Published in final edited form as:
Artif Intell Med. 2014 October ; 62(2): 91–104. doi:10.1016/j.artmed.2014.08.002.

Evaluating the effects of cognitive support on psychiatric clinical comprehension

Venkata V. Dalai[a], **Sana Khalid**[b], **Dinesh Gottipati**[a], **Thomas Kannampallil**[b], **Vineeth John**[c], **Brett Blatter**[d], **Vimla L. Patel**[b], and **Trevor Cohen**[a,*]

[a]National Center for Cognitive Informatics and Decision Making in Healthcare, School of Biomedical Informatics, University of Texas Health Science Center, 7000 Fannin St, Houston, Texas 77030

[b]Center for Cognitive Studies in Medicine and Public Health, The New York Academy of Medicine, 1216 5th Avenue, New York, New York 10029

[c]Psychiatry and Behavioral Sciences, University of Texas Medical School at Houston, University of Texas Health Science Center, 6341 Fannin St, Houston, Texas 77030

[d]Psychiatric Emergency Services, New York Presbyterian Hospital, 622 West 168th St, New York, New York 10032

Abstract

Objective—Clinicians' attention is a precious resource, which in the current healthcare practice is consumed by the cognitive demands arising from complex patient conditions, information overload, time pressure, and the need to aggregate and synthesize information from disparate sources. The ability to organize information in ways that facilitate the generation of effective diagnostic solutions is a distinguishing characteristic of expert physicians, suggesting that automated systems that organize clinical information in a similar manner may augment physicians' decision-making capabilities. In this paper, we describe the design and evaluation of a theoretically driven cognitive support system (CSS) that assists psychiatrists in their interpretation of clinical cases. The system highlights, and provides the means to navigate to, text that is organized in accordance with a set of diagnostically and therapeutically meaningful higher-level concepts.

Methods and Materials—To evaluate the interface, 16 psychiatry residents interpreted two clinical case scenarios, with and without the CSS. Think-aloud protocols captured during their interpretation of the cases were transcribed and analyzed qualitatively. In addition, the frequency and relative position of content related to key higher-level concepts in a verbal summary of the

*Corresponding author: Trevor Cohen, PhD, Associate Professor, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX 77030, trevor.cohen@uth.tmc.edu, Phone: 713-486-3675, Fax: 713-486-0117.

case were evaluated. In addition the transcripts from both groups were compared to an expert derived reference standard using latent semantic analysis (LSA).

Results—Qualitative analysis showed that users of the system better attended to specific clinically important aspects of both cases when these were highlighted by the system, and revealed ways in which the system mediates hypotheses generation and evaluation. Analysis of the summary data showed differences in emphasis with and without the system. The LSA analysis suggested users of the system were more “expert-like” in their emphasis, and that cognitive support was more effective in the more complex case.

Conclusions—Cognitive support impacts upon clinical comprehension. This appears to be largely helpful, but may also lead to neglect of information (such as the psychosocial history) that the system does not highlight. The results have implications for the design of CSSs for clinical narratives including the role of information organization and textual embellishments for more efficient clinical case presentation and comprehension.

Keywords

Biomedical Informatics; Cognitive Science; Clinical Comprehension; Cognitive Support; Latent Semantic Analysis; Propositional Analysis; Verbal Protocol Analysis; Psychiatry; Emergency Psychiatry

1 Introduction

In complex clinical environments, clinicians must cope with and manage multiple, voluminous, heterogeneous data sources to solve clinical problems [1, 2]. Both comprehension and problem solving capabilities of physicians affect their efficiency, as comprehension is a prerequisite to problem solving [3]. Previous studies have suggested that the process of clinical comprehension differs between expert and novice clinicians with respect to selective filtering, pattern recognition and accuracy of inferences generated [4]. Specifically, experts use knowledge structures called “intermediate constructs” that represent clinically meaningful clusters of observations that lead toward specific diagnoses. The ability to generate intermediate constructs is a distinguishing characteristic of expert clinical comprehension [5]. In contrast, non-experts (e.g., residents) and other trainees may possess a less organized, albeit large, knowledge base.

