamines		
12	submit hypothesis pheochromocytoma with high belief		

Case 1 Categorized Episodes for Expert 3

E3	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	can the new blood medication be causing the symptoms?	37 yr old woman	tests: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
2	cluster of episodes of palpitations, profuse sweating and flushing	lost 10 pounds in 4 months	and Triiodothyronine (T3) - 0.92-2.78 nmol/L
3	alternative endocrinology diagnosis	set hypothesis of hypertyroid (Grave's disease)	
4	to keep in mind and screen for:	reviewing evidence and hypothesis, planning for CT of abdomen	
5	cortisol increase, adrenal hormones		
6	hypertensive 180/103		
7	test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L		
8	test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr		
9	main hypothesis: Pheocromocytoma		
10	test: Urinary Catecholamines / Free - 1560 mmol / 24 hr		

Case 1 Categorized Episodes for Expert 4

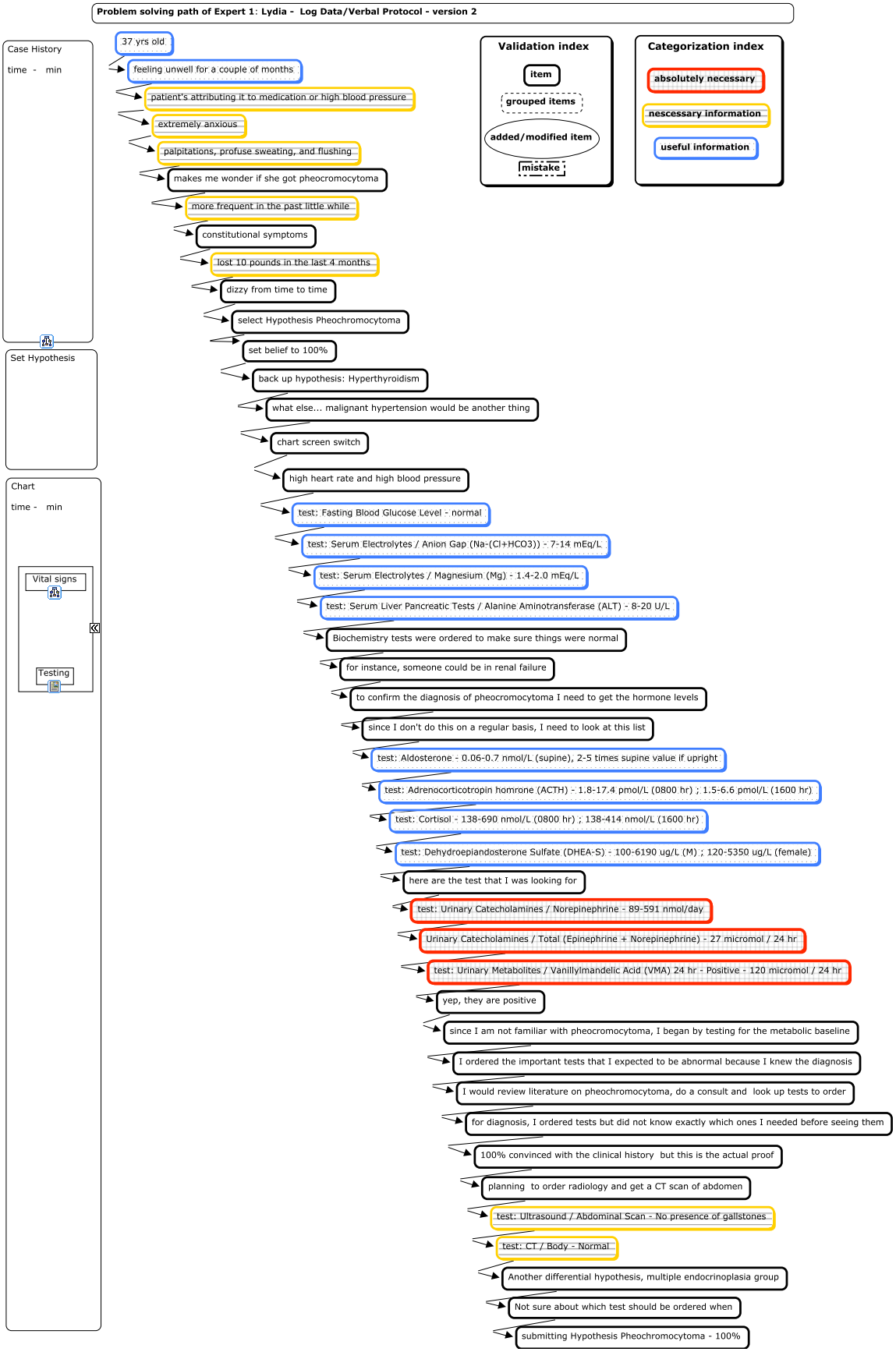
E4	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	taking medicaiton for high blood pressure	select Hypothesis Pheochromocytoma	symptoms more frequent
2	frequent headaches	tachycardic	lost 10 pounds and dizzy
3	extremely anxious with palpitations, sweating, and flushing	test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L	test: Plain ECG - Sinus rhythm with nonspecific ST-T wave changes, left bundle branch block
4	hypertensive	test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 8 - 25 mg/dL	test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
5	planning about checking blood pressure and doing physical exam	Alternative hypothesis: drug abuse?	alternative hypothesis: Essential hypertension
6	If teaching, need to tell them to rule out secondary cause of hypertension	Alternative hypothesis: related to medication?	
7	test: Urinary Catecholamines / Dopamine - 340-3134 nmol/day		
8	test: Urinary Catecholamines / Epinephrine - 11-131 nmol/day		
9	test: Urinary Catecholamines / Free - 1560 mmol / 24 hr		
10	test: Urinary Catecholamines / Norepinephrine - 89-591 nmol/day		
11	test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr		
12	add imaging test: CT abdomen		
13	reviewing results - cathecolmine total high !!!		

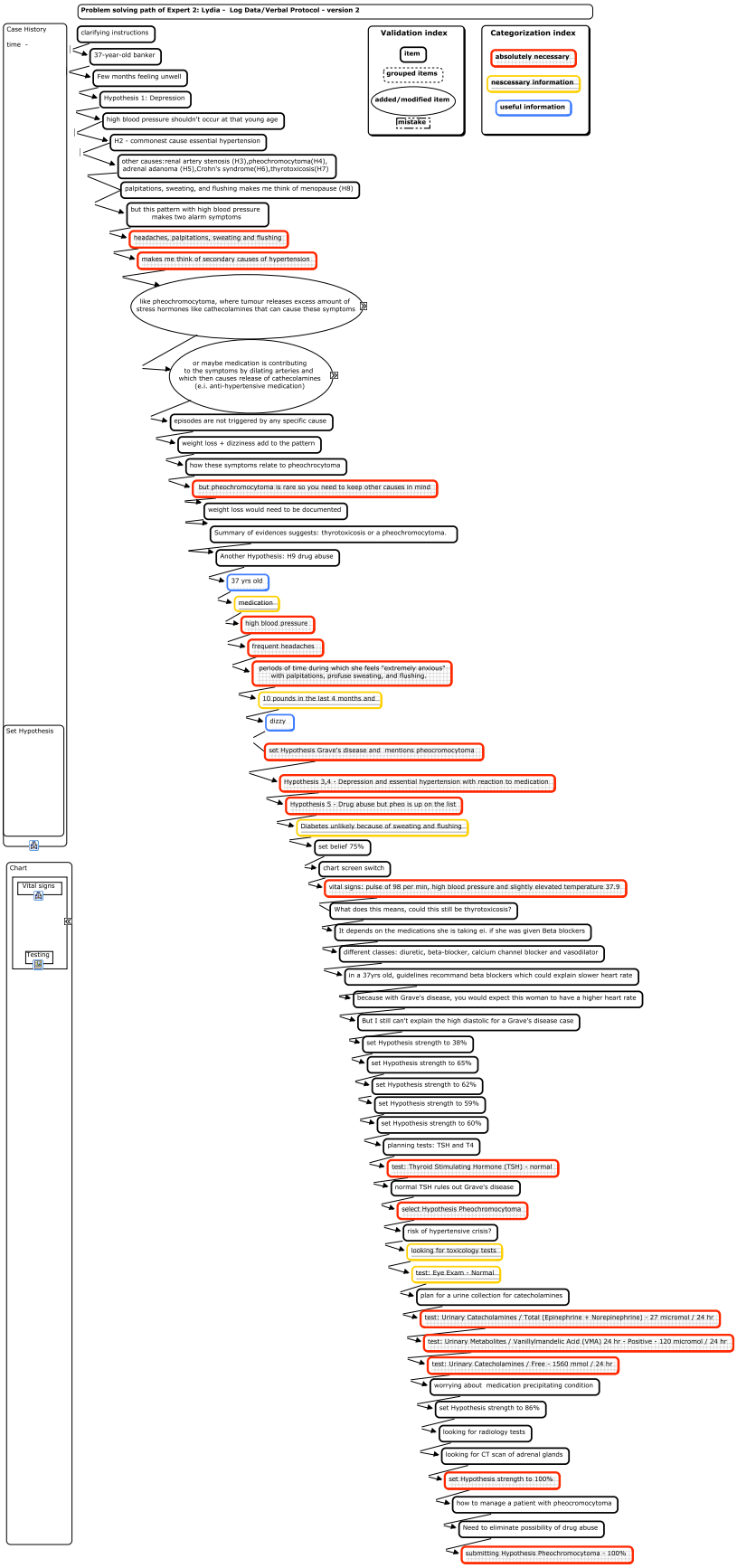
Case 1 Categorized Episodes for Expert 5

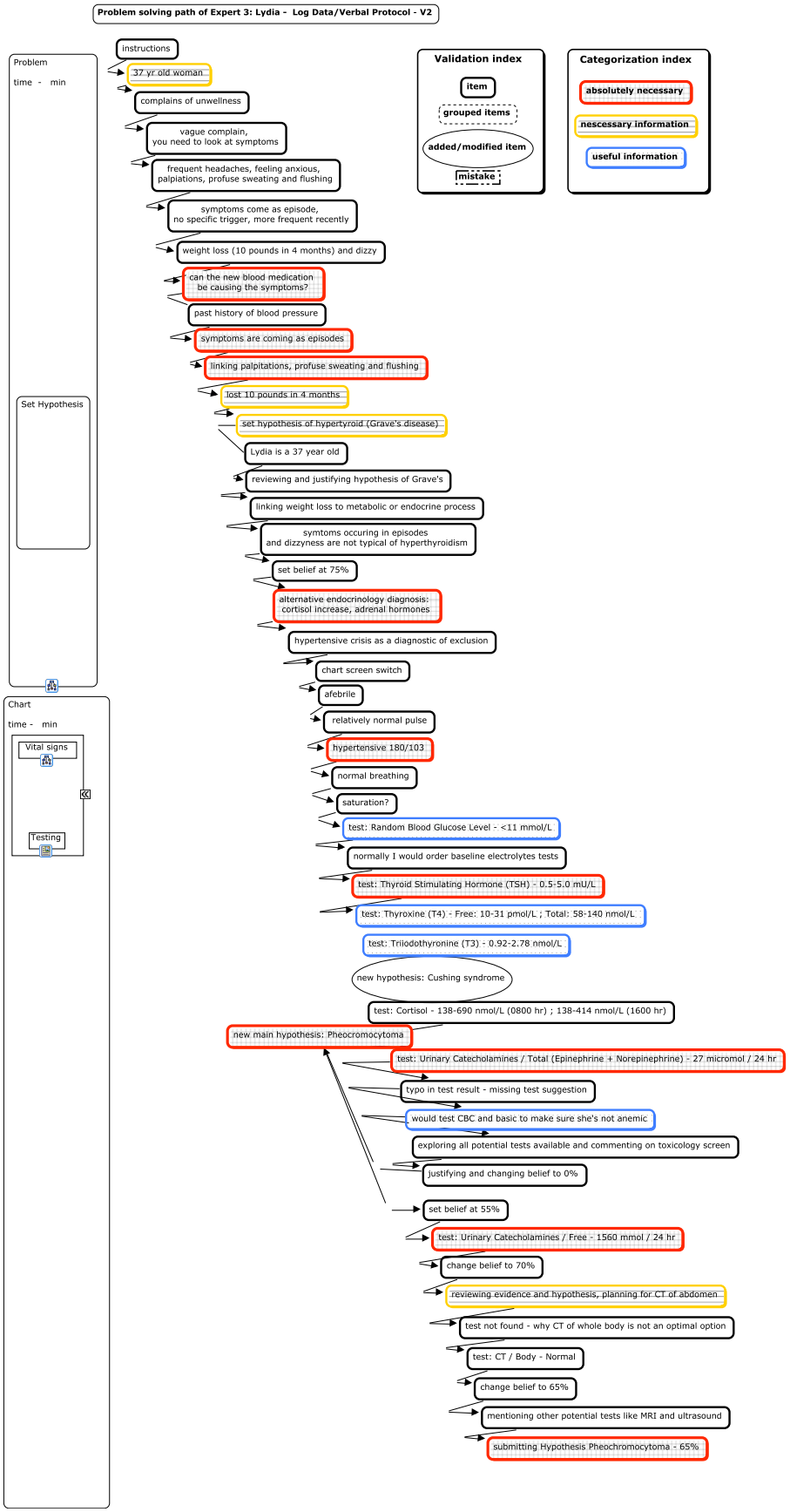
E5	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	37 yr old	feeling unwell for a few months	weight loss
2	female	problem may be related to medication	test: Triiodothyronine (T3) - 0.92-2.78 nmol/L
3	episodes	high blood pressure	ordinarily you do routine laboratory tests
4		headache is a non specific complaint	test: Fasting Blood Glucose Level
5		extreme anxiousness, palpitations, profuse sweating and flushing	test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 7-14 mEq/L
6		hypothesis anxiety	test: Serum Electrolytes / Bicarbonate (HCO3) - 22-26 mEq/L
7		Hypothesis of pheocromocytoma	test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
8		linking weight loss with anxiety and hyperthyroidism	test: WBC (total) - 4.5-11.0 X 10E9 /L
9		dizzyness is non-specific	
10		selecting hypotheses: hypertryroid, pheochromocytoma, depression, panic attack	
11		requesting hypothesis non available in menu: Anxiety	
12		selecting hyperthyroid as main hypothesis with 10% belief	
13		elevated blood pressure and pulse rate	
14		linking vital signs to hypotheses	
15		test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L	
16		test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L	
17		considers Pheochromocytoma and mentions related tests	
18		presumably the serum potassium is normal	
19		selecting hypothesis Panic Attack with 50 belief	

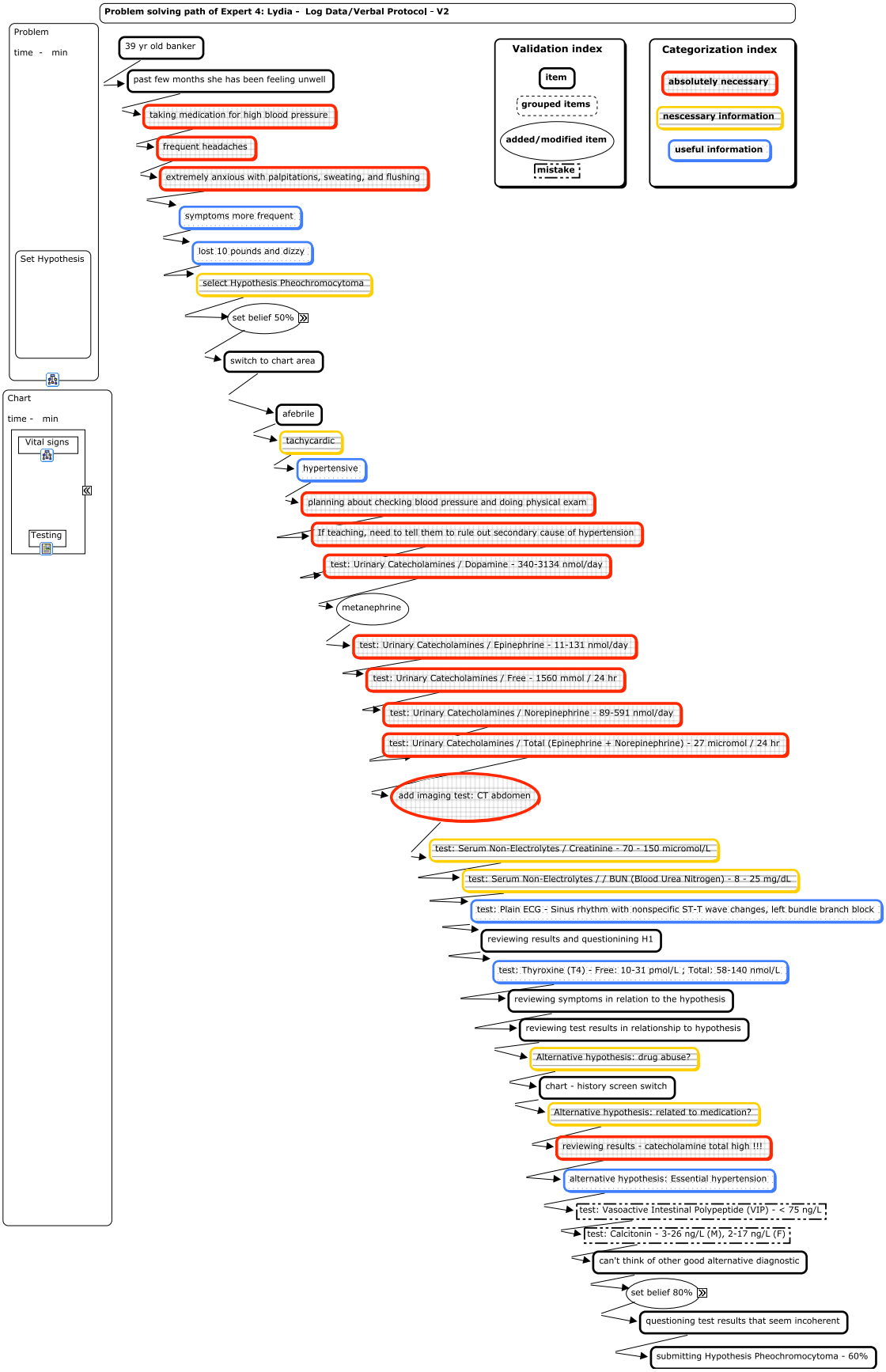
APPENDIX E

Individual Categorized Visual Representations for Case 1

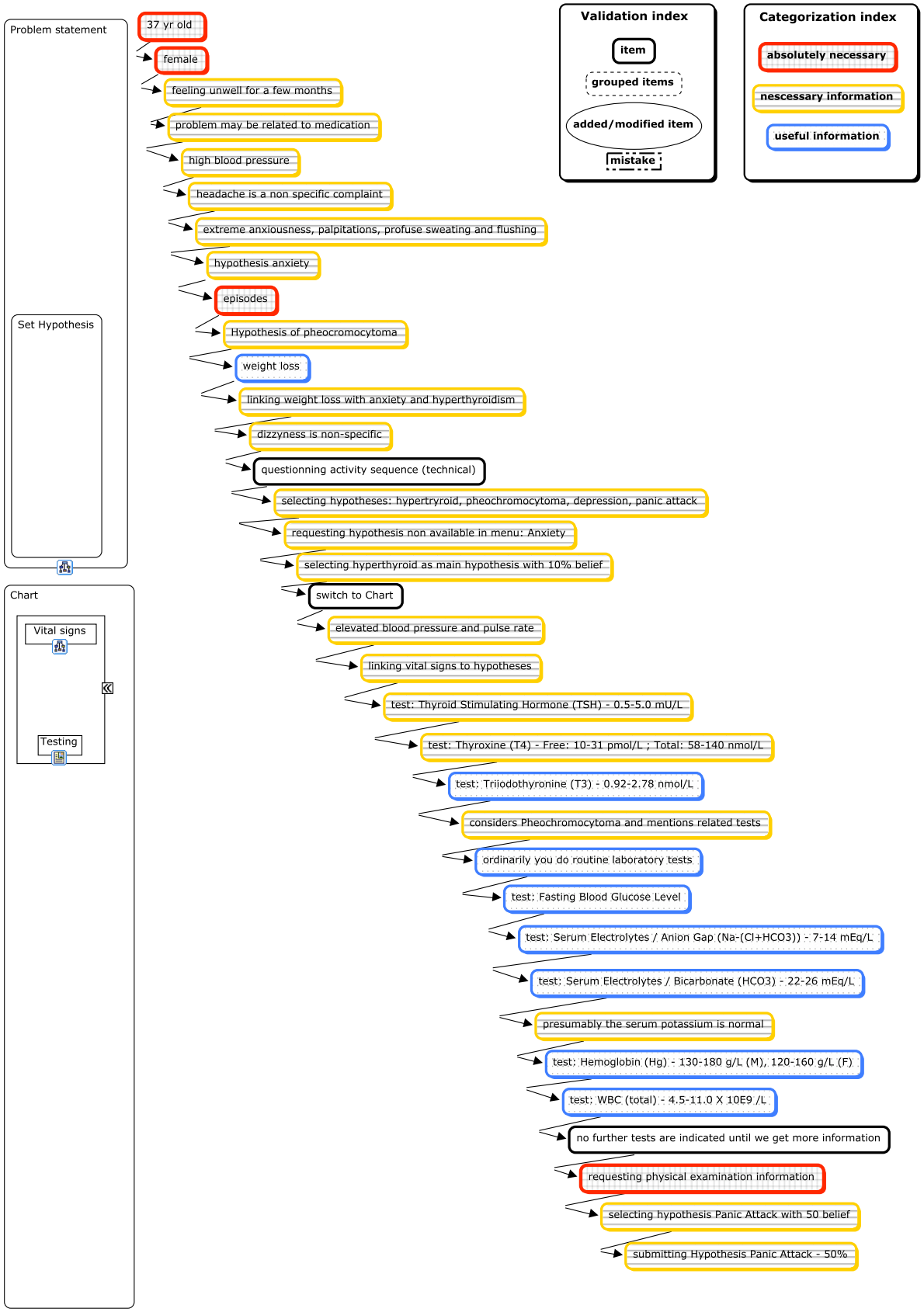

Problem solving path of Expert 1: Lydia - Log Data/Verbal Protocol - version 2
Case History
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
37 yrs old
feeling unwell for a couple of months
patient's attributing it to medication or high blood pressure
extremely anxious
palpitations, profuse sweating, and flushing
makes me wonder if she got pheocromocytoma
more frequent in the past little while
constitutional symptoms
lost 10 pounds in the last 4 months
dizzy from time to time
select Hypothesis Pheochromocytoma
set belief to 100%
back up hypothesis: Hyperthyroidism
what else... malignant hypertension would be another thing
chart screen switch
high heart rate and high blood pressure
test: Fasting Blood Glucose Level - normal
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 7-14 mEq/L
test: Serum Electrolytes / Magnesium (Mg) - 1.4-2.0 mEq/L
test: Serum Liver Pancreatic Tests / Alanine Aminotransferase (ALT) - 8-20 U/L
Biochemistry tests were ordered to make sure things were normal
for instance, someone could be in renal failure
to confirm the diagnosis of pheocromocytoma I need to get the hormone levels
since I don't do this on a regular basis, I need to look at this list
test: Aldosterone - 0.06-0.7 nmol/L (supine), 2-5 times supine value if upright
test: Adrenocorticotropin homrone (ACTH) - 1.8-17.4 pmol/L (0800 hr) ; 1.5-6.6 pmol/L (1600 hr)
test: Cortisol - 138-690 nmol/L (0800 hr) ; 138-414 nmol/L (1600 hr)
test: Dehydroepiandosterone Sulfate (DHEA-S) - 100-6190 ug/L (M) ; 120-5350 ug/L (female)
here are the test that I was looking for
test: Urinary Catecholamines / Norepinephrine - 89-591 nmol/day
Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
test: Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr
yep, they are positive
since I am not familiar with pheocromocytoma, I began by testing for the metabolic baseline
I ordered the important tests that I expected to be abnormal because I knew the diagnosis
I would review literature on pheochromocytoma, do a consult and look up tests to order
for diagnosis, I ordered tests but did not know exactly which ones I needed before seeing them
100% convinced with the clinical history but this is the actual proof
planning to order radiology and get a CT scan of abdomen
test: Ultrasound / Abdominal Scan - No presence of gallstones
test: CT / Body - Normal
Another differential hypothesis, multiple endocrinoplasia group
Not sure about which test should be ordered when
submitting Hypothesis Pheochromocytoma - 100%
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information


Problem solving path of Expert 2: Lydia - Log Data/Verbal Protocol - version 2
Case History
time -
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
clarifying instructions
37-year-old banker
Few months feeling unwell
Hypothesis 1: Depression
high blood pressure shouldn't occur at that young age
H2 - commonest cause essential hypertension
other causes:renal artery stenosis (H3),pheochromocytoma(H4), adrenal adanoma (H5),Crohn's syndrome(H6),thyrotoxicosis(H7)
palpitations, sweating, and flushing makes me think of menopause (H8)
but this pattern with high blood pressure makes two alarm symptoms
headaches, palpitations, sweating and flushing
makes me think of secondary causes of hypertension
like pheochromocytoma, where tumour releases excess amount of stress hormones like cathecolamines that can cause these symptoms
or maybe medication is contributing to the symptoms by dilating arteries and which then causes release of cathecolamines (e.i. anti-hypertensive medication)
episodes are not triggered by any specific cause
weight loss + dizziness add to the pattern
how these symptoms relate to pheochrocytoma
but pheochromocytoma is rare so you need to keep other causes in mind
weight loss would need to be documented
Summary of evidences suggests: thyrotoxicosis or a pheochromocytoma.
Another Hypothesis: H9 drug abuse
37 yrs old
medication
high blood pressure
frequent headaches
periods of time during which she feels "extremely anxious" with palpitations, profuse sweating, and flushing.
10 pounds in the last 4 months and
dizzy
set Hypothesis Grave's disease and mentions pheocromocytoma
Hypothesis 3,4 - Depression and essential hypertension with reaction to medication
Hypothesis 5 - Drug abuse but pheo is up on the list
Diabetes unlikely because of sweating and flushing
set belief 75%
chart screen switch
vital signs: pulse of 98 per min, high blood pressure and slightly elevated temperature 37.9
What does this means, could this still be thyrotoxicosis?
It depends on the medications she is taking ei. if she was given Beta blockers
different classes: diuretic, beta-blocker, calcium channel blocker and vasodilator
in a 37yrs old, guidelines recommand beta blockers which could explain slower heart rate
because with Grave's disease, you would expect this woman to have a higher heart rate
But I still can't explain the high diastolic for a Grave's disease case
set Hypothesis strength to 38%
set Hypothesis strength to 65%
set Hypothesis strength to 62%
set Hypothesis strength to 59%
set Hypothesis strength to 60%
planning tests: TSH and T4
test: Thyroid Stimulating Hormone (TSH) - normal
normal TSH rules out Grave's disease
select Hypothesis Pheochromocytoma
risk of hypertensive crisis?
looking for toxicology tests
test: Eye Exam - Normal
plan for a urine collection for catecholamines
test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
test: Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr
test: Urinary Catecholamines / Free - 1560 mmol / 24 hr
worrying about medication precipitating condition
set Hypothesis strength to 86%
looking for radiology tests
looking for CT scan of adrenal glands
set Hypothesis strength to 100%
how to manage a patient with pheocromocytoma
Need to eliminate possibility of drug abuse
submitting Hypothesis Pheochromocytoma - 100%
Set Hypothesis
Chart
Vital signs
Testing


Problem solving path of Expert 3: Lydia - Log Data/Verbal Protocol - V2
Problem
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
instructions
37 yr old woman
complains of unwellness
vague complain, you need to look at symptoms
frequent headaches, feeling anxious, palpiations, profuse sweating and flushing
symptoms come as episode, no specific trigger, more frequent recently
weight loss (10 pounds in 4 months) and dizzy
can the new blood medication be causing the symptoms?
past history of blood pressure
symptoms are coming as episodes
linking palpitations, profuse sweating and flushing
lost 10 pounds in 4 months
set hypothesis of hypertyroid (Grave's disease)
Lydia is a 37 year old
reviewing and justifying hypothesis of Grave's
linking weight loss to metabolic or endocrine process
symtoms occuring in episodes and dizzyness are not typical of hyperthyroidism
set belief at 75%
alternative endocrinology diagnosis: cortisol increase, adrenal hormones
hypertensive crisis as a diagnostic of exclusion
chart screen switch
afebrile
relatively normal pulse
hypertensive 180/103
normal breathing
saturation?
test: Random Blood Glucose Level - <11 mmol/L
normally I would order baseline electrolytes tests
test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L
test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
test: Triiodothyronine (T3) - 0.92-2.78 nmol/L
new hypothesis: Cushing syndrome
test: Cortisol - 138-690 nmol/L (0800 hr) ; 138-414 nmol/L (1600 hr)
new main hypothesis: Pheocromocytoma
test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
typo in test result - missing test suggestion
would test CBC and basic to make sure she's not anemic
exploring all potential tests available and commenting on toxicology screen
justifying and changing belief to 0%
set belief at 55%
test: Urinary Catecholamines / Free - 1560 mmol / 24 hr
change belief to 70%
reviewing evidence and hypothesis, planning for CT of abdomen
test not found - why CT of whole body is not an optimal option
test: CT / Body - Normal
change belief to 65%
mentioning other potential tests like MRI and ultrasound
submitting Hypothesis Pheochromocytoma - 65%

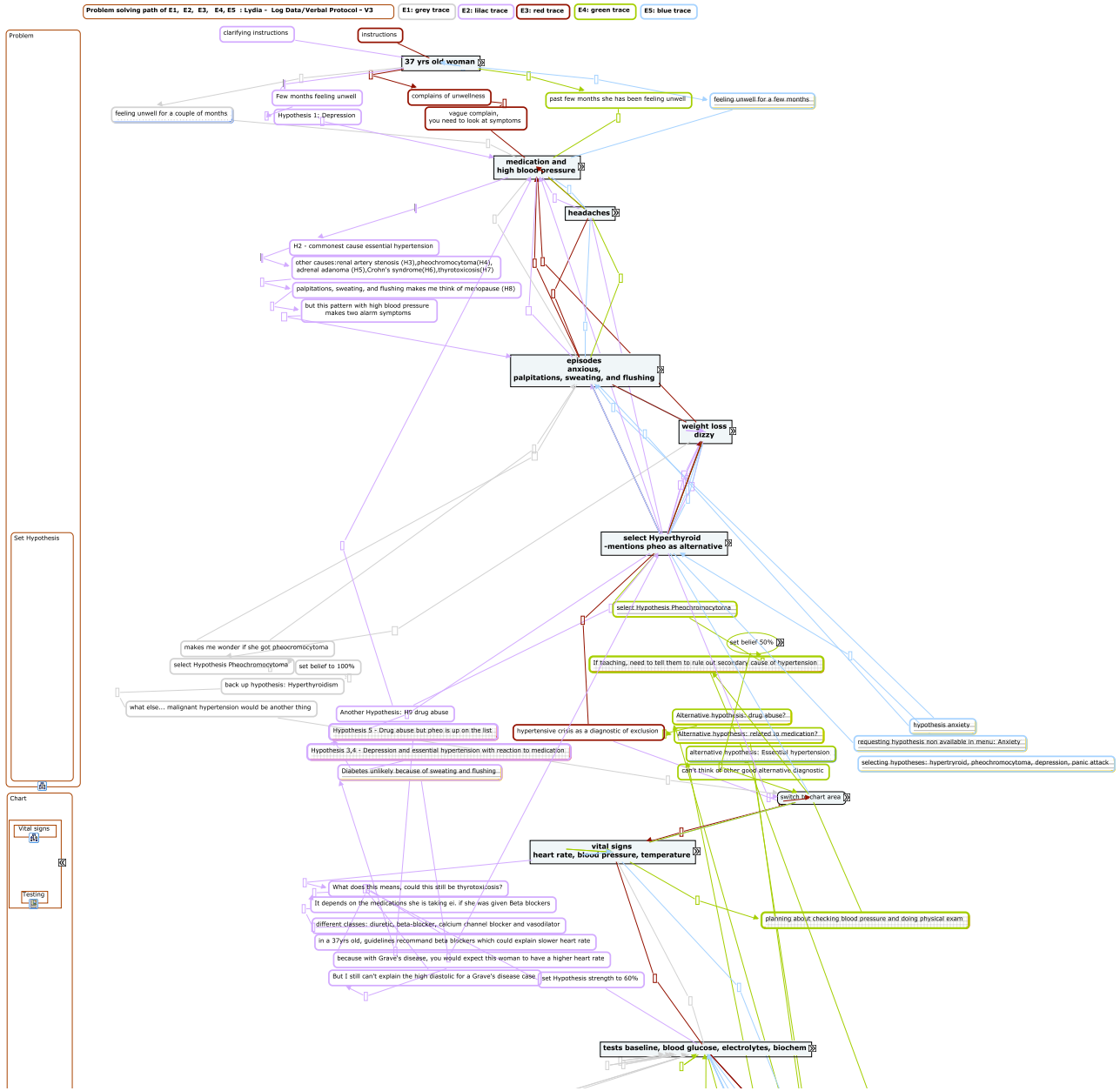

Problem solving path of Expert 4: Lydia - Log Data/Verbal Protocol - V2
Problem
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
39 yr old banker
past few months she has been feeling unwell
taking medication for high blood pressure
frequent headaches
extremely anxious with palpitations, sweating, and flushing
symptoms more frequent
lost 10 pounds and dizzy
select Hypothesis Pheochromocytoma
set belief 50%
switch to chart area
afebrile
tachycardic
hypertensive
planning about checking blood pressure and doing physical exam
If teaching, need to tell them to rule out secondary cause of hypertension
test: Urinary Catecholamines / Dopamine - 340-3134 nmol/day
metanephrine
test: Urinary Catecholamines / Epinephrine - 11-131 nmol/day
test: Urinary Catecholamines / Free - 1560 mmol / 24 hr
test: Urinary Catecholamines / Norepinephrine - 89-591 nmol/day
test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
add imaging test: CT abdomen
test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 8 - 25 mg/dL
test: Plain ECG - Sinus rhythm with nonspecific ST-T wave changes, left bundle branch block
reviewing results and questionining H1
test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
reviewing symptoms in relation to the hypothesis
reviewing test results in relationship to hypothesis
Alternative hypothesis: drug abuse?
chart - history screen switch
Alternative hypothesis: related to medication?
reviewing results - catecholamine total high !!!
alternative hypothesis: Essential hypertension
test: Vasoactive Intestinal Polypeptide (VIP) - < 75 ng/L
test: Calcitonin - 3-26 ng/L (M), 2-17 ng/L (F)
can't think of other good alternative diagnostic
set belief 80%
questioning test results that seem incoherent
submitting Hypothesis Pheochromocytoma - 60%

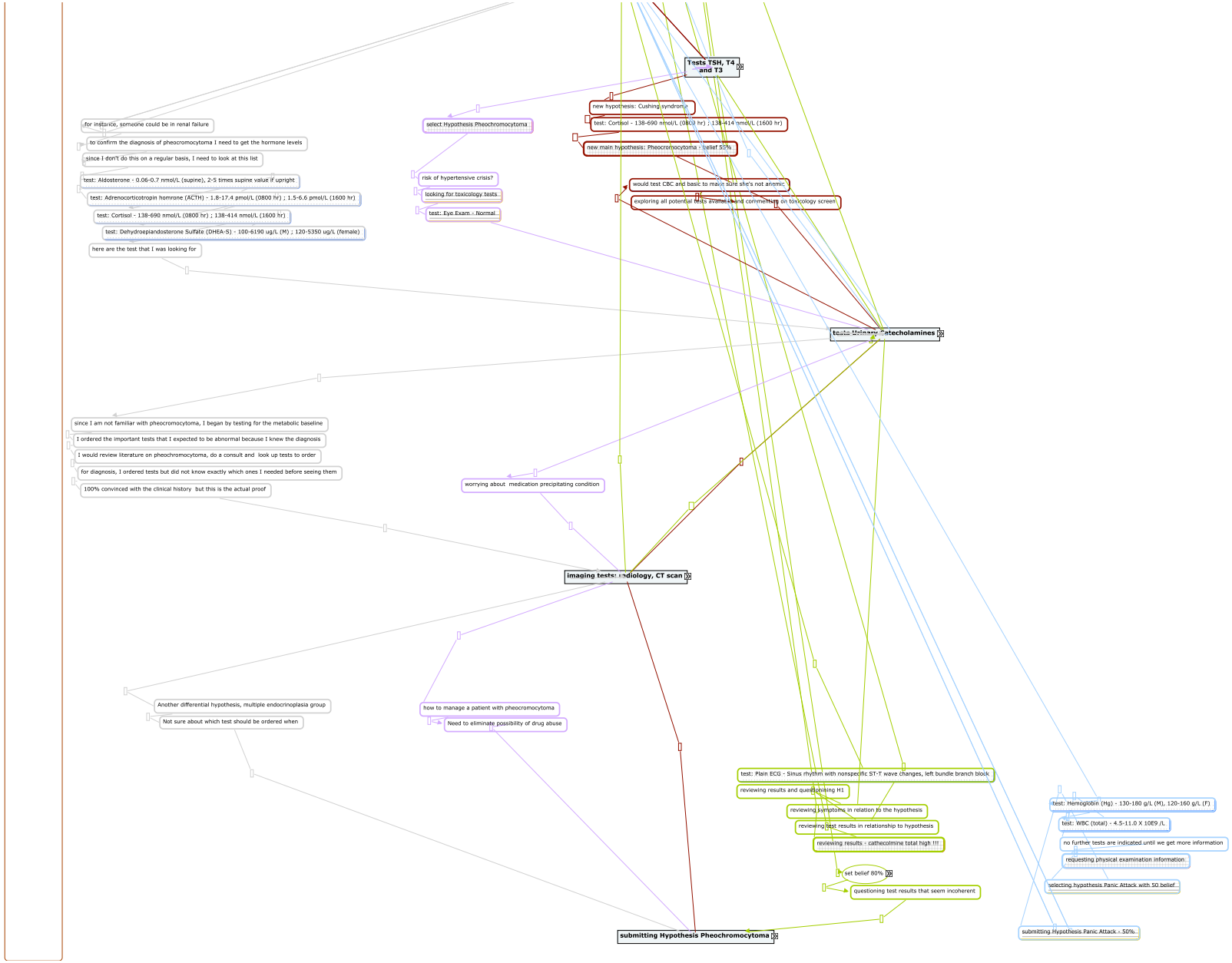
Problem solving path of Expert 5: Lydia - Log Data/Verbal Protocol - V2