It has been argued that the application of information technology to simulate aspects of expert comprehension in order to provide cognitive support may allow trainees to reason in an expert-like manner [6]. Therefore, a cognitive support system (CSS) that organizes the information in a manner that mediates efficient problem solving may improve the quality and efficiency of patient care. While we have chosen the narratively rich clinical specialty of psychiatry as our problem domain, the problem we describe is related to human information processing in general. As such, this work has implications for the organization of information in any knowledge-intensive domain.

In this paper, we describe the development and evaluation of a CSS based on intermediate constructs. The problem solving processes of users of this interface are characterized, and compared to those of users of another interface without cognitive support. The interface and

its evaluation provide insights for the design of technology that can help clinicians organize information in a manner conducive to efficient decision-making.

2 Background

Experts have the ability to perceive the features of a problem that are most pertinent to its solution [7]. Seminal research from the chess domain showed that expert players are distinguished by their ability to recognize and reconstruct strategically meaningful configurations of chess pieces [8]. Similar studies conducted in various other fields of medicine such as radiology [9] and dermatology [10], demonstrated the expert's pattern recognition ability, especially in visually-oriented domains. Analogously, it has been found that expert physicians were proficient at recognizing diagnostically relevant patterns of symptoms in a clinical narrative [11], where information is presented verbally rather than visually.

Patel and Groen identified three important characteristics that differentiate experts from non-experts [11]. The *first* characteristic is a pattern of reasoning. In routine problems, experts use a data-driven pattern of reasoning where observations pertinent to problem data lead to an accurate diagnostic hypothesis, often progressing through pre-diagnostic hypotheses (e.g., "a cardiac problem") before reaching a final diagnosis (e.g., "left ventricular failure secondary to a myocardial infarction"). In contrast, non-experts and experts in unfamiliar situations use a hypothesis-driven pattern of reasoning, where a hypothesis, or set of hypotheses, guides data collection and interpretation.

The *second* characteristic that differentiates experts from non-experts is the organization of their knowledge base. Experts have a highly organized knowledge base that allows them to partition a problem into manageable "chunks." In the context of diagnostic reasoning, these "chunks" consist of intermediate constructs – diagnostically meaningful clusters of signs and symptoms that are not in and of themselves diagnoses, but serve to partition the diagnostic problem space and lead the way toward a correct diagnosis [12]. The recognition of a cardiac problem before reaching a more specific diagnosis is an example of the application of an intermediate construct. As an example drawn from the domain of psychiatry, psychotic symptoms such as hallucinations and delusions would be considered components of an intermediate construct indicating a psychotic episode. The organization of clinical findings into intermediate constructs provides a support structure for the ultimate diagnosis. While trainees may have large knowledge bases, these tend to be less organized than an expert's knowledge base. This may lead to the generation of diagnostic hypotheses without adequate supporting evidence. The *third* characteristic is the approach to a clinical problem. Experts typically generate a small set of relevant diagnoses at a high level of abstraction and quickly narrow down to the most accurate one, while non-experts tend to generate a large number of irrelevant diagnostic hypotheses [11].

Sharda and colleagues investigated the effect of expertise on comprehension of psychiatric narratives [13]. They found differences in knowledge organization between experts and non-experts. Experts approached a diagnostic solution using relevant intermediate constructs, while non-experts failed to generate key constructs, a finding consistent with those obtained

in other clinical domains [8, 9]. This raises the question of how the explicit presentation of intermediate constructs may affect clinical reasoning. This question motivates the current work, in which we evaluate the effects of such an interface on clinical comprehension and diagnostic reasoning.

3 A CSS for psychiatry

In this section, we describe a prototype user interface that presents psychiatric narrative in a manner conducive to the recognition of key intermediate constructs. In contrast to traditional decision support systems that seek to emulate expert performance of a decision-making task, this system supports the decision-making process at a point that is proximal to the decision itself. The basis for this design is the thought process of experts, as revealed through cognitive methods for the study of comprehension [4].