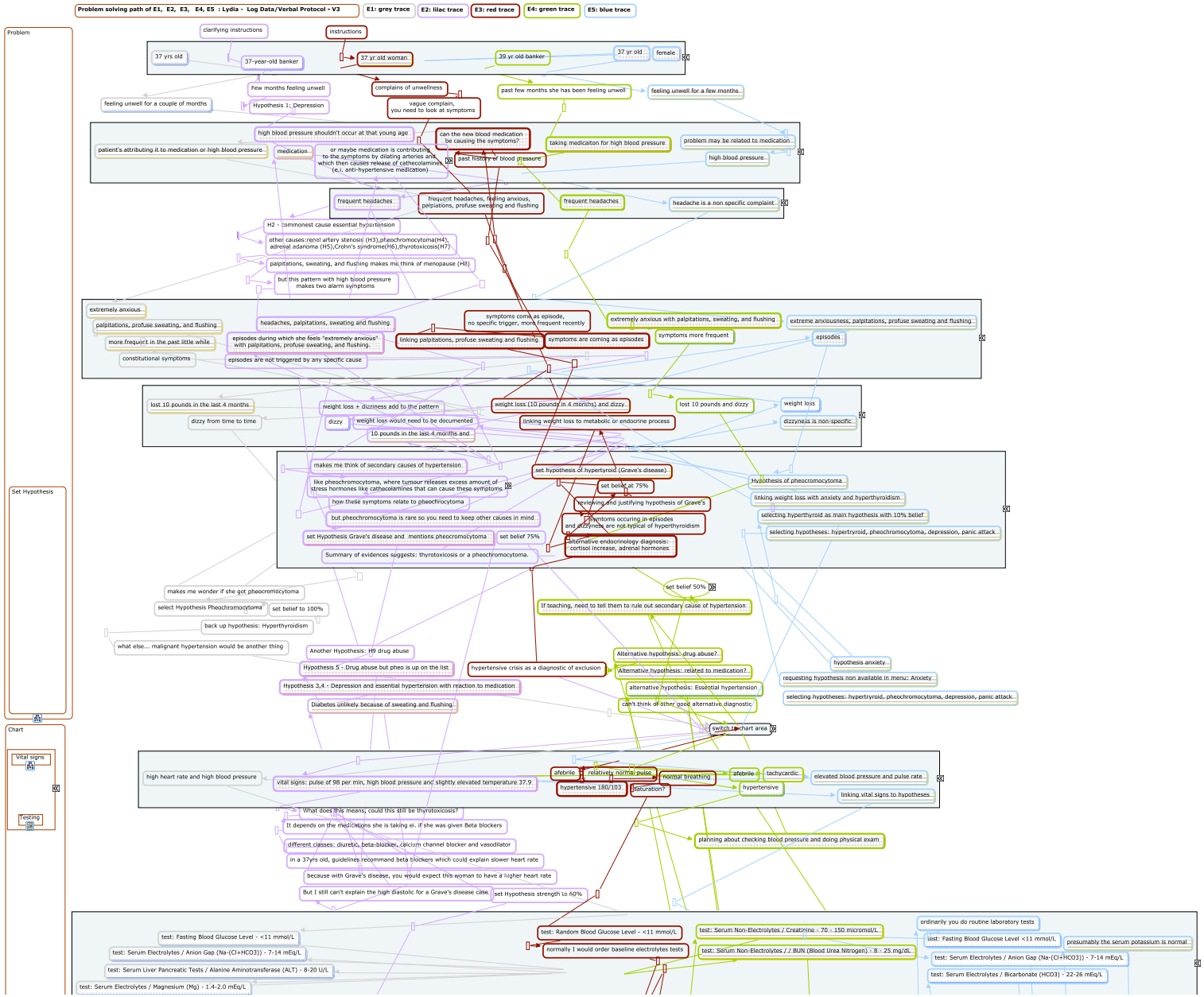


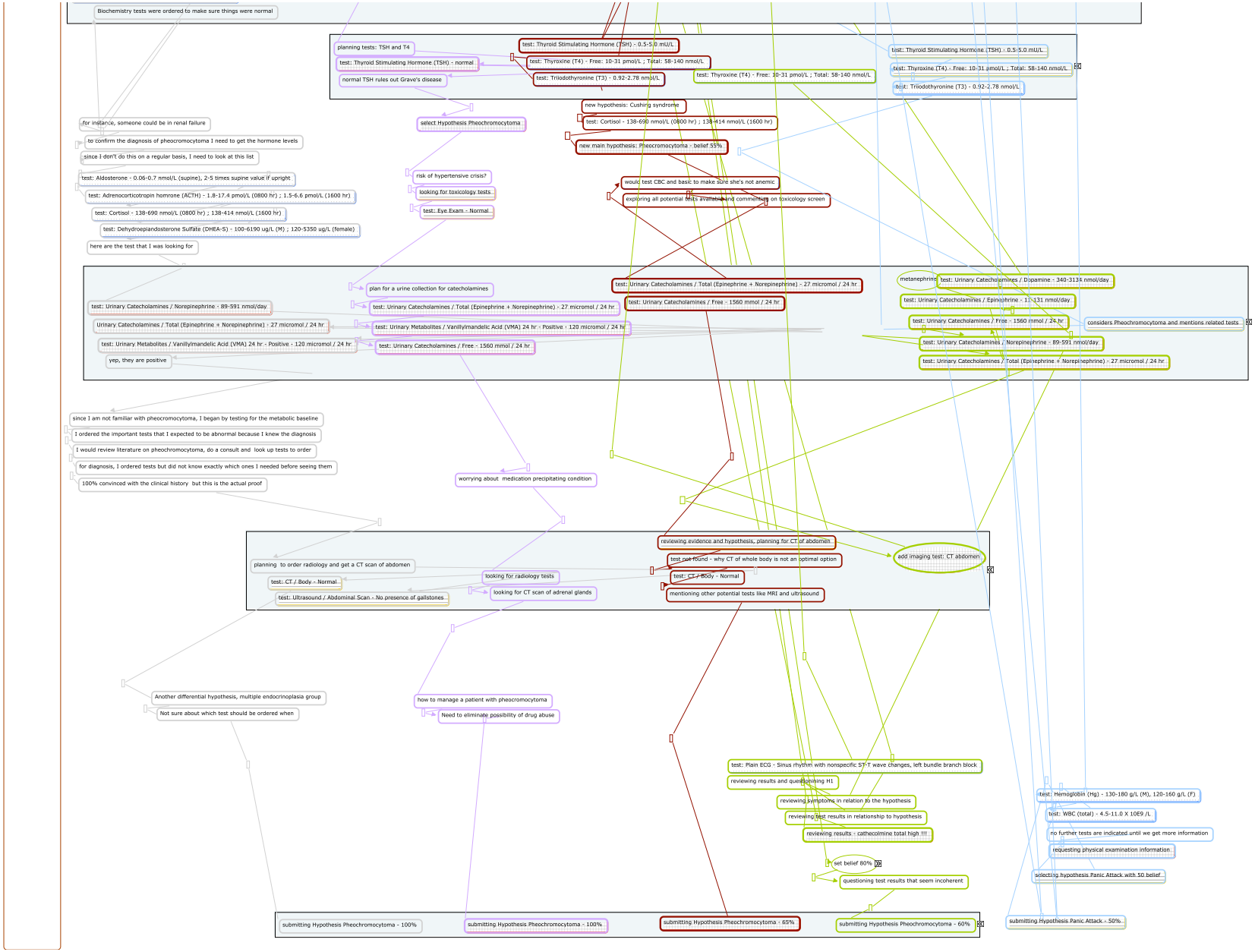
APPENDIX F

Merged Representations for Case 1 at Level 1 and Level 2


Problem solving path of E1, E2, E3, E4, E5 : Lydia - Log Data/Verbal Protocol - V3
E1: grey trace
E2: lilac trace
E3: red trace
E4: green trace
E5: blue trace
Problem
Set Hypothesis
Chart
Vital signs
Testing
clarifying instructions
instructions
37 yrs old woman
Few months feeling unwell
complains of unwellness
past few months she has been feeling unwell
feeling unwell for a few months
feeling unwell for a couple of months
Hypothesis 1: Depression
vague complain, you need to look at symptoms
medication and high blood pressure
headaches
H2 - commonest cause essential hypertension
other causes:renal artery stenosis (H3),pheochromocytoma(H4), adrenal adenoma (H5),Crohn's syndrome(H6),thyrotoxicosis(H7)
palpitations, sweating, and flushing makes me think of menopause (H8)
but this pattern with high blood pressure makes two alarm symptoms
episodes anxious, palpitations, sweating, and flushing
weight loss dizzy
select Hyperthyroid -mentions pheo as alternative
select Hypothesis Pheochromocytoma
set belief 50%
If teaching, need to tell them to rule out secondary cause of hypertension
makes me wonder if she got pheocromocytoma
select Hypothesis Pheochromocytoma
set belief to 100%
back up hypothesis: Hyperthyroidism
what else... malignant hypertension would be another thing
Another Hypothesis: H9 drug abuse
Hypothesis 5 - Drug abuse but pheo is up on the list
hypertensive crisis as a diagnostic of exclusion
Alternative hypothesis: drug abuse?
Alternative hypothesis: related to medication?
hypothesis anxiety
requesting hypothesis non available in menu: Anxiety
Hypothesis 3,4 - Depression and essential hypertension with reaction to medication
alternative hypothesis: Essential hypertension
selecting hypotheses: hypertryroid, pheochromocytoma, depression, panic attack
Diabetes unlikely because of sweating and flushing
can't think of other good alternative diagnostic
switch to chart area
vital signs heart rate, blood pressure, temperature
What does this means, could this still be thyrotoxicosis?
It depends on the medications she is taking ei. if she was given Beta blockers
different classes: diuretic, beta-blocker, calcium channel blocker and vasodilator
in a 37yrs old, guidelines recommand beta blockers which could explain slower heart rate
because with Grave's disease, you would expect this woman to have a higher heart rate
But I still can't explain the high diastolic for a Grave's disease case
set Hypothesis strength to 60%
planning about checking blood pressure and doing physical exam
tests baseline, blood glucose, electrolytes, biochem


Tests TSH, T4 and T3
new hypothesis: Cushing syndrom
test: Cortisol - 138-690 nmol/L (0800 hr) ; 138-414 nmol/L (1600 hr)
new main hypothesis: Pheocromocytoma - belief 55%
select Hypothesis Pheochromocytoma
risk of hypertensive crisis?
looking for toxicology tests
test: Eye Exam - Normal
would test CBC and basic to make sure she's not anemic
exploring all potential tests availble and commentum on toxicology screen
for instance, someone could be in renal failure
to confirm the diagnosis of pheocromocytoma I need to get the hormone levels
since I don't do this on a regular basis, I need to look at this list
test: Aldosterone - 0.06-0.7 nmol/L (supine), 2-5 times supine value if upright
test: Adrenocorticotropin hormone (ACTH) - 1.8-17.4 pmol/L (0800 hr) ; 1.5-6.6 pmol/L (1600 hr)
test: Cortisol - 138-690 nmol/L (0800 hr) ; 138-414 nmol/L (1600 hr)
test: Dehydroepiandosterone Sulfate (DHEA-S) - 100-6190 ug/L (M) ; 120-5350 ug/L (female)
here are the test that I was looking for
tests Urinary Catecholamines
since I am not familiar with pheocromocytoma, I began by testing for the metabolic baseline
I ordered the important tests that I expected to be abnormal because I knew the diagnosis
I would review literature on pheochromocytoma, do a consult and look up tests to order
for diagnosis, I ordered tests but did not know exactly which ones I needed before seeing them
100% convinced with the clinical history but this is the actual proof
worrying about medication precipitating condition
imaging tests: radiology, CT scan
Another differential hypothesis, multiple endocrinoplasia group
Not sure about which test should be ordered when
how to manage a patient with pheocromocytoma
Need to eliminate possibility of drug abuse
test: Plain ECG - Sinus rhythm with nonspecific ST-T wave changes, left bundle branch block
reviewing results and questionning H1
reviewing symptoms in relation to the hypothesis
reviewing test results in relationship to hypothesis
reviewing results - cathecolmine total high !!!
set belief 80%
questioning test results that seem incoherent
submitting Hypothesis Pheochromocytoma
test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
test: WBC (total) - 4.5-11.0 X 10E9 /L
no further tests are indicated until we get more information
requesting physical examination information
selecting hypothesis Panic Attack with 50 belief
submitting Hypothesis Panic Attack - 50%


Problem solving path of E1, E2, E3, E4, E5 : Lydia - Log Data/Verbal Protocol - V3
E1: grey trace
E2: lilac trace
E3: red trace
E4: green trace
E5: blue trace
Problem
Set Hypothesis
Chart
Vital signs
Testing
clarifying instructions
instructions
37 yrs old
37-year-old banker
37 yr old woman
39 yr old banker
37 yr old
female
Few months feeling unwell
complains of unwellness
past few months she has been feeling unwell
feeling unwell for a few months
feeling unwell for a couple of months
Hypothesis 1: Depression
vague complain, you need to look at symptoms
high blood pressure shouldn't occur at that young age
can the new blood medication be causing the symptoms?
taking medicaiton for high blood pressure
problem may be related to medication
patient's attributing it to medication or high blood pressure
medication
or maybe medication is contributing to the symptoms by dilating arteries and which then causes release of cathecolamines (e.i. anti-hypertensive medication)
past history of blood pressure
high blood pressure
frequent headaches
frequent headaches, feeling anxious, palpitations, profuse sweating and flushing
frequent headaches
headache is a non specific complaint
H2 - commonest cause essential hypertension
other causes:renal artery stenosis (H3),pheochromocytoma(H4), adrenal adanoma (H5),Crohn's syndrome(H6),thyrotoxicosis(H7)
palpitations, sweating, and flushing makes me think of menopause (H8)
but this pattern with high blood pressure makes two alarm symptoms
extremely anxious
palpitations, profuse sweating, and flushing
more frequent in the past little while
constitutional symptoms
headaches, palpitations, sweating and flushing
episodes during which she feels "extremely anxious" with palpitations, profuse sweating, and flushing.
episodes are not triggered by any specific cause
symptoms come as episode, no specific trigger, more frequent recently
linking palpitations, profuse sweating and flushing
symptoms are coming as episodes
extremely anxious with palpitations, sweating, and flushing
symptoms more frequent
extreme anxiousness, palpitations, profuse sweating and flushing
episodes
lost 10 pounds in the last 4 months
dizzy from time to time
weight loss + dizziness add to the pattern
dizzy
weight loss would need to be documented
10 pounds in the last 4 months and
weight loss (10 pounds in 4 months) and dizzy
linking weight loss to metabolic or endocrine process
lost 10 pounds and dizzy
weight loss
dizzyness is non-specific
makes me think of secondary causes of hypertension
like pheochromocytoma, where tumour releases excess amount of stress hormones like cathecolamines that can cause these symptoms
how these symptoms relate to pheochrocytoma
but pheochromocytoma is rare so you need to keep other causes in mind
set Hypothesis Grave's disease and mentions pheocromocytoma
set belief 75%
Summary of evidences suggests: thyrotoxicosis or a pheochromocytoma.
set hypothesis of hypertyroid (Grave's disease)
set belief at 75%
reviewing and justifying hypothesis of Grave's
symptoms occuring in episodes and dizzyness are not typical of hyperthyroidism
alternative endocrinology diagnosis: cortisol increase, adrenal hormones
Hypothesis of pheocromocytoma
linking weight loss with anxiety and hyperthyroidism
selecting hyperthyroid as main hypothesis with 10% belief
selecting hypotheses: hypertryroid, pheochromocytoma, depression, panic attack
makes me wonder if she got pheocromocytoma
select Hypothesis Pheochromocytoma
set belief to 100%
back up hypothesis: Hyperthyroidism
what else... malignant hypertension would be another thing
set belief 50%
If teaching, need to tell them to rule out secondary cause of hypertension
Another Hypothesis: H9 drug abuse
Hypothesis 5 - Drug abuse but pheo is up on the list
Hypothesis 3,4 - Depression and essential hypertension with reaction to medication
Diabetes unlikely because of sweating and flushing
hypertensive crisis as a diagnosis of exclusion
Alternative hypothesis: drug abuse?
Alternative hypothesis: related to medication?
alternative hypothesis: Essential hypertension
can't think of other good alternative diagnostic
hypothesis anxiety
requesting hypothesis non available in menu: Anxiety
selecting hypotheses: hypertryroid, pheochromocytoma, depression, panic attack
switch to chart area
high heart rate and high blood pressure
vital signs: pulse of 98 per min, high blood pressure and slightly elevated temperature 37.9
afebrile
relatively normal pulse
normal breathing
hypertensive 180/103
saturation?
afebrile
tachycardic
hypertensive
elevated blood pressure and pulse rate
linking vital signs to hypotheses
What does this means, could this still be thyrotoxicosis?
It depends on the medications she is taking ei. if she was given Beta blockers
different classes: diuretic, beta-blocker, calcium channel blocker and vasodilator
in a 37yrs old, guidelines recommand beta blockers which could explain slower heart rate
because with Grave's disease, you would expect this woman to have a higher heart rate
But I still can't explain the high diastolic for a Grave's disease case
set Hypothesis strength to 60%
planning about checking blood pressure and doing physical exam
test: Fasting Blood Glucose Level - <11 mmol/L
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 7-14 mEq/L
test: Serum Liver Pancreatic Tests / Alanine Aminotransferase (ALT) - 8-20 U/L
test: Serum Electrolytes / Magnesium (Mg) - 1.4-2.0 mEq/L
test: Random Blood Glucose Level - <11 mmol/L
normally I would order baseline electrolytes tests
test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 8 - 25 mg/dL
ordinarily you do routine laboratory tests
test: Fasting Blood Glucose Level <11 mmol/L
presumably the serum potassium is normal
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 7-14 mEq/L
test: Serum Electrolytes / Bicarbonate (HCO3) - 22-26 mEq/L


Biochemistry tests were ordered to make sure things were normal
planning tests: TSH and T4
test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L
test: Thyroid Stimulating Hormone (TSH) - normal
test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
normal TSH rules out Grave's disease
test: Triiodothyronine (T3) - 0.92-2.78 nmol/L
test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L
test: Thyroxine (T4) - Free: 10-31 pmol/L ; Total: 58-140 nmol/L
test: Triiodothyronine (T3) - 0.92-2.78 nmol/L
for instance, someone could be in renal failure
to confirm the diagnosis of pheocromocytoma I need to get the hormone levels
since I don't do this on a regular basis, I need to look at this list
test: Aldosterone - 0.06-0.7 nmol/L (supine), 2-5 times supine value if upright
test: Adrenocorticotropin homrone (ACTH) - 1.8-17.4 pmol/L (0800 hr) ; 1.5-6.6 pmol/L (1600 hr)
test: Cortisol - 138-690 nmol/L (0800 hr) ; 138-414 nmol/L (1600 hr)
test: Dehydroepiandosterone Sulfate (DHEA-S) - 100-6190 ug/L (M) ; 120-5350 ug/L (female)
here are the test that I was looking for
select Hypothesis Pheochromocytoma
new hypothesis: Cushing syndrome
test: Cortisol - 138-690 nmol/L (0800 hr) ; 138-414 nmol/L (1600 hr)
new main hypothesis: Pheocromocytoma - belief 55%
risk of hypertensive crisis?
looking for toxicology tests
test: Eye Exam - Normal
would test CBC and basic to make sure she's not anemic
exploring all potential tests available and comments on toxicology screen
plan for a urine collection for catecholamines
test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
test: Urinary Catecholamines / Free - 1560 mmol / 24 hr
test: Urinary Catecholamines / Norepinephrine - 89-591 nmol/day
test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
test: Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr
test: Urinary Metabolites / Vanillylmandelic Acid (VMA) 24 hr - Positive - 120 micromol / 24 hr
test: Urinary Catecholamines / Free - 1560 mmol / 24 hr
yep, they are positive
metanephrine
test: Urinary Catecholamines / Dopamine - 340-3134 nmol/day
test: Urinary Catecholamines / Epinephrine - 11-131 nmol/day
test: Urinary Catecholamines / Free - 1560 mmol / 24 hr
considers Pheochromocytoma and mentions related tests
test: Urinary Catecholamines / Norepinephrine - 89-591 nmol/day
test: Urinary Catecholamines / Total (Epinephrine + Norepinephrine) - 27 micromol / 24 hr
since I am not familiar with pheocromocytoma, I began by testing for the metabolic baseline
I ordered the important tests that I expected to be abnormal because I knew the diagnosis
I would review literature on pheochromocytoma, do a consult and look up tests to order
for diagnosis, I ordered tests but did not know exactly which ones I needed before seeing them
100% convinced with the clinical history but this is the actual proof
worrying about medication precipitating condition
reviewing evidence and hypothesis, planning for CT of abdomen
add imaging test: CT abdomen
test not found - why CT of whole body is not an optimal option
test: CT / Body - Normal
mentioning other potential tests like MRI and ultrasound
planning to order radiology and get a CT scan of abdomen
test: CT / Body - Normal
test: Ultrasound / Abdominal Scan - No presence of gallstones
looking for radiology tests
looking for CT scan of adrenal glands
Another differential hypothesis, multiple endocrinoplasia group
Not sure about which test should be ordered when
how to manage a patient with pheocromocytoma
Need to eliminate possibility of drug abuse
test: Plain ECG - Sinus rhythm with nonspecific ST-T wave changes, left bundle branch block
reviewing results and questioning H1
reviewing symptoms in relation to the hypothesis
reviewing test results in relationship to hypothesis
reviewing results - cathecolmine total high !!!
set belief 80%
questioning test results that seem incoherent
test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
test: WBC (total) - 4.5-11.0 X 10E9 /L
no further tests are indicated until we get more information
requesting physical examination information
selecting hypothesis Panic Attack with 50 belief
submitting Hypothesis Pheochromocytoma - 100%
submitting Hypothesis Pheochromocytoma - 100%
submitting Hypothesis Pheochromocytoma - 65%
submitting Hypothesis Pheochromocytoma - 60%
submitting Hypothesis Panic Attack - 50%

APPENDIX G

Case 2 Categorized Episodes for Expert 1

E1	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	urinate more frequently	Fatigue has progressed	16 yrs old
2	feeling excessively thirsty	nausea	test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21
3	select Hypothesis Diabetes Mellitus (type I)	difficulty seeing	test: Serum Electrolytes / Sodium (Na) - 130 mEq/L
4	Random Blood Glucose Level - 18.2 mmol/L	6 pound weight loss	test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L
5	test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L	Nausea, vomiting and abdominal pain	test: WBC (total) - 12 x 10E9 / L
6	test: pH - 7	possible complications - diabetic ketoacidosis, pancreatitus or gastroentirits or gull stones	test: Chest and content / Chest X - Ray - No bone, heart, or lung abnormalities seen
7	test: Serum Electrolytes / Osmolilty - 320 mmol/Kg H20	heart rate is up	
8	test: Serum Ketones - Present	blood pressure is lower than it should be	
9		test: HbA1C - 12.5%	
10		test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 8 - 25 mg/dL	
11		test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L (relative to BMI)	

Case 2 Categorized Episodes for Expert 2

E2	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	fatigue + urinate more frequently make me think of diabetes or maybe urinary infection	6 pound weight loss in the past month	16 year old
2	feeling excessively thirsty strenghens the endocrine hypotheses	increased respiratory rate support a diagnosis of acidosis	fatigue
3	nauseated and difficulty seeing are even more specific	investigating abdominal pain, nausea and vomiting: surgical abdomen, intra pelvis abscess, pancreatitis, gall stones	
4	having to urinate more frequently	test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L (relative to BMI)	
5	excessively thirsty	test: Serum Electrolytes / Sodium (Na) - 130 mEq/L	
6	difficulty seeing	test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)	
7	Today Stephanie is experiencing nausea, vomiting, and abdominal pain	looking for infection - planning septic workup, checking abdomen, physical, x-ray, ultrasound	
8	pulse is high	test: Ultrasound / Abdominal Scan - No presence of gallstones	
9	blood pressure is not as low	test: Chest and content / Chest X - Ray - No bone, heart, or lung abnormalities seen	
10	test: Random Blood Glucose Level - 18.2 mmol/L	test: Urinalysis / Leukocyte Esterase - Negative	
11	test: Serum Ketones - Present		
12	test: WBC (total) - 12 x 10E9 / L		
13	test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21		
14	test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L		
15	submitting Hypothesis Diabetes Mellitus (type I) - 100%		

Case 2 Categorized Episodes for Expert 3

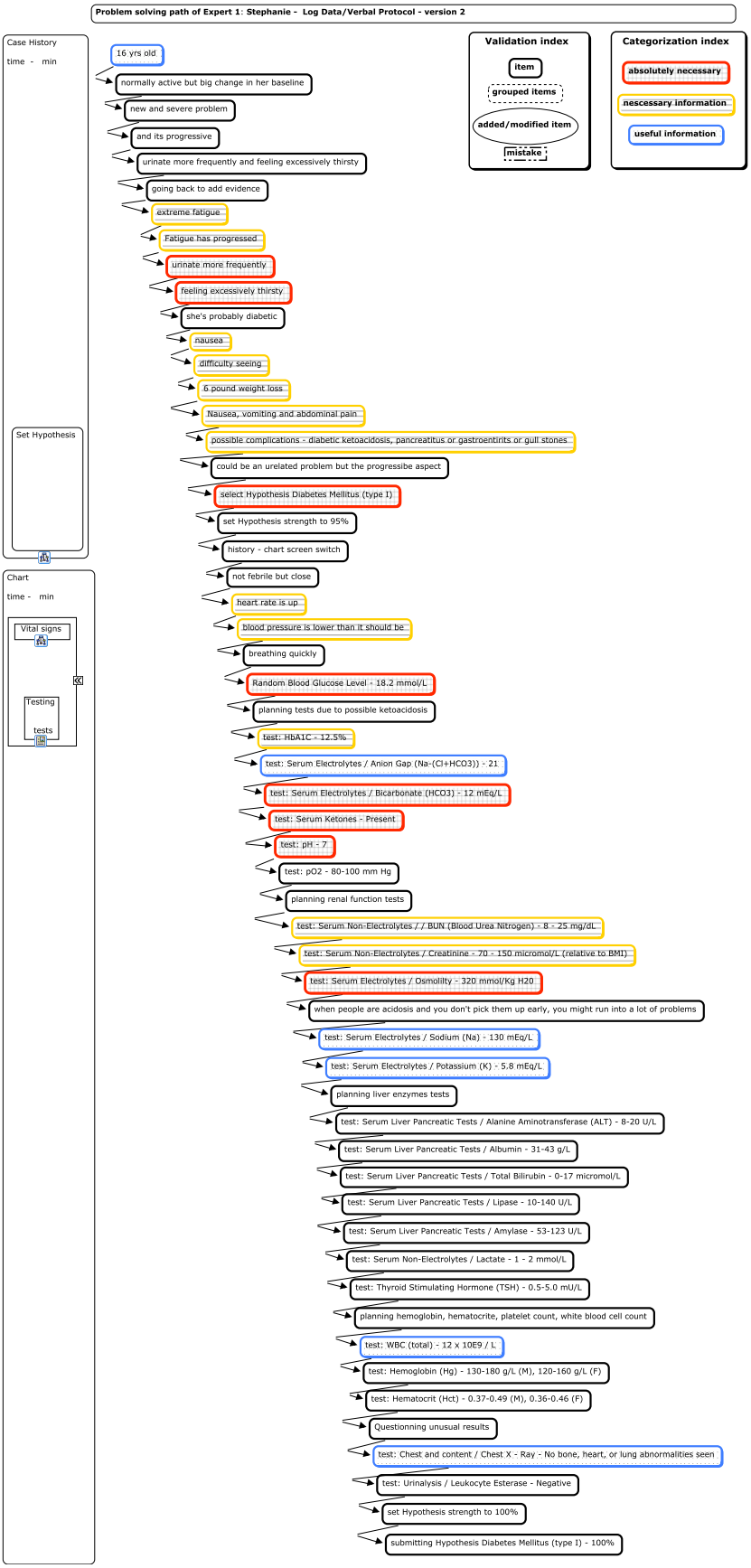
E3	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	16 yr old teenaged girl	main symptom = extreme fatigue	difficulty seeing
2	another very imp. symptom - urinate more frequently	6 pound weight loss in the past month	RR is bit fast at 22
3	excessively thirsty	test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L - need to be monitored	
4	today's set of symptoms - nausea, vomiting and abdominal pain	test: Serum Electrolytes / Osmolilty - 320 mmol/Kg H20 - it's high	
5	hypothesis 1 - Diabetes Mellitus (type I)	test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21 - elevated	
6	tachycardic at 110 probably due to dehydratation	test: Serum Electrolytes / Sodium (Na) - 130 mEq/L	
7	almost normal blood pressure 95/72	test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L	
8	test: Random Blood Glucose Level - 18.2 mmol/L - expected higher result	test: pH - 7	
9	submitting Hypothesis Diabetes Mellitus (type I) - 100%	test: Urinalysis / Ketones - Present	
10		checking for infection	
11		test: Serum Non-Electrolytes / Creatinine - 67 micromol/L	
12		test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 5.4 mmol/L	
13		could still check for infection with x-ray	
14		DKA	

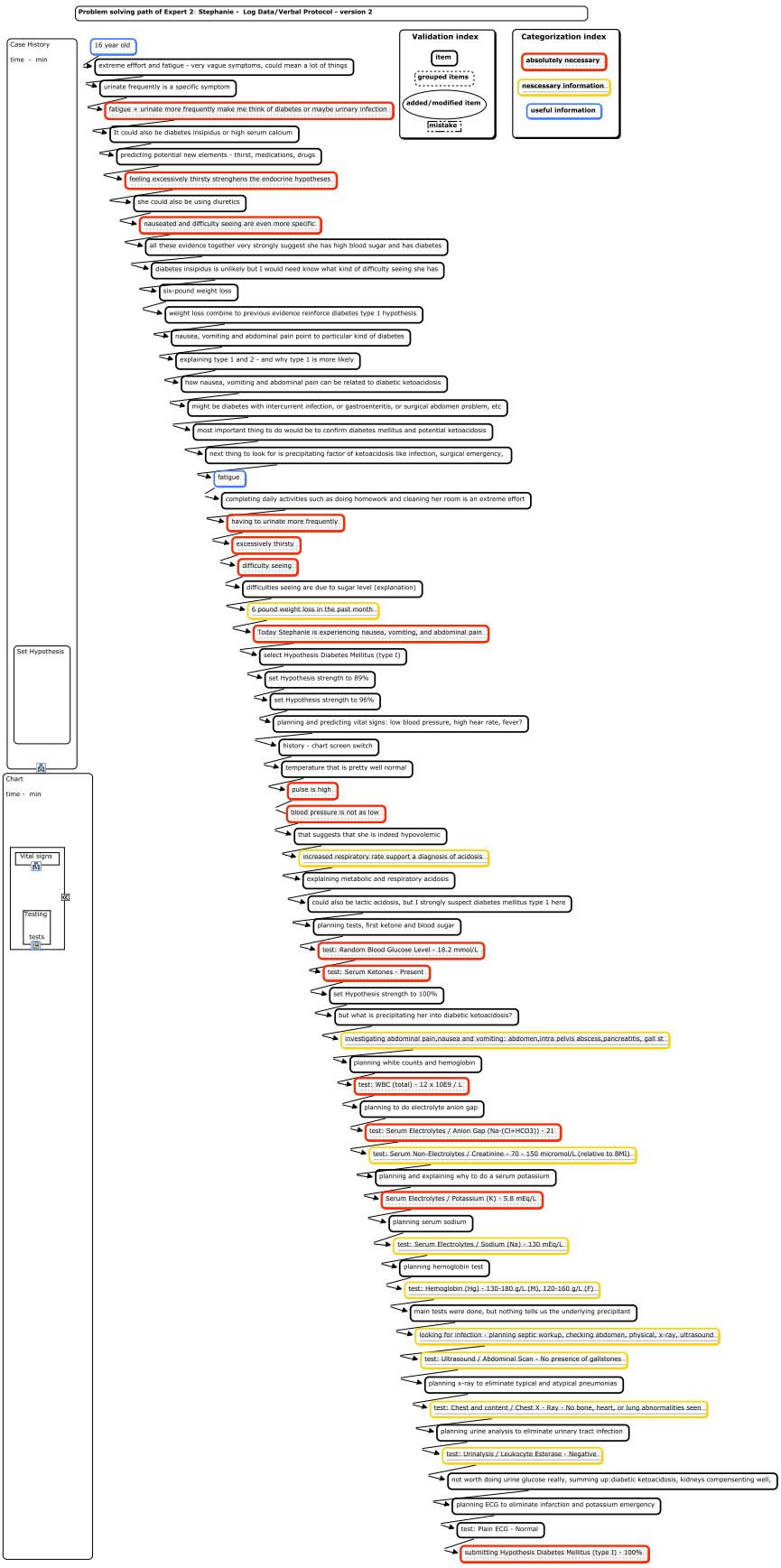
Case 2 Categorized Episodes for Expert 4

E4	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	symptoms: fatigue, urinate frequently, excessively thirsty	H1: juvenile onset diabetic	16 year old teenage girl
2	select Hypothesis Diabetes Mellitus (type I)	concerned with nausea	reviewing evidence, stating alternative hypothesis - anemia
3	Polyurea and Polydipsia	difficulty seeing = blurred vision due to glucose	extreme fatigue
4	today's symptoms - nausea, vomiting and abdominal pain	so nausea + vomiting + abdominal pain = type 1 diabetes with ketoacidosis	nauseated
5	tachycardic, hypotensive and tachypeic as expected	difficulty seeing	6 pound weight loss
6	test: Random Blood Glucose Level - 18.2 mmol/L	test: HbA1C - 12.5%	
7	test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L	test: Serum Electrolytes / Phosphate - 1.8 mg/dL	
8	test: Serum Ketones - Present	test: Serum Electrolytes / Magnesium (Mg) - 1.7 mEq/L	
9	test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L	summarizing and reviewing clinical picture	
10	test: Serum Non-Electrolytes / Creatinine - 67 micromol/L		
11	test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 5.4 mmol/L		
12	test: pH - 7		
13	test: pCO2 - 24 mmHg		
14	Looking for infection that could be precipitating diabetes		
15	test: WBC (total) - 12 x 10E9 / L		
16	submitting Hypothesis Diabetes Mellitus (type I) - 100%		

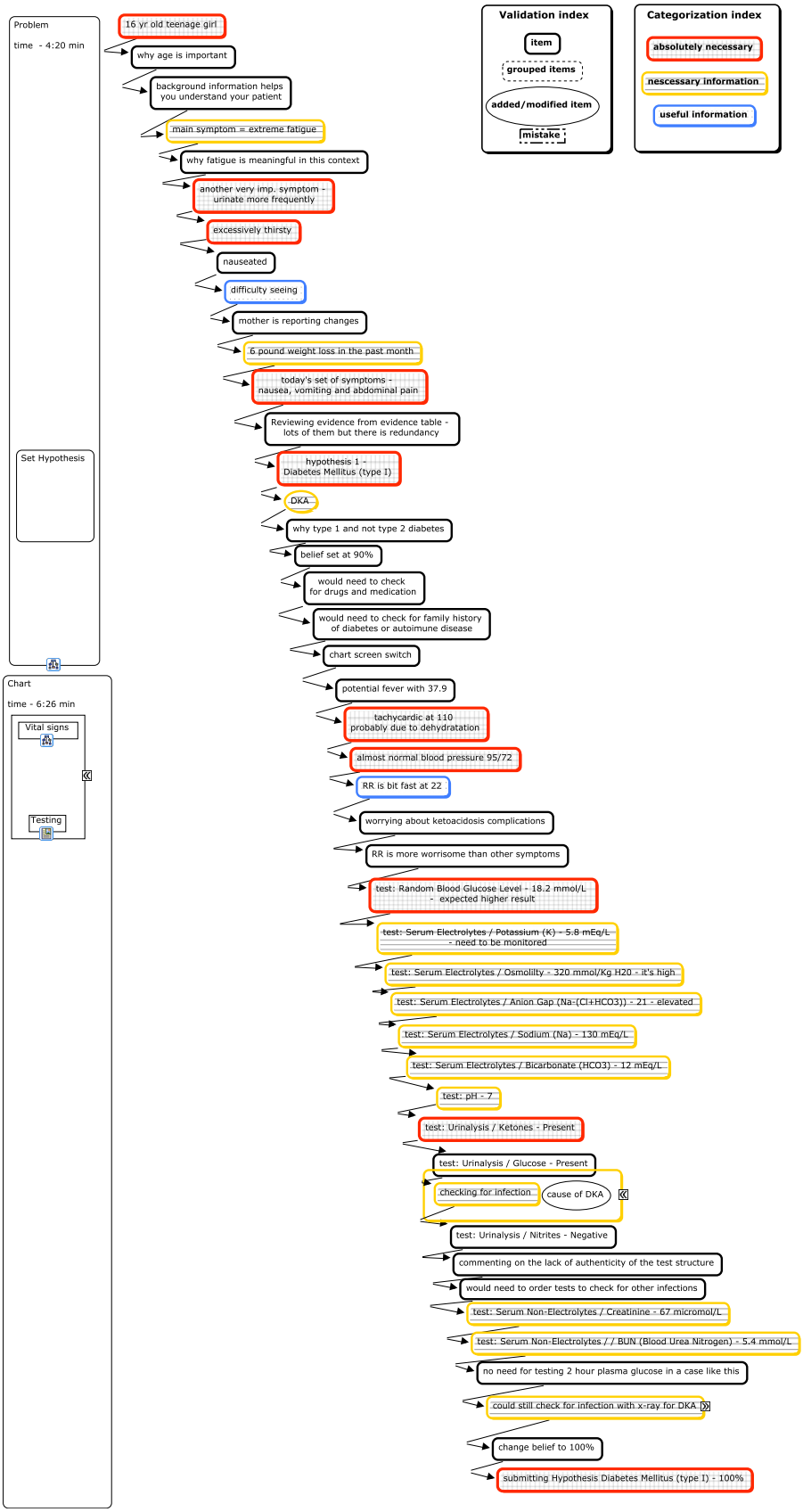
APPENDIX H

Individual Categorized Visual Representations for Case 2


Problem solving path of Expert 1: Stephanie - Log Data/Verbal Protocol - version 2
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
Case History
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
tests
16 yrs old
normally active but big change in her baseline
new and severe problem
and its progressive
urinate more frequently and feeling excessively thirsty
going back to add evidence
extreme fatigue
Fatigue has progressed
urinate more frequently
feeling excessively thirsty
she's probably diabetic
nausea
difficulty seeing
6 pound weight loss
Nausea, vomiting and abdominal pain
possible complications - diabetic ketoacidosis, pancreatitus or gastroentirits or gull stones
could be an urelated problem but the progressibe aspect
select Hypothesis Diabetes Mellitus (type I)
set Hypothesis strength to 95%
history - chart screen switch
not febrile but close
heart rate is up
blood pressure is lower than it should be
breathing quickly
Random Blood Glucose Level - 18.2 mmol/L
planning tests due to possible ketoacidosis
test: HbA1C - 12.5%
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21
test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L
test: Serum Ketones - Present
test: pH - 7
test: pO2 - 80-100 mm Hg
planning renal function tests
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 8 - 25 mg/dL
test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L (relative to BMI)
test: Serum Electrolytes / Osmolilty - 320 mmol/Kg H20
when people are acidosis and you don't pick them up early, you might run into a lot of problems
test: Serum Electrolytes / Sodium (Na) - 130 mEq/L
test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L
planning liver enzymes tests
test: Serum Liver Pancreatic Tests / Alanine Aminotransferase (ALT) - 8-20 U/L
test: Serum Liver Pancreatic Tests / Albumin - 31-43 g/L
test: Serum Liver Pancreatic Tests / Total Bilirubin - 0-17 micromol/L
test: Serum Liver Pancreatic Tests / Lipase - 10-140 U/L
test: Serum Liver Pancreatic Tests / Amylase - 53-123 U/L
test: Serum Non-Electrolytes / Lactate - 1 - 2 mmol/L
test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L
planning hemoglobin, hematocrite, platelet count, white blood cell count
test: WBC (total) - 12 x 10E9 / L
test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
test: Hematocrit (Hct) - 0.37-0.49 (M), 0.36-0.46 (F)
Questionning unusual results
test: Chest and content / Chest X - Ray - No bone, heart, or lung abnormalities seen
test: Urinalysis / Leukocyte Esterase - Negative
set Hypothesis strength to 100%
submitting Hypothesis Diabetes Mellitus (type I) - 100%


Problem solving path of Expert 2: Stephanie - Log Data/Verbal Protocol - version 2
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
Case History
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
tests
16 year old
extreme efffort and fatigue - very vague symptoms, could mean a lot of things
urinate frequently is a specific symptom
fatigue + urinate more frequently make me think of diabetes or maybe urinary infection
It could also be diabetes insipidus or high serum calcium
predicting potential new elements - thirst, medications, drugs
feeling excessively thirsty strenghens the endocrine hypotheses
she could also be using diuretics
nauseated and difficulty seeing are even more specific
all these evidence together very strongly suggest she has high blood sugar and has diabetes
diabetes insipidus is unlikely but I would need know what kind of difficulty seeing she has
six-pound weight loss
weight loss combine to previous evidence reinforce diabetes type 1 hypothesis
nausea, vomiting and abdominal pain point to particular kind of diabetes
explaining type 1 and 2 - and why type 1 is more likely
how nausea, vomiting and abdominal pain can be related to diabetic ketoacidosis
might be diabetes with intercurrent infection, or gastroenteritis, or surgical abdomen problem, etc
most important thing to do would be to confirm diabetes mellitus and potential ketoacidosis
next thing to look for is precipitating factor of ketoacidosis like infection, surgical emergency,
fatigue
completing daily activities such as doing homework and cleaning her room is an extreme effort
having to urinate more frequently
excessively thirsty
difficulty seeing
difficulties seeing are due to sugar level (explanation)
6 pound weight loss in the past month
Today Stephanie is experiencing nausea, vomiting, and abdominal pain
select Hypothesis Diabetes Mellitus (type I)
set Hypothesis strength to 89%
set Hypothesis strength to 96%
planning and predicting vital signs: low blood pressure, high hear rate, fever?
history - chart screen switch
temperature that is pretty well normal
pulse is high
blood pressure is not as low
that suggests that she is indeed hypovolemic
increased respiratory rate support a diagnosis of acidosis
explaining metabolic and respiratory acidosis
could also be lactic acidosis, but I strongly suspect diabetes mellitus type 1 here
planning tests, first ketone and blood sugar
test: Random Blood Glucose Level - 18.2 mmol/L
test: Serum Ketones - Present
set Hypothesis strength to 100%
but what is precipitating her into diabetic ketoacidosis?
investigating abdominal pain,nausea and vomiting: abdomen,intra pelvis abscess,pancreatitis, gall st
planning white counts and hemoglobin
test: WBC (total) - 12 x 10E9 / L
planning to do electrolyte anion gap
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21
test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L (relative to BMI)
planning and explaining why to do a serum potassium
Serum Electrolytes / Potassium (K) - 5.8 mEq/L
planning serum sodium
test: Serum Electrolytes / Sodium (Na) - 130 mEq/L
planning hemoglobin test
test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
main tests were done, but nothing tells us the underlying precipitant
looking for infection - planning septic workup, checking abdomen, physical, x-ray, ultrasound
test: Ultrasound / Abdominal Scan - No presence of gallstones
planning x-ray to eliminate typical and atypical pneumonias
test: Chest and content / Chest X - Ray - No bone, heart, or lung abnormalities seen
planning urine analysis to eliminate urinary tract infection
test: Urinalysis / Leukocyte Esterase - Negative
not worth doing urine glucose really, summing up:diabetic ketoacidosis, kidneys compensenting well,
planning ECG to eliminate infarction and potassium emergency
test: Plain ECG - Normal
submitting Hypothesis Diabetes Mellitus (type I) - 100%