This approach is motivated by the theory of distributed cognition [14], which views cognition as the product of a distributed system involving both human actors and the external media that support them in their cognitive tasks. Rather than being confined to the mind of a single clinician, clinical comprehension can be viewed as a distributed process involving, for example, a human reader and a textual display (See Figure 1). Comprehension involves the construction of a mental representation of a clinical case that is influenced by structured knowledge stored in the mind of the clinician [5]. By organizing the information presented in accordance with a simulation of the structure of expert knowledge, a system can redistribute part of the cognitive work of expert comprehension from man to machine.

3.1 System description

We provide a brief account of the system design and development, but refer the interested reader to [6, 15] for further details of the development and evaluation of the back end of the system, which provides the means to draw associations between short segments of clinical narrative and a set of four diagnostically and/or prognostically relevant intermediate constructs, “psychosis”, “mood”, “substance abuse” and “dangerousness.” We refer to these constructs as “facets” in accordance with terminology developed in [16]. These facets were selected based on their clinical importance for patient assessment in emergency psychiatry. The selection of facets on this basis was informed by discussion with author BB, an expert in the domain of emergency psychiatry, as well as by our observation of an emergency psychiatry unit during the course of qualitative research conducted prior to the commencement of this project. For a detailed description of the unit concerned, we refer the interested reader to [6]. To link text in a discharge summary to each of these facets, we used a combination of latent semantic analysis (LSA) [17] and a training mechanism motivated by the conceptual spaces framework proposed by Peter Gardenfors [18]. LSA provides the means to derive high-dimensional semantic vector representations of terms from large text corpora such that the representations of terms that are semantically related occur close to one another in the semantic vector space. The conceptual spaces framework provides a geometric interpretation of conceptual categorization based on region connection calculus. The learning mechanism that underlies the system learns regions of the LSA semantic vector space that correspond to the facet models from positive and negative training examples. While a detailed description of this learning mechanism is beyond the scope of the current

paper, we refer the interested reader to [6, 15]. This paper also includes an evaluation of reliability of automated categorization using this approach, which a test set of 100 previously unseen psychiatry discharge summaries was annotated by two psychiatry residents. These summaries were segmented into phrases, and each phrase was annotated for its relevance to each of the five facets using a web-based annotation system developed for this purpose. The system agreed with the annotators in the majority of cases, with an average F-measure of 0.86.

For the current research, the LSA space and trained facet model representations that were used for our previous research were utilized to highlight phrases within the case scenarios. As was the case previously, phrases were represented as the vector average (or normalized sum) of the vectors representing their component terms, aside from terms that occur on the stopword list distributed with the General Text Parser package [19], which was used to derive the semantic space in our previous research. If the vector representation for the phrase as a whole, or any of the vector representations of component terms fell within the region of semantic space demarcated by a trained facet model representation, it was highlighted as relevant to the facet concerned.

These and other aspects of this work, including a tool created to perform propositional segmentation (an aspect of an analysis of natural language known as propositional analysis involving the segmentation of text into units corresponding approximately to one concept-relation-concept triplet, or proposition) are integrated into a system that performs three tasks [6, 15]: (1) propositional segmentation of discharge summary text; (2) association between text segments and trained facet models; and (3) generation of HTML code to produce a series of frame-based web pages that provide facet-specific views of discharge summaries.

3.2 Interface elements

The discharge summary used to illustrate the components of interface is one of two clinical cases created by Sharda and his colleagues, which were used in the study of expert and novice differences in comprehension previously [13], and are used for the studies documented in this paper. The various components of the interface are presented in Figure 2.

The facet tabs in this figure are situated at the top of the browser window, and provide a way to switch between facet-specific views of the same summary (e.g., between “psychosis” and “danger”). Clicking on one of these tabs switches the perspective to emphasize a particular facet. In the illustration in Figure 2, text elements relevant to the facet model for psychosis are emphasized by (1) increased font size and (2) facet-specific color-coding. The graphical summary on the top right hand corner provides an overview of the amount of content in this summary that the system has linked to each of the four facet models, using a bar graph. The facet-specific summary, in the right-most frame, contains a list of all of the text segments associated with a particular facet, broken down by discharge summary section.