Problem solving path of Expert 3: Stephanie - Log Data/Verbal Protocol - V2



Problem solving path of Expert 4: Stephanie - Data Protocol - V2

Problem

time - min

Chart

time - min

Vital signs

Testing

16 year old teenage girl

good fitness level
fit - implies previously good health

symptoms: fatigue, urinate frequently, excessively thirsty

H1: juvenile onset diabetic

reviewing evidence, stating alternative hypothesis - anemia

but diabetes is more likely

concerned with nausea

difficulty seeing = blurred vision due to glucose

so nausea + vomiting + abdominal pain = type 1 diabetes with ketoacidosis

select Hypothesis Diabetes Mellitus

set belief: 79%

what I mean by 80%

extreme fatigue

Polyurea and Polydipsia

nauseated

difficulty seeing

6 pounds weight loss

today's symptoms - nausea, vomiting and abdominal pain

history - chart screen switch

tachycardic, hypotensive and tachypeic as expected

planning tests

test: Random Blood Glucose Level - 18.2 mmol/L

test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L

test: Serum Ketones - Present

change belief to 99%

planning to test for complications

test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L

test: Serum Non-Electrolytes / Creatinine - 67 micromol/L

test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 5.4 mmol/L

test: HbA1C - 12.5%

test: Ph - 7

test: pCO2 - 24 mmHg

Looking for infection that could be precipitating diabetes

test: WBC (total) - 12 x 10E9 / L

test: Serum Electrolytes / Phosphate - 1.8 mg/dL

test: Serum Electrolytes / Magnesium (Mg) - 1.7 mEq/L

summarizing and reviewing clinical picture

submitting Hypothesis Diabetes Mellitus (type I) - 99%

Validation index

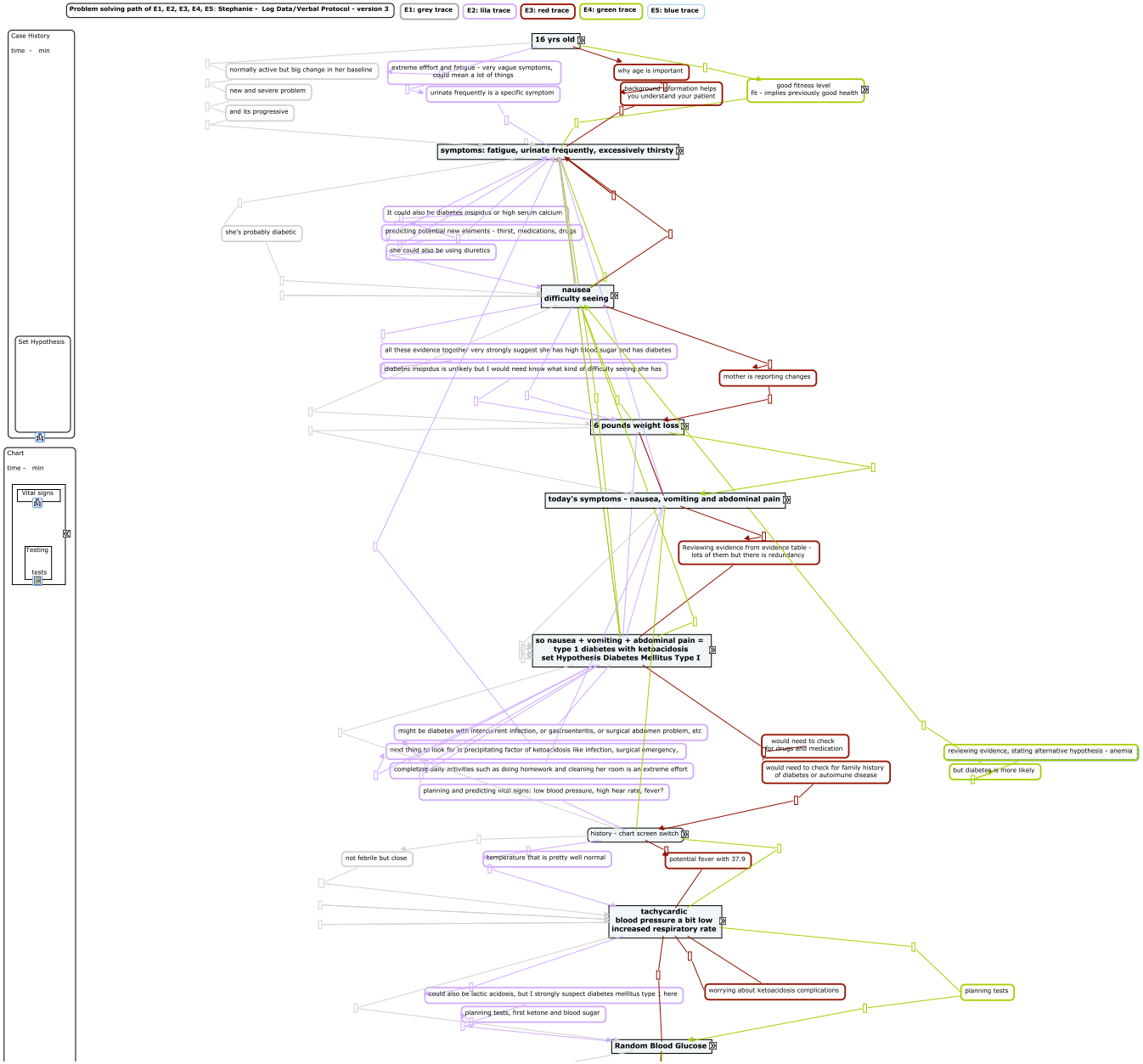
- item
- grouped items
- added/modified item
- mistake

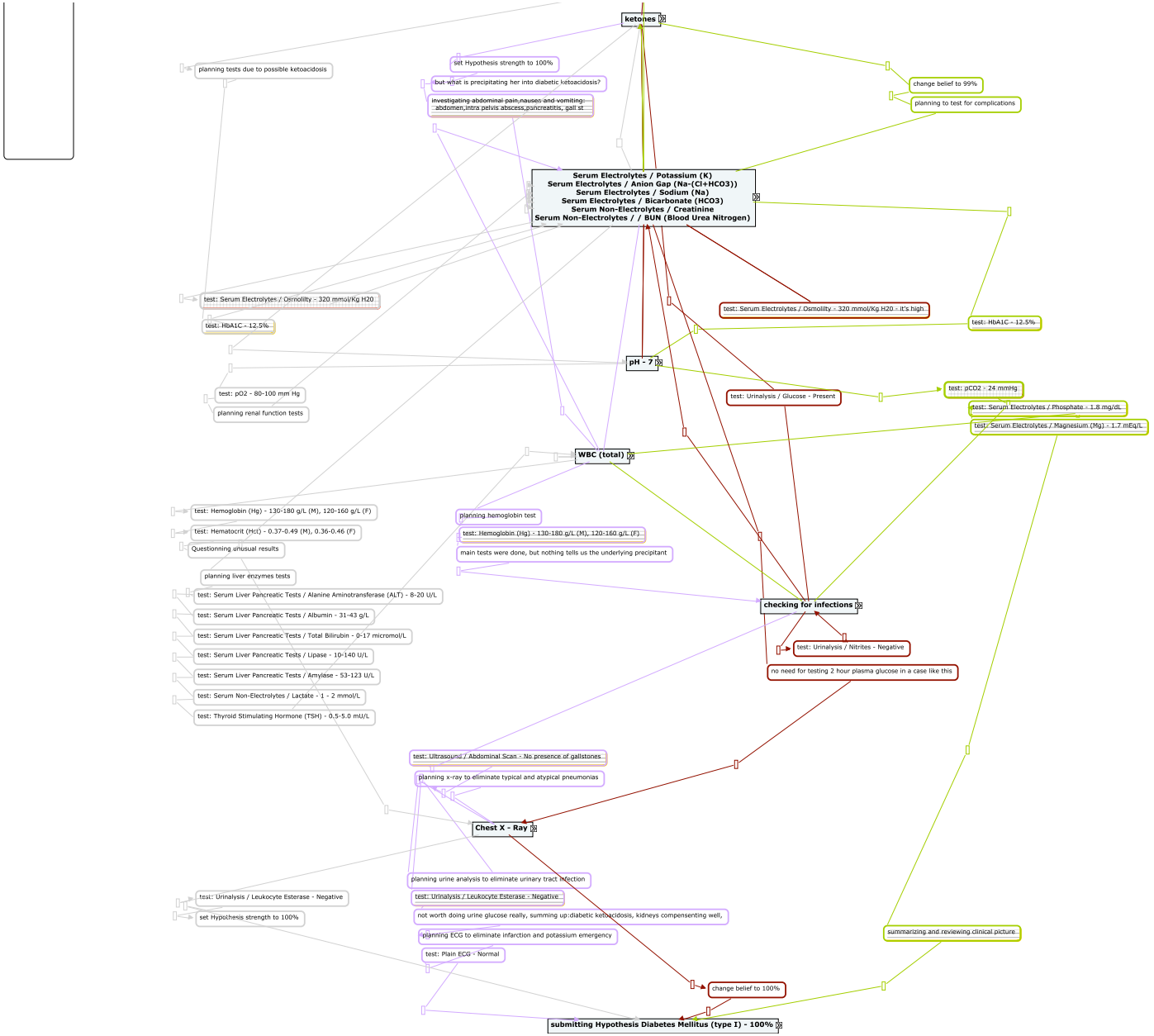
Categorization index

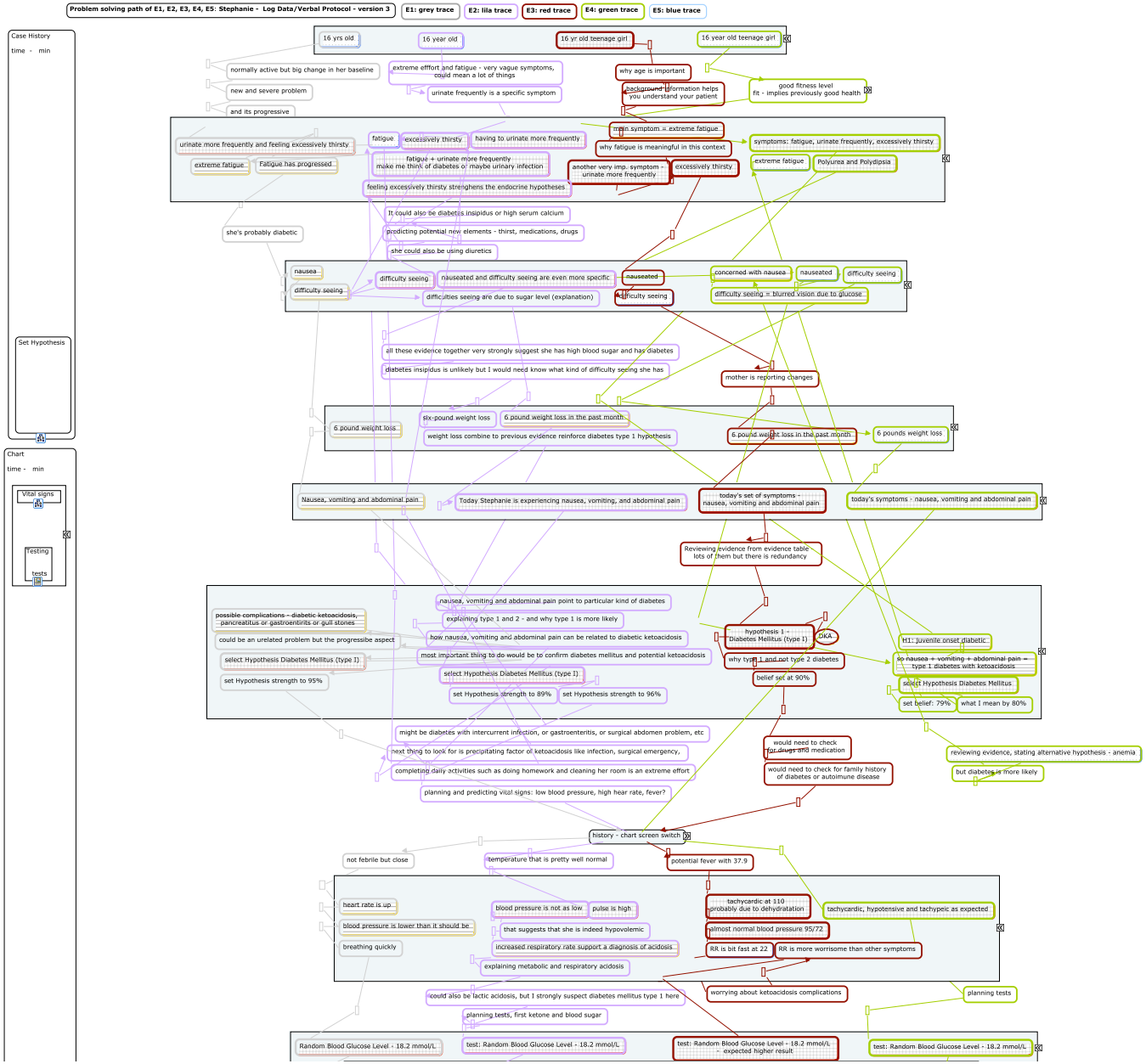
- absolutely necessary
- nescessary information
- useful information

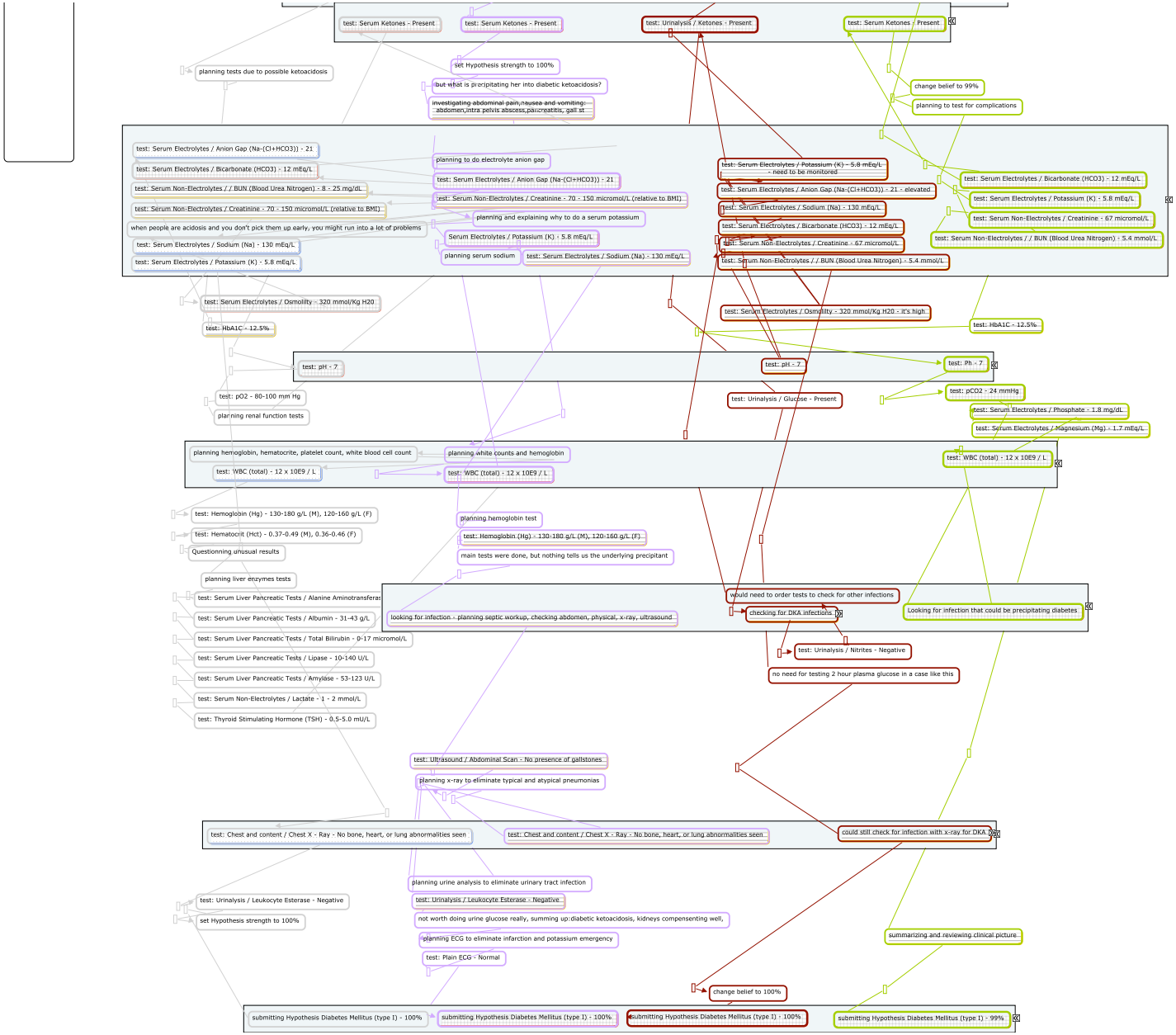
APPENDIX I

Merged Representations for Case 2 at Level 1 and Level 2


Problem solving path of E1, E2, E3, E4, E5: Stephanie - Log Data/Verbal Protocol - version 3
E1: grey trace
E2: lila trace
E3: red trace
E4: green trace
E5: blue trace
Case History
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
tests
16 yrs old
normally active but big change in her baseline
new and severe problem
and its progressive
extreme efffort and fatigue - very vague symptoms, could mean a lot of things
urinate frequently is a specific symptom
why age is important
background information helps you understand your patient
good fitness level
fit - implies previously good health
symptoms: fatigue, urinate frequently, excessively thirsty
she's probably diabetic
It could also be diabetes insipidus or high serum calcium
predicting potential new elements - thirst, medications, drugs
she could also be using diuretics
nausea
difficulty seeing
all these evidence together very strongly suggest she has high blood sugar and has diabetes
diabetes insipidus is unlikely but I would need know what kind of difficulty seeing she has
mother is reporting changes
6 pounds weight loss
today's symptoms - nausea, vomiting and abdominal pain
Reviewing evidence from evidence table - lots of them but there is redundancy
so nausea + vomiting + abdominal pain =
type 1 diabetes with ketoacidosis
set Hypothesis Diabetes Mellitus Type I
might be diabetes with intercurrent infection, or gastroenteritis, or surgical abdomen problem, etc
next thing to look for is precipitating factor of ketoacidosis like infection, surgical emergency,
completing daily activities such as doing homework and cleaning her room is an extreme effort
planning and predicting vital signs: low blood pressure, high hear rate, fever?
would need to check for drugs and medication
would need to check for family history of diabetes or autoimmune disease
reviewing evidence, stating alternative hypothesis - anemia
but diabetes is more likely
history - chart screen switch
not febrile but close
temperature that is pretty well normal
potential fever with 37.9
tachycardic
blood pressure a bit low
increased respiratory rate
could also be lactic acidosis, but I strongly suspect diabetes mellitus type 1 here
planning tests, first ketone and blood sugar
worrying about ketoacidosis complications
planning tests
Random Blood Glucose


ketones
planning tests due to possible ketoacidosis
set Hypothesis strength to 100%
but what is precipitating her into diabetic ketoacidosis?
investigating abdominal pain,nauses and vomiting: abdomen,intra pelvis abscess,puncreatitis, gall st
change belief to 99%
planning to test for complications
Serum Electrolytes / Potassium (K)
Serum Electrolytes / Anion Gap (Na-(Cl+HCO3))
Serum Electrolytes / Sodium (Na)
Serum Electrolytes / Bicarbonate (HCO3)
Serum Non-Electrolytes / Creatinine
Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen)
test: Serum Electrolytes / Osmolilty - 320 mmol/Kg H20
test: HbA1C - 12.5%
test:. Serum Electrolytes / Osmolilty - 320 mmol/Kg H20 - it's high
test:. HbA1C - 12.5%
pH - 7
test: pO2 - 80-100 mm Hg
planning renal function tests
test: Urinalysis / Glucose - Present
test: pCO2 - 24 mmHg
test:. Serum Electrolytes / Phosphate - 1.8 mg/dL
test:. Serum Electrolytes / Magnesium (Mg) - 1.7 mEq/L
WBC (total)
test:. Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
test:. Hematocrit (Hct) - 0.37-0.49 (M), 0.36-0.46 (F)
Questionning unusual results
planning hemoglobin test
test:. Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
main tests were done, but nothing tells us the underlying precipitant
planning liver enzymes tests
test: Serum Liver Pancreatic Tests / Alanine Aminotransferase (ALT) - 8-20 U/L
test: Serum Liver Pancreatic Tests / Albumin - 31-43 g/L
test: Serum Liver Pancreatic Tests / Total Bilirubin - 0-17 micromol/L
test: Serum Liver Pancreatic Tests / Lipase - 10-140 U/L
test: Serum Liver Pancreatic Tests / Amylase - 53-123 U/L
test: Serum Non-Electrolytes / Lactate - 1 - 2 mmol/L
test: Thyroid Stimulating Hormone (TSH) - 0.5-5.0 mU/L
checking for infections
test: Urinalysis / Nitrites - Negative
no need for testing 2 hour plasma glucose in a case like this
test: Ultrasound / Abdominal Scan - No presence of gallstones
planning x-ray to eliminate typical and atypical pneumonias
Chest X - Ray
planning urine analysis to eliminate urinary tract infection
test:. Urinalysis / Leukocyte Esterase - Negative
test: Urinalysis / Leukocyte Esterase - Negative
set Hypothesis strength to 100%
not worth doing urine glucose really, summing up:diabetic ketoacidosis, kidneys compensenting well,
planning ECG to eliminate infarction and potassium emergency
test: Plain ECG - Normal
summarizing and reviewing clinical picture
change belief to 100%
submitting Hypothesis Diabetes Mellitus (type I) - 100%


Problem solving path of E1, E2, E3, E4, E5: Stephanie - Log Data/Verbal Protocol - version 3
E1: grey trace
E2: lila trace
E3: red trace
E4: green trace
E5: blue trace
Case History
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
tests
16 yrs old
16 year old
16 yr old teenage girl
16 year old teenage girl
normally active but big change in her baseline
new and severe problem
and its progressive
extreme efffort and fatigue - very vague symptoms, could mean a lot of things
urinate frequently is a specific symptom
why age is important
background information helps you understand your patient
good fitness level
fit - implies previously good health
urinate more frequently and feeling excessively thirsty
fatigue
excessively thirsty
having to urinate more frequently
extreme fatigue
Fatigue has progressed
fatigue + urinate more frequently
make me think of diabetes or maybe urinary infection
feeling excessively thirsty strenghens the endocrine hypotheses
main symptom = extreme fatigue
why fatigue is meaningful in this context
another very imp. symptom - urinate more frequently
excessively thirsty
symptoms: fatigue, urinate frequently, excessively thirsty
extreme fatigue
Polyurea and Polydipsia
It could also be diabetes insipidus or high serum calcium
predicting potential new elements - thirst, medications, drugs
she's probably diabetic
she could also be using diuretics
nausea
difficulty seeing
difficulty seeing
nauseated and difficulty seeing are even more specific
difficulties seeing are due to sugar level (explanation)
nauseated
difficulty seeing
concerned with nausea
nauseated
difficulty seeing
difficulty seeing = blurred vision due to glucose
all these evidence together very strongly suggest she has high blood sugar and has diabetes
diabetes insipidus is unlikely but I would need know what kind of difficulty seeing she has
mother is reporting changes
6 pound weight loss
six-pound weight loss
6 pound weight loss in the past month
weight loss combine to previous evidence reinforce diabetes type 1 hypothesis
6 pound weight loss in the past month
6 pounds weight loss
Nausea, vomiting and abdominal pain
Today Stephanie is experiencing nausea, vomiting, and abdominal pain
today's set of symptoms - nausea, vomiting and abdominal pain
today's symptoms - nausea, vomiting and abdominal pain
Reviewing evidence from evidence table - lots of them but there is redundancy
nausea, vomiting and abdominal pain point to particular kind of diabetes
possible complications - diabetic ketoacidosis, pancreatitus or gastroentints or gull stones
explaining type 1 and 2 - and why type 1 is more likely
could be an urelated problem but the progressibe aspect
how nausea, vomiting and abdominal pain can be related to diabetic ketoacidosis
most important thing to do would be to confirm diabetes mellitus and potential ketoacidosis
select Hypothesis Diabetes Mellitus (type I)
set Hypothesis strength to 95%
select Hypothesis Diabetes Mellitus (type I)
set Hypothesis strength to 89%
set Hypothesis strength to 96%
hypothesis 1 - Diabetes Mellitus (type I)
DKA
why type 1 and not type 2 diabetes
belief set at 90%
H1: juvenile onset diabetic
so nausea + vomiting + abdominal pain = type 1 diabetes with ketoacidosis
select Hypothesis Diabetes Mellitus
set belief: 79%
what I mean by 80%
might be diabetes with intercurrent infection, or gastroenteritis, or surgical abdomen problem, etc
next thing to look for is precipitating factor of ketoacidosis like infection, surgical emergency,
completing daily activities such as doing homework and cleaning her room is an extreme effort
planning and predicting vital signs: low blood pressure, high hear rate, fever?
would need to check for diuretic and medication
would need to check for family history of diabetes or autoimune disease
reviewing evidence, stating alternative hypothesis - anemia
but diabetes is more likely
history - chart screen switch
not febrile but close
temperature that is pretty well normal
potential fever with 37.9
heart rate is up
blood pressure is lower than it should be
breathing quickly
blood pressure is not as low
pulse is high
that suggests that she is indeed hypovolemic
increased respiratory rate support a diagnosis of acidosis
explaining metabolic and respiratory acidosis
tachycardic at 110 - probably due to dehydratation
tachycardic, hypotensive and tachypeic as expected
almost normal blood pressure 95/72
RR is bit fast at 22
RR is more worrisome than other symptoms
could also be lactic acidosis, but I strongly suspect diabetes mellitus type 1 here
planning tests, first ketone and blood sugar
worrying about ketoacidosis complications
planning tests
Random Blood Glucose Level - 18.2 mmol/L
test: Random Blood Glucose Level - 18.2 mmol/L
test: Random Blood Glucose Level - 18.2 mmol/L - expected higher result
test: Random Blood Glucose Level - 18.2 mmol/L


test: Serum Ketones - Present
test: Serum Ketones - Present
test: Urinalysis / Ketones - Present
test: Serum Ketones - Present
planning tests due to possible ketoacidosis
set Hypothesis strength to 100%
but what is precipitating her into diabetic ketoacidosis?
investigating abdominal pain,nausea and vomiting: abdomen,intra pelvis abscess,pancreatitis, gall st
change belief to 99%
planning to test for complications
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21
test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 8 - 25 mg/dL
test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L (relative to BMI)
when people are acidosis and you don't pick them up early, you might run into a lot of problems
test: Serum Electrolytes / Sodium (Na) - 130 mEq/L
test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L
planning to do electrolyte anion gap
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21
test: Serum Non-Electrolytes / Creatinine - 70 - 150 micromol/L (relative to BMI)
planning and explaining why to do a serum potassium
Serum Electrolytes / Potassium (K) - 5.8 mEq/L
planning serum sodium
test: Serum Electrolytes / Sodium (Na) - 130 mEq/L
test: Serum Electrolytes / Potassium (K) – 5.8 mEq/L - need to be monitored
test: Serum Electrolytes / Anion Gap (Na-(Cl+HCO3)) - 21 - elevated
test: Serum Electrolytes / Sodium (Na) - 130 mEq/L
test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L
test: Serum Non-Electrolytes / Creatinine - 67 micromol/L
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 5.4 mmol/L
test: Serum Electrolytes / Bicarbonate (HCO3) - 12 mEq/L
test: Serum Electrolytes / Potassium (K) - 5.8 mEq/L
test: Serum Non-Electrolytes / Creatinine - 67 micromol/L
test: Serum Non-Electrolytes / / BUN (Blood Urea Nitrogen) - 5.4 mmol/L
test: Serum Electrolytes / Osmolility - 320 mmol/Kg H20
test: HbA1C - 12.5%
test: Serum Electrolytes / Osmolility - 320 mmol/Kg H20 - it's high
test: HbA1C - 12.5%
test: pH - 7
test: pH - 7
test: Ph - 7
test: pO2 - 80-100 mm Hg
planning renal function tests
test: Urinalysis / Glucose - Present
test: pCO2 - 24 mmHg
test: Serum Electrolytes / Phosphate - 1.8 mg/dL
test: Serum Electrolytes / Magnesium (Mg) - 1.7 mEq/L
planning hemoglobin, hematocrite, platelet count, white blood cell count
test: WBC (total) - 12 x 10E9 / L
planning white counts and hemoglobin
test: WBC (total) - 12 x 10E9 / L
test: WBC (total) - 12 x 10E9 / L
test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
test: Hematocrit (Hct) - 0.37-0.49 (M), 0.36-0.46 (F)
Questionning unusual results
planning hemoglobin test
test: Hemoglobin (Hg) - 130-180 g/L (M), 120-160 g/L (F)
main tests were done, but nothing tells us the underlying precipitant
planning liver enzymes tests
test: Serum Liver Pancreatic Tests / Alanine Aminotransferas
test: Serum Liver Pancreatic Tests / Albumin - 31-43 g/L
test: Serum Liver Pancreatic Tests / Total Bilirubin - 0-17 micromol/L
test: Serum Liver Pancreatic Tests / Lipase - 10-140 U/L
test: Serum Liver Pancreatic Tests / Amylase - 53-123 U/L
test: Serum Non-Electrolytes / Lactate - 1 - 2 mmol/L
test: Thyroid Stimulating Hormone (TSH) - 0,5-5.0 mU/L
would need to order tests to check for other infections
looking for infection - planning septic workup, checking abdomen, physical, x-ray, ultrasound
checking for DKA infections
Looking for infection that could be precipitating diabetes
test: Urinalysis / Nitrites - Negative
no need for testing 2 hour plasma glucose in a case like this
test: Ultrasound / Abdominal Scan - No presence of gallstones
planning x-ray to eliminate typical and atypical pneumonias
test: Chest and content / Chest X - Ray - No bone, heart, or lung abnormalities seen
test: Chest and content / Chest X - Ray - No bone, heart, or lung abnormalities seen
could still check for infection with x-ray for DKA
planning urine analysis to eliminate urinary tract infection
test: Urinalysis / Leukocyte Esterase - Negative
test: Urinalysis / Leukocyte Esterase - Negative
set Hypothesis strength to 100%
not worth doing urine glucose really, summing up:diabetic ketoacidosis, kidneys compensenting well,
planning ECG to eliminate infarction and potassium emergency
test: Plain ECG - Normal
summarizing and reviewing clinical picture
change belief to 100%
submitting Hypothesis Diabetes Mellitus (type I) - 100%
submitting Hypothesis Diabetes Mellitus (type I) - 100%
submitting Hypothesis Diabetes Mellitus (type I) - 100%
submitting Hypothesis Diabetes Mellitus (type I) - 99%

APPENDIX J

Case 3 Categorized Episodes for Expert 1

E1	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	excessive sweating, hand tremors, and a sensation "as if her heart was racing."	anxiety attacks	valium
2	12 pound weight loss in two months	test: Triiodothyronine (T3) - 4.2 nmol/L	past medical history is unremarkable
3	test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L	test: Thyroxine (T4) - 89 pmol/L (free)	she's efebrile, her high blood pressure
4	test: Radionuclide Scan of Thyroid Gland - diffuse high uptake		

Case 3 Categorized Episodes for Expert 2

E2	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	episodes of excessive sweating, hand tremors, and a sensation "as if her heart was racing."	two months anxiety attacks	34 yrs old woman
2	12 pound weight loss in two months	feels very nervous for no particular reason	Valium she has been borrowing from her mother to "calm her nerves
3	test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L	panic attack is a very common diagnosis, but it is a diagnosis of exclusion	test: WBC (total) - 4.5-11.0 X 10E9 /L
4	test: Thyroxine (T4) - 89 pmol/L (free)	weight loss is a problem going against the anxiety or panic disorder	
5	test: Triiodothyronine (T3) - 4.2 nmol/L	summing up important elements: anxiety attacks, episodes she's having and weight loss	
6	test: Radionuclide Scan of Thyroid Gland - diffuse high uptake	It could be a pheocromocytome, but it's rare; hyperthyroid because of pattern of symptoms	
7	submitting Hypothesis Hyperthyroid (Grave's disease) - 100%	differential would include: drug abuse, alcohol withdrawal, cardia arrhythmia and panic disorder	
8		select Hypothesis Hyperthyroid (Grave's disease)	
9		pulse of 120, widened pulse pressure, ok RR and no temperature	
10		test: Plain ECG - Sinus tachycardia	
11		test: B - HCG - < 5.0 IU/L	

Case 3 Categorized Episodes for Expert 3

E3	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	healthy 34 yrs old woman	anxiety attacks for two months	test: Radionuclide Scan of Thyroid Gland - diffuse high uptake
2	feels nervous	Hypothesis 1: Grave's disease	
3	excessive sweating, hand tremors, and a sensation "as if her heart was racing."		
4	12 pound weight loss in two months		
5	tachycardic at 120		
6	a bit hypertensive		
7	test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L		
8	test: Thyroxine (T4) - 89 pmol/L (free)		
9	test: Triiodothyronine (T3) - 4.2 nmol/L		
10	test: Thyroid Stimulating Immunoglobulin Assay - Present		
11	No other test needed to submit diagnosis of Grave's disease		

Case 3 Categorized Episodes for Expert 4

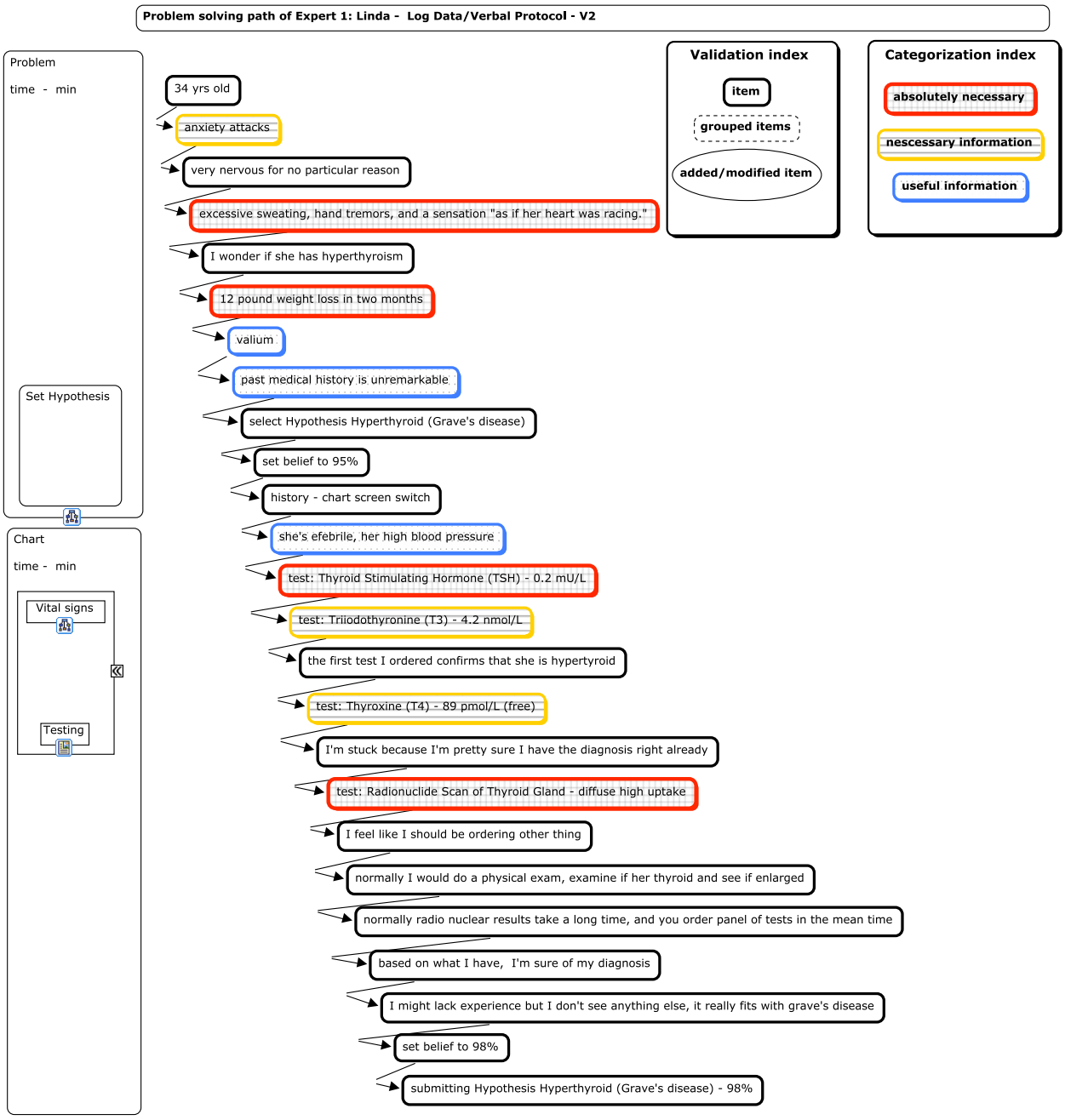
E4	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	symptom complex of sweating, tremors, ah palpitations, weight loss, and anxiety	alternative hypothesis: anxiety or other psychiatric diagnosis	young woman
2	select Hypothesis Hyperthyroid (Grave's disease)	alternative hypothesis: I'd wonder about non-prescribed drugs	otherwise healthy
3	summary of evidence	sweating	alternative hypothesis Pheo but hand tremors and lack of episodic symptoms don't fit
4	test: Thyroxine (T4) - 89 pmol/L (free)	tremors	low probability and if her blood pressure is normal I would not pursue
5	test: Triiodothyronine (T3) - 4.2 nmol/L	heart racing	would expect tachychardic and higher temperature
6	test: Thyroid Stimulating Hormone [TSH) - 0.2 mU/L	weight loss	
7	test: Thyroid Stimulating Immunoglobulin Assay - Present	anxiety	
8		vital signs: no high temperature, tachycardia and a little cystolic hypertension	

Case 3 Categorized Episodes for Expert 5

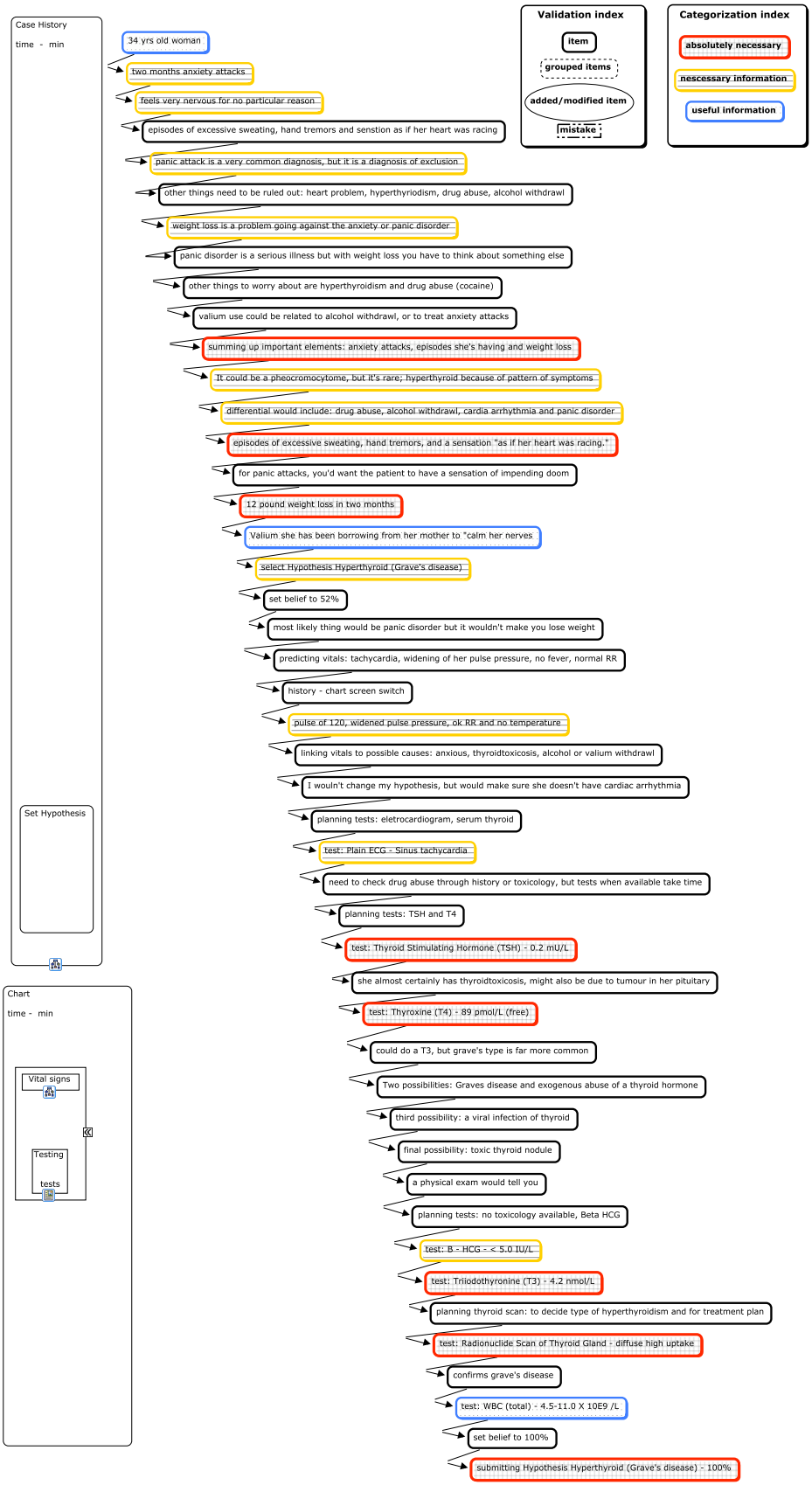
E5	Absolutely necessary (+5)	Necessary (+3)	Useful information (+1)
1	episodes of excessive sweating, hand tremors and heart racing	34 yr old woman	her concerns about the medications are in no way helpful
2	select Hypothesis Hyperthyroid (Grave's disease)	anxiety attacks suggests a psychosocial problem	
3	submitting Hypothesis Hyperthyroid (Grave's disease) - 90%	very nervous for no particular reason	
4		possible cardiac tachyarrhythmia	
5		12 pound weight loss is significant	
6		considering endocrinologic disorders	
7		select Hypothesis Panic Attack although anxiety would be a more serious consideration	
8		select Hypothesis Arrhythmia	
9		select Hypothesis Hyperthyroid (Grave's disease)	
10		select Hypothesis Pheochromocytoma	
11		main hypothesis remains psychosocial: select Hypothesis Panic Attack	
12		vital signs are relatively impressive: pulse rate at 120, blood pressure at 160	
13		ordinarily we do routine testing but given concerns about thyroid disease we should do T4 and T3	
14		test: Thyroxine (T4) - 89 pmol/L (free)	
15		test: Triiodothyronine (T3) - 4.2 nmol/L	
16		test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L	
17		test: Plain ECG - Sinus tachycardia	