4 Method

4.1 Study setting and participants

Sixteen (*N*=16) third- and fourth-year residents at an academic psychiatry program participated in the study. Participants were recruited with the help of a clinical collaborator and were given a $10 gift certificate for their participation. The study was conducted in the resident offices and took about an hour per participant to complete. The institutional review board (IRB) approved the study and written consents were obtained from all participants.

4.2 Study design and materials

We used a 2×2 mixed design study with two factors: *Interface type* (with intermediate constructs: IC, without intermediate constructs: No-IC) and *Case complexity* (simple, complex) where the interface type was a between-subjects variable and case complexity was a within-subjects repeated measure. Participants were randomly assigned to one of two interface types and completed two clinical cases (i.e., simple and complex). The order of the cases was counterbalanced to mitigate learning effects.

Interface type—We used two interface types: IC, the intermediate construct-based interface shown in Figure 2, and No-IC, an interface that followed a narrative style without any embellishments.

As previously described, the IC-based interface divides the text into segments, and assigns relevant segments to one of the four intermediate constructs (“psychosis”, “mood”, “substance” and “danger”) automatically (See Figure 2). The elements of the text that were relevant to one of the four intermediate constructs were highlighted in the text, and appeared in the frame on the right. The top frame of the interface indicates the extracted features of all four intermediate constructs for comparison and review. Clicking on one of these brings the focus to the point in the text at which this feature occurs, and highlights all elements of the text deemed relevant to this intermediate construct. The No-IC interface presents the case narrative in a browser without any text embellishments.

Case complexity—Two levels of case complexity, simple and complex, were used. The variation in the complexity relates to the number of clinical conditions that were represented in the narrative. For example, the patient presented in the complex case had significantly more past diagnostic conditions that were enmeshed within a web of social and emotional issues. In other words, for the complex case, development of a diagnosis and management plan would require a more exhaustive exploration of the diagnostic problem space, thereby imposing more cognitive load on the physician.

Two hypothetical case scenarios that were developed for the purpose of previous research were used in our study. The case scenarios were based on clinical cases in the DSM-IV casebook [13], an educational resource that aims to teach psychiatric diagnosis. The case scenarios were converted to text narratives, each describing a fictitious yet realistic psychiatric case.

Case 1 is relatively simple. The case describes a 27-year-old female with a past diagnosis of psychotic depression. In contrast to this diagnosis, her history suggests that she has experienced manic episodes. Given this combination of mood and psychotic symptoms, bipolar disorder and schizoaffective disorder are reasonable differential diagnoses for this case. However, the correct diagnosis ultimately is schizoaffective disorder as mood and psychotic symptoms overlap; psychotic symptoms occur independently (rather than as secondary symptoms in the context of a manic or depressive episode only); and mood symptoms are an enduring feature of the case.

Case 2 was designed to be considerably more complex than Case 1. The case describes a 29-year-old female with many past diagnoses, including bipolar disorder with psychotic features, schizoaffective disorder, post-traumatic stress disorder, and borderline personality disorder. The history of multiple diagnoses adds a layer of complexity to the case. The key to accurately diagnosing the case is recognizing that the haphazard distribution of signs and symptoms do not correspond to any recognized psychiatric category when viewed together, as this patient is malingering for the purpose of personal gain.

4.3 Procedure

A researcher provided the participants with an overview of the study, and obtained written consent. The participants were also provided with instructions on the use of the interface (in both IC and No-IC cases). Participants in the IC condition in addition watched a video tutorial describing the features of the interface and how it can be used. Following this, they were asked to complete a practice case.

Participants were then instructed to think aloud as they read the presented case with the intention to generate a diagnosis and construct a management plan. Thinking aloud involves verbalizing thoughts regarding the case without editing or interpreting them [20]. For example, several participants described the presenting condition (e.g., “this is a 27-year-old female patient with a past history of…”) and, in some cases, the rationale as to why this information was considered important. Verbal think-aloud techniques have been widely used in biomedical informatics research (for a review see [21]) and are considered to be an effective approach toward understanding human cognition and reasoning. Think-aloud data (in the form of verbal protocols) gathered during the task of reading through the case reflect the nature of the cognitive processes occurring during clinical comprehension.