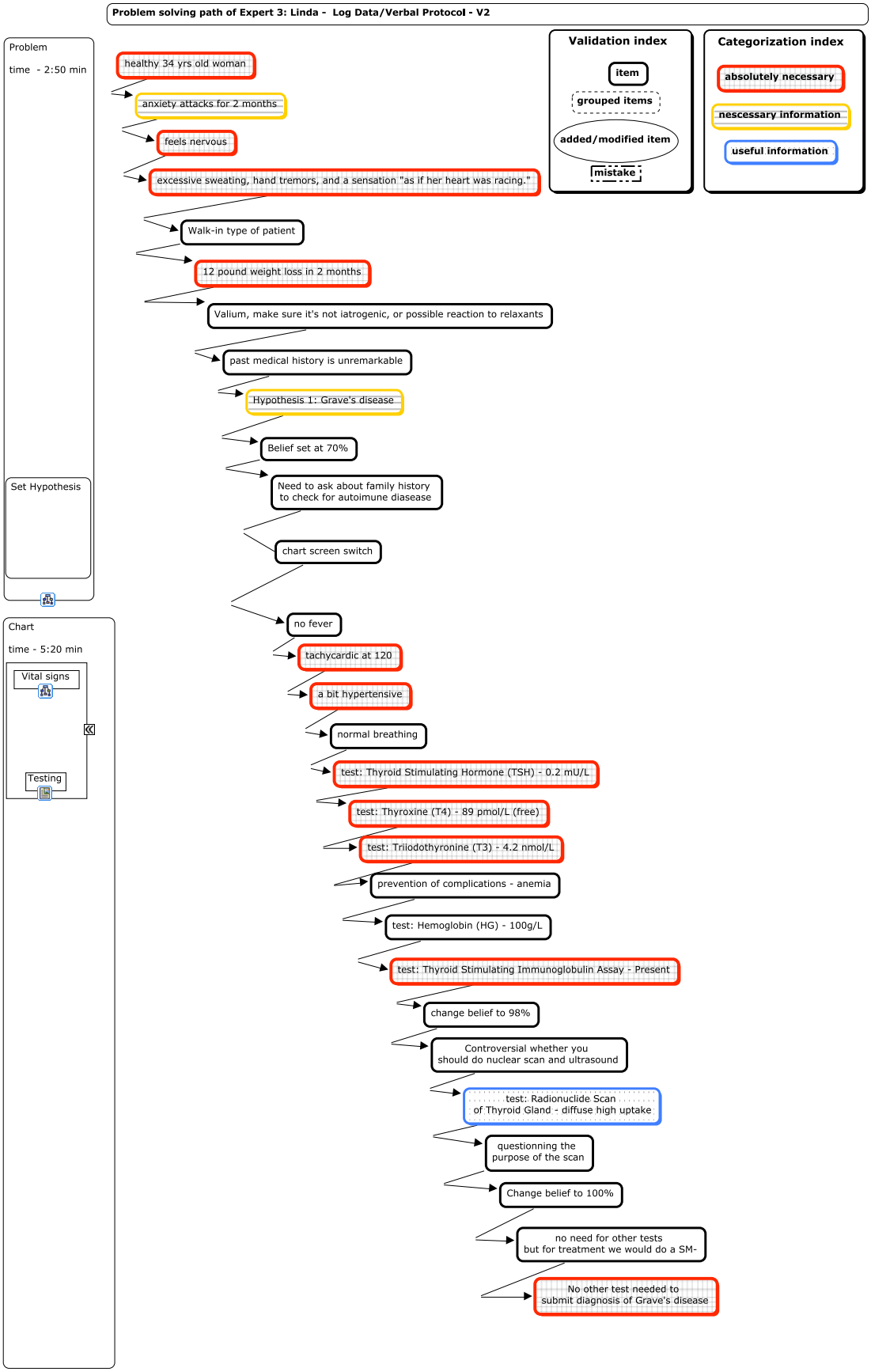
APPENDIX K

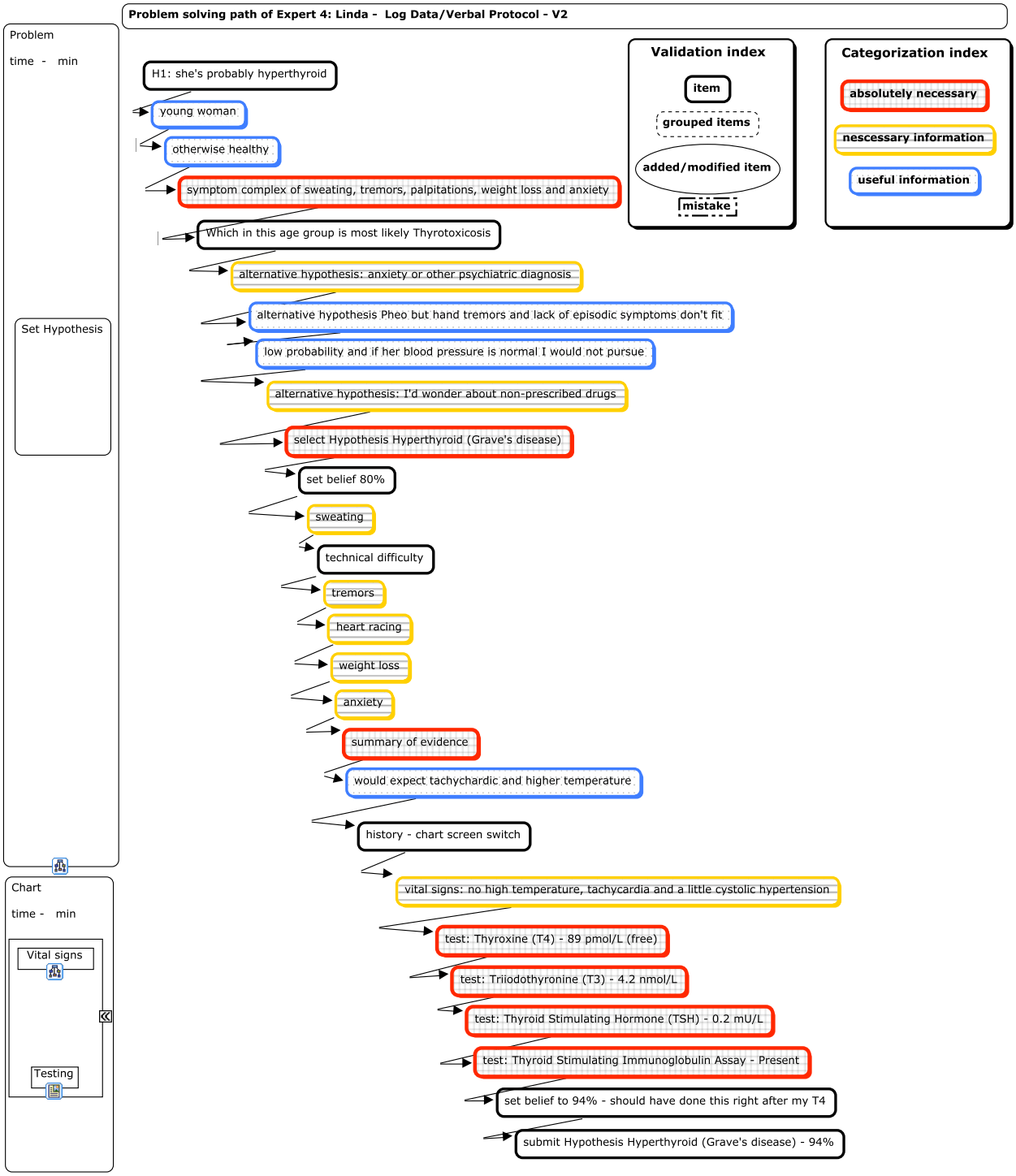
Individual Categorized Visual Representations for Case 3

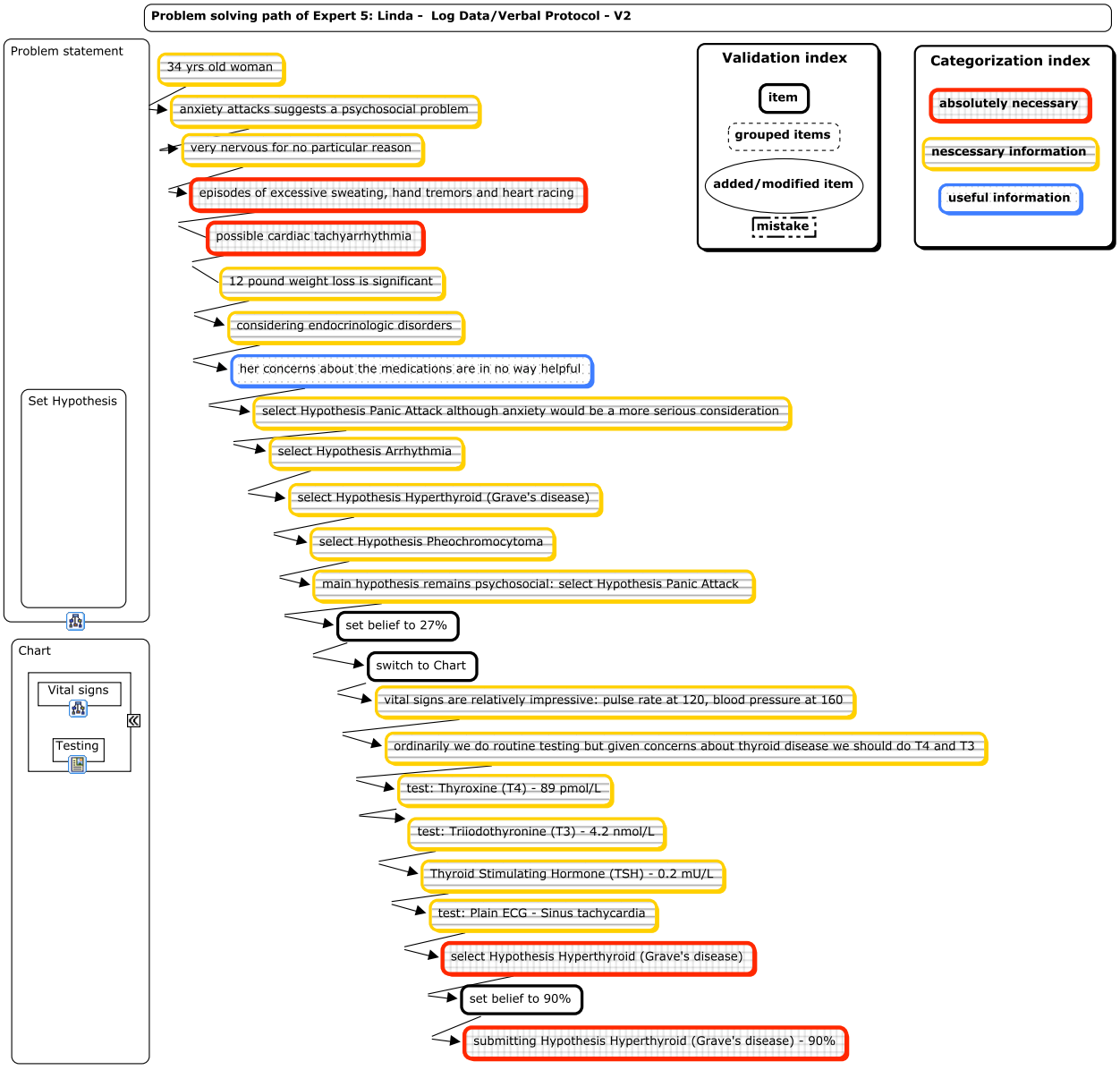

Problem solving path of Expert 1: Linda - Log Data/Verbal Protocol - V2
Problem
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
Validation index
item
grouped items
added/modified item
Categorization index
absolutely necessary
nescessary information
useful information
34 yrs old
anxiety attacks
very nervous for no particular reason
excessive sweating, hand tremors, and a sensation "as if her heart was racing."
I wonder if she has hyperthyroism
12 pound weight loss in two months
valium
past medical history is unremarkable
select Hypothesis Hyperthyroid (Grave's disease)
set belief to 95%
history - chart screen switch
she's efebrile, her high blood pressure
test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
test: Triiodothyronine (T3) - 4.2 nmol/L
the first test I ordered confirms that she is hypertyroid
test: Thyroxine (T4) - 89 pmol/L (free)
I'm stuck because I'm pretty sure I have the diagnosis right already
test: Radionuclide Scan of Thyroid Gland - diffuse high uptake
I feel like I should be ordering other thing
normally I would do a physical exam, examine if her thyroid and see if enlarged
normally radio nuclear results take a long time, and you order panel of tests in the mean time
based on what I have, I'm sure of my diagnosis
I might lack experience but I don't see anything else, it really fits with grave's disease
set belief to 98%
submitting Hypothesis Hyperthyroid (Grave's disease) - 98%

Problem solving path of Expert 2: Linda - Log Data/Verbal Protocol - version 2



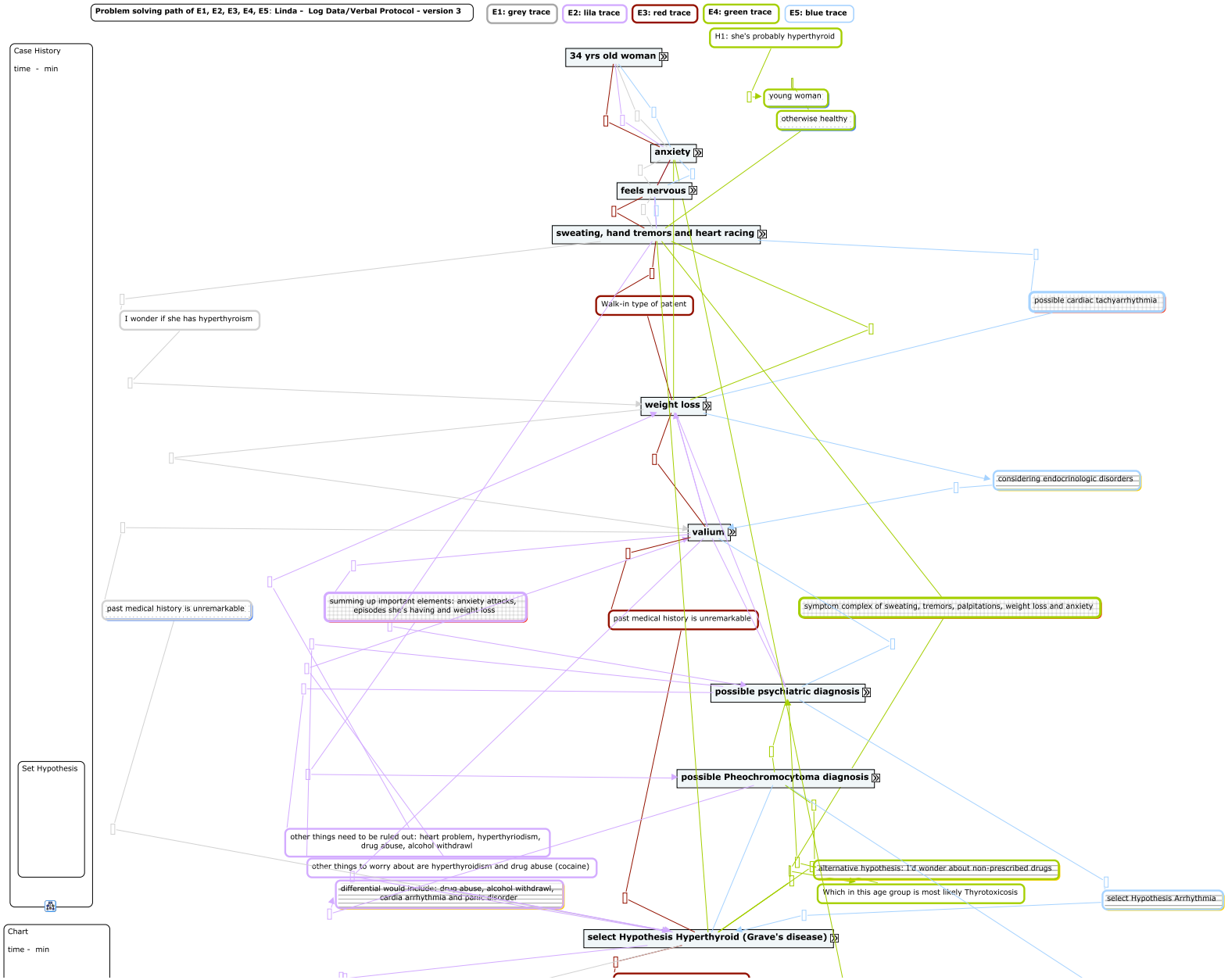

Problem solving path of Expert 3: Linda - Log Data/Verbal Protocol - V2
Problem
time - 2:50 min
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
healthy 34 yrs old woman
anxiety attacks for 2 months
feels nervous
excessive sweating, hand tremors, and a sensation "as if her heart was racing."
Walk-in type of patient
12 pound weight loss in 2 months
Valium, make sure it's not iatrogenic, or possible reaction to relaxants
past medical history is unremarkable
Hypothesis 1: Grave's disease
Belief set at 70%
Need to ask about family history to check for autoimune diasease
Set Hypothesis
chart screen switch
Chart
time - 5:20 min
Vital signs
no fever
tachycardic at 120
a bit hypertensive
normal breathing
Testing
test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
test: Thyroxine (T4) - 89 pmol/L (free)
test: Triiodothyronine (T3) - 4.2 nmol/L
prevention of complications - anemia
test: Hemoglobin (HG) - 100g/L
test: Thyroid Stimulating Immunoglobulin Assay - Present
change belief to 98%
Controversial whether you should do nuclear scan and ultrasound
test: Radionuclide Scan of Thyroid Gland - diffuse high uptake
questionning the purpose of the scan
Change belief to 100%
no need for other tests but for treatment we would do a SM-
No other test needed to submit diagnosis of Grave's disease

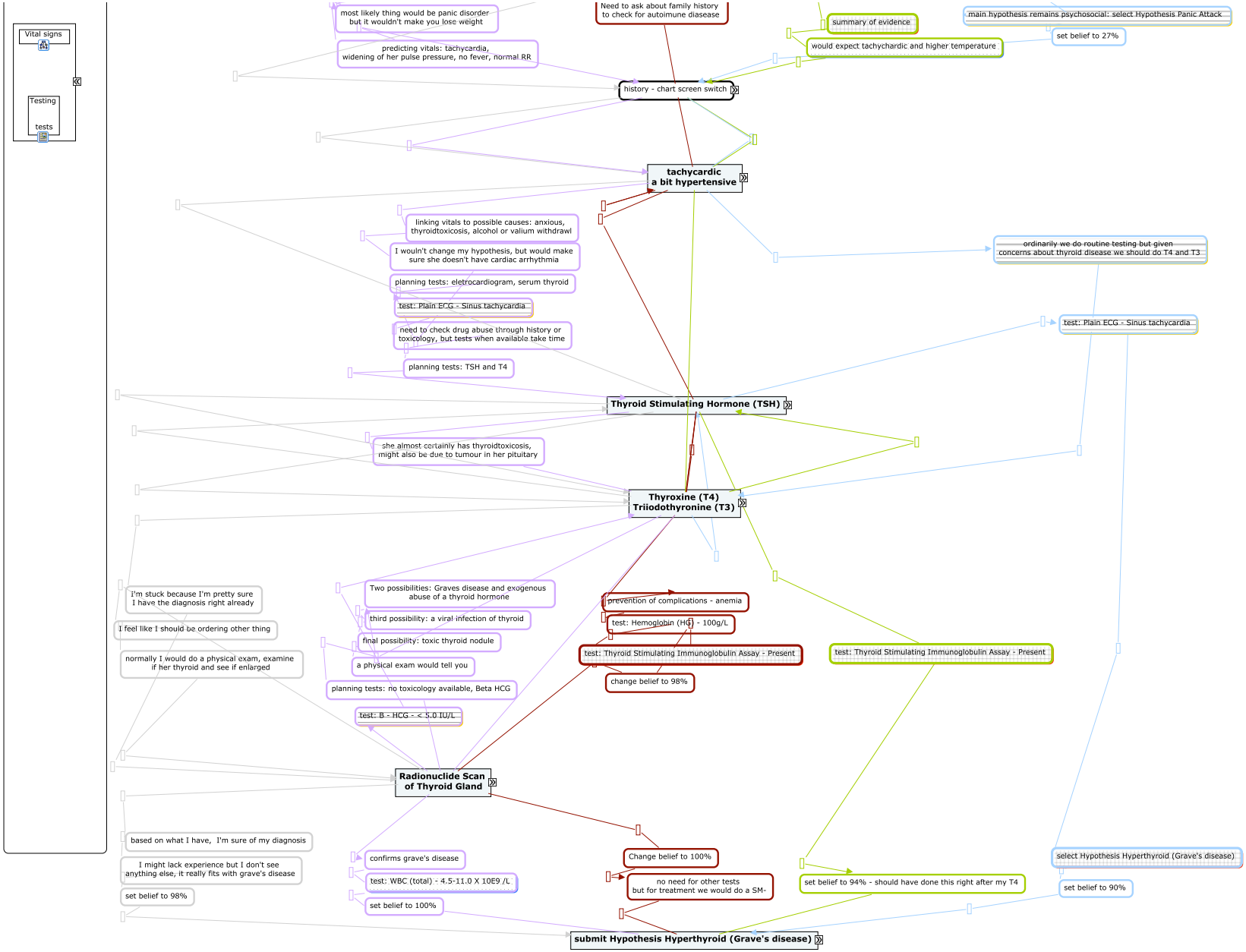

Problem solving path of Expert 4: Linda - Log Data/Verbal Protocol - V2
Problem
time - min
Set Hypothesis
Chart
time - min
Vital signs
Testing
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
H1: she's probably hyperthyroid
young woman
otherwise healthy
symptom complex of sweating, tremors, palpitations, weight loss and anxiety
Which in this age group is most likely Thyrotoxicosis
alternative hypothesis: anxiety or other psychiatric diagnosis
alternative hypothesis Pheo but hand tremors and lack of episodic symptoms don't fit
low probability and if her blood pressure is normal I would not pursue
alternative hypothesis: I'd wonder about non-prescribed drugs
select Hypothesis Hyperthyroid (Grave's disease)
set belief 80%
sweating
technical difficulty
tremors
heart racing
weight loss
anxiety
summary of evidence
would expect tachychardic and higher temperature
history - chart screen switch
vital signs: no high temperature, tachycardia and a little cystolic hypertension
test: Thyroxine (T4) - 89 pmol/L (free)
test: Triiodothyronine (T3) - 4.2 nmol/L
test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
test: Thyroid Stimulating Immunoglobulin Assay - Present
set belief to 94% - should have done this right after my T4
submit Hypothesis Hyperthyroid (Grave's disease) - 94%

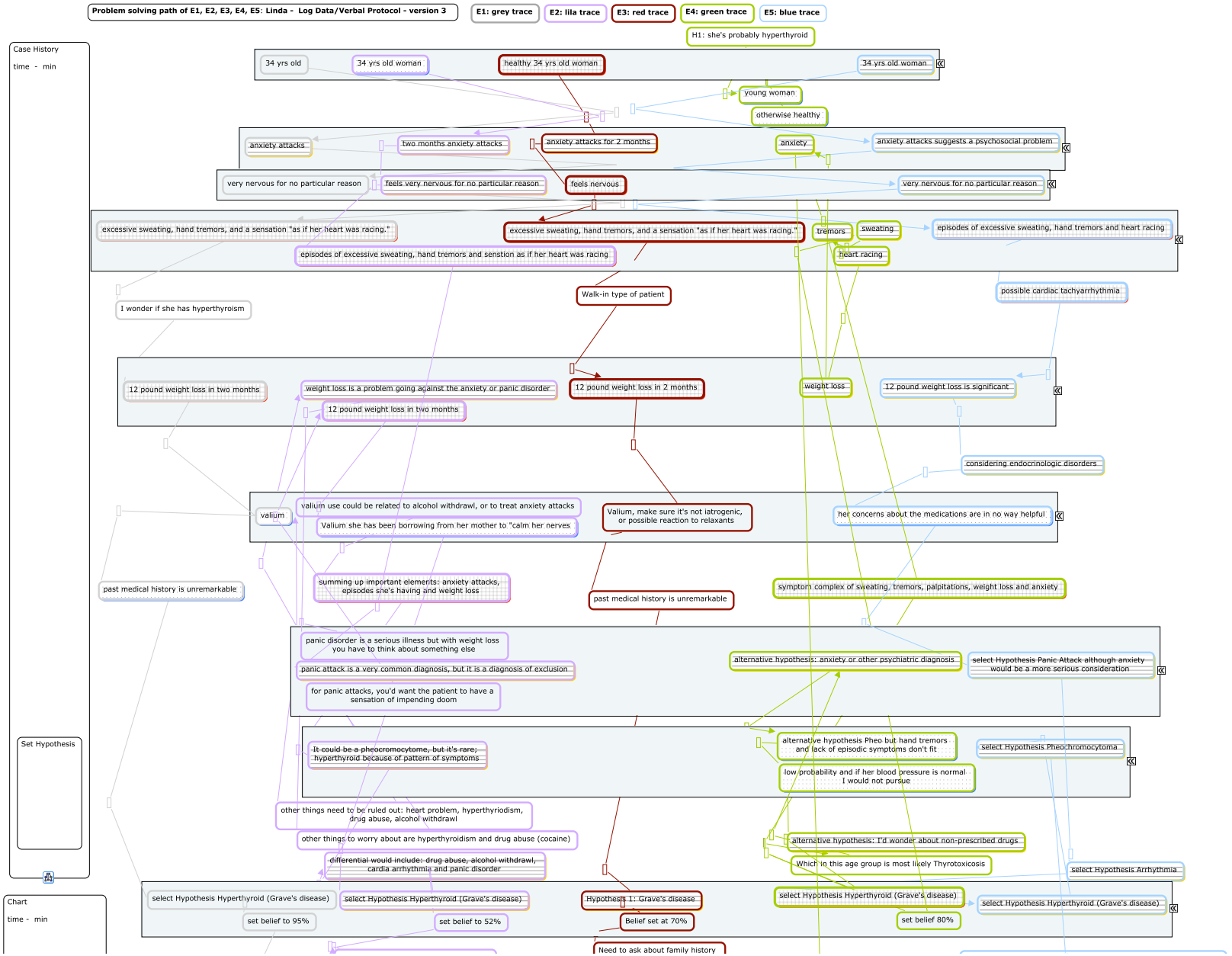

Problem solving path of Expert 5: Linda - Log Data/Verbal Protocol - V2
Problem statement
Set Hypothesis
Chart
Vital signs
Testing
Validation index
item
grouped items
added/modified item
mistake
Categorization index
absolutely necessary
nescessary information
useful information
34 yrs old woman
anxiety attacks suggests a psychosocial problem
very nervous for no particular reason
episodes of excessive sweating, hand tremors and heart racing
possible cardiac tachyarrhythmia
12 pound weight loss is significant
considering endocrinologic disorders
her concerns about the medications are in no way helpful
select Hypothesis Panic Attack although anxiety would be a more serious consideration
select Hypothesis Arrhythmia
select Hypothesis Hyperthyroid (Grave's disease)
select Hypothesis Pheochromocytoma
main hypothesis remains psychosocial: select Hypothesis Panic Attack
set belief to 27%
switch to Chart
vital signs are relatively impressive: pulse rate at 120, blood pressure at 160
ordinarily we do routine testing but given concerns about thyroid disease we should do T4 and T3
test: Thyroxine (T4) - 89 pmol/L
test: Triiodothyronine (T3) - 4.2 nmol/L
Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
test: Plain ECG - Sinus tachycardia
select Hypothesis Hyperthyroid (Grave's disease)
set belief to 90%
submitting Hypothesis Hyperthyroid (Grave's disease) - 90%

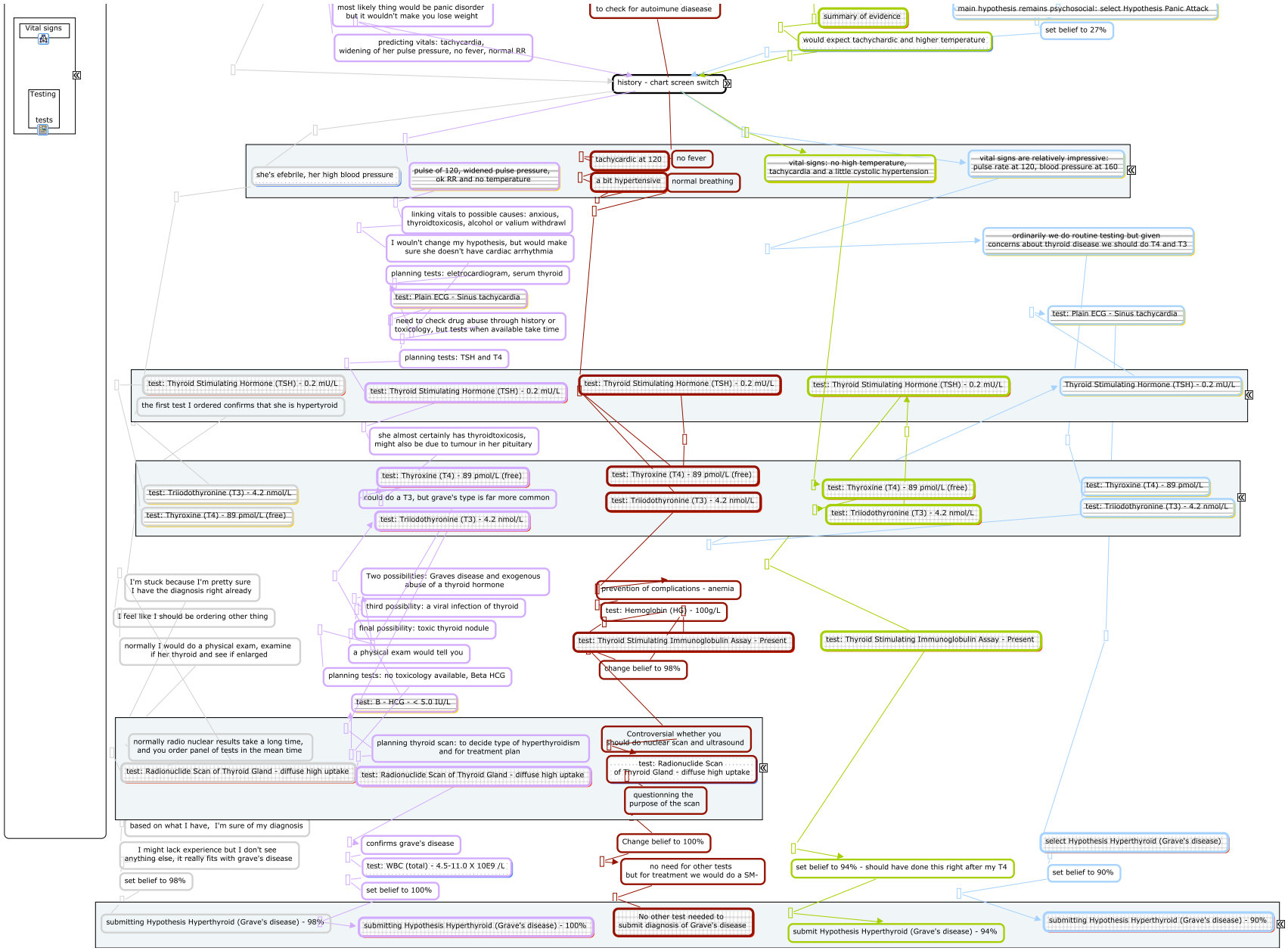
APPENDIX L

Merged Representations for Case 3 at Level 1 and Level 2


Problem solving path of E1, E2, E3, E4, E5: Linda - Log Data/Verbal Protocol - version 3
E1: grey trace
E2: lila trace
E3: red trace
E4: green trace
E5: blue trace
Case History
time - min
H1: she's probably hyperthyroid
34 yrs old woman
young woman
otherwise healthy
anxiety
feels nervous
sweating, hand tremors and heart racing
Walk-in type of patient
possible cardiac tachyarrhythmia
I wonder if she has hyperthyroism
weight loss
considering endocrinologic disorders
valium
past medical history is unremarkable
summing up important elements: anxiety attacks, episodes she's having and weight loss
past medical history is unremarkable
symptom complex of sweating, tremors, palpitations, weight loss and anxiety
possible psychiatric diagnosis
Set Hypothesis
possible Pheochromocytoma diagnosis
other things need to be ruled out: heart problem, hyperthyriodism, drug abuse, alcohol withdrawl
other things to worry about are hyperthyroidism and drug abuse (cocaine)
alternative hypothesis: I'd wonder about non-prescribed drugs
differential would include: drug abuse, alcohol withdrawl, cardia arrhythmia and panic disorder
Which in this age group is most likely Thyrotoxicosis
select Hypothesis Arrhythmia
Chart
time - min
select Hypothesis Hyperthyroid (Grave's disease)


Vital signs
Testing
tests
most likely thing would be panic disorder but it wouldn't make you lose weight
predicting vitals: tachycardia, widening of her pulse pressure, no fever, normal RR
Need to ask about family history to check for autoimune disease
summary of evidence
would expect tachychardic and higher temperature
main hypothesis remains psychosocial: select Hypothesis Panic Attack
set belief to 27%
history - chart screen switch
tachycardic a bit hypertensive
linking vitals to possible causes: anxious, thyroidtoxicosis, alcohol or valium withdrawl
I wouln't change my hypothesis, but would make sure she doesn't have cardiac arrhythmia
planning tests: eletrocardiogram, serum thyroid
test: Plain ECG - Sinus tachycardia
need to check drug abuse through history or toxicology, but tests when available take time
planning tests: TSH and T4
ordinarily we do routine testing but given concerns about thyroid disease we should do T4 and T3
test: Plain ECG - Sinus tachycardia
Thyroid Stimulating Hormone (TSH)
she almost certainly has thyroidtoxicosis, might also be due to tumour in her pituitary
Thyroxine (T4) Triiodothyronine (T3)
I'm stuck because I'm pretty sure I have the diagnosis right already
I feel like I should be ordering other thing
normally I would do a physical exam, examine if her thyroid and see if enlarged
Two possibilities: Graves disease and exogenous abuse of a thyroid hormone
third possibility: a viral infection of thyroid
final possibility: toxic thyroid nodule
a physical exam would tell you
planning tests: no toxicology available, Beta HCG
test: B - HCG - < 5.0 IU/L
prevention of complications - anemia
test: Hemoglobin (Hb) - 100g/L
test: Thyroid Stimulating Immunoglobulin Assay - Present
change belief to 98%
test: Thyroid Stimulating Immunoglobulin Assay - Present
Radionuclide Scan of Thyroid Gland
based on what I have, I'm sure of my diagnosis
I might lack experience but I don't see anything else, it really fits with grave's disease
set belief to 98%
confirms grave's disease
test: WBC (total) - 4.5-11.0 X 10E9 /L
set belief to 100%
Change belief to 100%
no need for other tests but for treatment we would do a SM-
set belief to 94% - should have done this right after my T4
select Hypothesis Hyperthyroid (Grave's disease)
set belief to 90%
submit Hypothesis Hyperthyroid (Grave's disease)


Problem solving path of E1, E2, E3, E4, E5: Linda - Log Data/Verbal Protocol - version 3
E1: grey trace
E2: lila trace
E3: red trace
E4: green trace
E5: blue trace
H1: she's probably hyperthyroid
Case History
time - min
34 yrs old
34 yrs old woman
healthy 34 yrs old woman
34 yrs old woman
young woman
otherwise healthy
anxiety attacks
two months anxiety attacks
anxiety attacks for 2 months
anxiety
anxiety attacks suggests a psychosocial problem
very nervous for no particular reason
feels very nervous for no particular reason
feels nervous
very nervous for no particular reason
excessive sweating, hand tremors, and a sensation "as if her heart was racing."
excessive sweating, hand tremors, and a sensation "as if her heart was racing."
tremors
sweating
episodes of excessive sweating, hand tremors and heart racing
episodes of excessive sweating, hand tremors and senstion as if her heart was racing
heart racing
Walk-in type of patient
possible cardiac tachyarrhythmia
I wonder if she has hyperthyroism
12 pound weight loss in two months
weight loss is a problem going against the anxiety or panic disorder
12 pound weight loss in 2 months
weight loss
12 pound weight loss is significant
12 pound weight loss in two months
considering endocrinologic disorders
valium
valium use could be related to alcohol withdrawl, or to treat anxiety attacks
Valium she has been borrowing from her mother to "calm her nerves
Valium, make sure it's not iatrogenic, or possible reaction to relaxants
her concerns about the medications are in no way helpful
past medical history is unremarkable
summing up important elemerits: anxiety attacks, episodes she's having and weight loss
past medical history is unremarkable
symptom complex of sweating, tremors, palpitations, weight loss and anxiety
panic disorder is a serious illness but with weight loss you have to think about something else
panic attack is a very common diagnosis, but it is a diagnosis of exclusion
for panic attacks, you'd want the patient to have a sensation of impending doom
alternative hypothesis: anxiety or other psychiatric diagnosis
select Hypothesis Panic Attack although anxiety would be a more serious consideration
Set Hypothesis
It could be a pheocromocytome, but it's rare; hyperthyroid because of pattern of symptoms
alternative hypothesis Pheo but hand tremors and lack of episodic symptoms don't fit
select Hypothesis Pheochromocytoma
low probability and if her blood pressure is normal I would not pursue
other things need to be ruled out: heart problem, hyperthyriodism, drug abuse, alcohol withdrawl
other things to worry about are hyperthyroidism and drug abuse (cocaine)
alternative hypothesis: I'd wonder about non-prescribed drugs
differential would include: drug abuse, alcohol withdrawl, cardia arrhythmia and panic disorder
Which in this age group is most likely Thyrotoxicosis
select Hypothesis Arrhythmia
select Hypothesis Hyperthyroid (Grave's disease)
select Hypothesis Hyperthyroid (Grave's disease)
Hypothesis 1: Grave's disease
select Hypothesis Hyperthyroid (Grave's disease)
select Hypothesis Hyperthyroid (Grave's disease)
set belief to 95%
set belief to 52%
Belief set at 70%
set belief 80%
Need to ask about family history
Chart
time - min


Vital signs
Testing
tests
most likely thing would be panic disorder but it wouldn't make you lose weight
predicting vitals: tachycardia, widening of her pulse pressure, no fever, normal RR
to check for autoimune diasease
summary of evidence
would expect tachychardic and higher temperature
main hypothesis remains psychosocial: select Hypothesis Panic Attack
set belief to 27%
history - chart screen switch
she's efebrile, her high blood pressure
pulse of 120, widened pulse pressure, ok RR and no temperature
tachycardic at 120
no fever
a bit hypertensive
normal breathing
vital signs: no high temperature, tachycardia and a little cystolic hypertension
vital signs are relatively impressive: pulse rate at 120, blood pressure at 160
linking vitals to possible causes: anxious, thyroidtoxicosis, alcohol or valium withdrawl
I wouln't change my hypothesis, but would make sure she doesn't have cardiac arrhythmia
planning tests: eletrocardiogram, serum thyroid
test: Plain ECG - Sinus tachycardia
need to check drug abuse through history or toxicology, but tests when available take time
ordinarily we do routine testing but given concerns about thyroid disease we should do T4 and T3
test: Plain ECG - Sinus tachycardia
planning tests: TSH and T4
test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
the first test I ordered confirms that she is hypertyroid
test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
test: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
test:: Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
Thyroid Stimulating Hormone (TSH) - 0.2 mU/L
she almost certainly has thyroidtoxicosis, might also be due to tumour in her pituitary
test: Triiodothyronine (T3) - 4.2 nmol/L
test: Thyroxine (T4) - 89 pmol/L (free)
test: Thyroxine (T4) - 89 pmol/L (free)
could do a T3, but grave's type is far more common
test: Triiodothyronine (T3) - 4.2 nmol/L
test: Thyroxine (T4) - 89 pmol/L (free)
test: Triiodothyronine (T3) - 4.2 nmol/L
test: Thyroxine (T4) - 89 pmol/L (free)
test: Triiodothyronine (T3) - 4.2 nmol/L
test: Thyroxine (T4) - 89 pmol/L
test: Triiodothyronine (T3) - 4.2 nmol/L
I'm stuck because I'm pretty sure I have the diagnosis right already
I feel like I should be ordering other thing
normally I would do a physical exam, examine if her thyroid and see if enlarged
Two possibilities: Graves disease and exogenous abuse of a thyroid hormone
third possibility: a viral infection of thyroid
final possibility: toxic thyroid nodule
a physical exam would tell you
planning tests: no toxicology available, Beta HCG
test: B - HCG - < 5.0 IU/L
prevention of complications - anemia
test: Hemoglobin (Hb) - 100g/L
test: Thyroid Stimulating Immunoglobulin Assay - Present
change belief to 98%
test: Thyroid Stimulating Immunoglobulin Assay - Present
normally radio nuclear results take a long time, and you order panel of tests in the mean time
test: Radionuclide Scan of Thyroid Gland - diffuse high uptake
planning thyroid scan: to decide type of hyperthyroidism and for treatment plan
test: Radionuclide Scan of Thyroid Gland - diffuse high uptake
Controversial whether you should do nuclear scan and ultrasound
test: Radionuclide Scan of Thyroid Gland - diffuse high uptake
questionning the purpose of the scan
based on what I have, I'm sure of my diagnosis
I might lack experience but I don't see anything else, it really fits with grave's disease
set belief to 98%
confirms grave's disease
test: WBC (total) - 4.5-11.0 X 10E9 /L
set belief to 100%
Change belief to 100%
no need for other tests but for treatment we would do a SM-
set belief to 94% - should have done this right after my T4
select Hypothesis Hyperthyroid (Grave's disease)
set belief to 90%
submitting Hypothesis Hyperthyroid (Grave's disease) - 98%
submitting Hypothesis Hyperthyroid (Grave's disease) - 100%
No other test needed to submit diagnosis of Grave's disease
submit Hypothesis Hyperthyroid (Grave's disease) - 94%
submitting Hypothesis Hyperthyroid (Grave's disease) - 90%