After completing the case evaluation, participants provided a diagnosis, case summary and management plan. This part of the process was conducted based on their memory of the case – participants were not given access to the case itself during this task. Consequently this part of the think-aloud protocol reflects the representation of the case that was generated by the process of clinical comprehension that was completed previously. After completing each case, all participants filled out the NASA-TLX workload survey [22]. They also filled out a system usability survey (SUS) [23] after completing each case. At the end of the experiment, the participants were de-briefed and provided a $10 gift certificate for their participation.

4.4 Data collection

We collected the verbal think-aloud of the participants as they worked through both their assigned cases, time required for case completion, diagnostic accuracy, workload measures using the NASA-TLX survey and the usability measures using the SUS scale. Think-aloud sessions were audio-recorded and were later transcribed verbatim for analysis. Additionally, the use of the interface was captured using the Techsmith Morae screen capture tool.

4.5 Data analysis

4.5.1 Overview of analytic methods—The transcribed audio recordings included both transcriptions of verbalizations captured during the process of reading through the case (think-aloud data) and transcriptions of their verbal case summaries at the conclusion of this task. The think-aloud data reflect the process of comprehension, and reveal foci of attention and interactions with the system during interpretation of the case. The mental result of this process is a case representation, which informed the spoken summary.

We analyzed the think-aloud data using qualitative methods. We evaluated whether set of diagnostically relevant elements identified during the course of previous research had been mentioned by each participant. In addition, for the participant that made the most extensive use of the system, we provide a detailed analysis of their reasoning and interaction with the system. Our analysis of the summary data aimed to characterize the way in which participants' mental representations of the case were organized at facet level. To achieve this aim, we manually assigned the case elements in their summaries to a set of clinically relevant facets, and evaluated the prevalence and sequential organization of elements relevant to each of these facets. We also reviewed their diagnoses and management plans. Finally, we performed an evaluation of the relatedness between each participant's entire transcript (think-aloud and summary) and a canonical model of the relevant aspects of the case derived from expert verbal protocols. This last analysis was performed using LSA.

4.5.2 Analysis of think-aloud data: divergent recall—In previous work using the same clinical case scenarios, Sharda and his colleagues highlighted nine points of "divergent recall" [13], clinically relevant propositions that both expert participants had mentioned in their think-aloud protocols, but one or both non-expert participants had failed to mention. This analysis involved the verbal protocol captured as the participant read through the case concerned ("recall" here refers to mention of an element of the case during the think aloud protocol, rather than a memory-related task), and as such can be considered as indicative of the focus of participant attention during construction of their mental representation of the case. We applied the same methodology to our think-aloud protocols, reviewing the protocols to see if each of these nine points appeared in the recall protocol of each of our participants.

4.5.3 Analysis of think-aloud data: qualitative analysis of patterns of navigation—In order to characterize the ways in which the system mediated problem solving, we performed a qualitative analysis of the think-aloud protocols and video screen capture data generated by selected participants during the course of their problem-solving process. Combining these data sources enabled us to characterize both the actions taken

(such as selection of a particular facet-level perspective, or navigation to a specific facet-related element in the text) and the thought process underlying these actions.

While the degree of usage differed across participants, all participants in the IC group used the interface to some degree – some participants used the presented facets to review the case after reading through it and others reviewed the facets ahead of reading the case narrative. To provide a more granular account of the ways in which the interface supported clinical comprehension, we performed a detailed qualitative analysis of the verbal protocol of the participant in the IC group who used the system most extensively.

Using video analysis we identified instances of interface use, and selected events where a mouse-click led to the navigation from one facet to another facet, or from text narrative to facet findings. Also, we used the think-aloud protocols as additional data to infer the purpose underlying the actions we observed. The patterns of navigation on the interfaces were analyzed using a combination of screen capture and verbal data. One of the researchers (VVD) conducted this part of the analysis and was verified by another researcher [14]. Any disagreements were resolved through discussion.

4.5.4 Analysis of summary data: identification of facets—In order to assess the extent to which the clinical summaries were organized in accordance with clinically relevant facets, it was necessary to identify how each unit of information in the summaries relates to the key facets in both cases. To achieve this, all verbal think-aloud data were transcribed, and then manually segmented into proposition-sized segments (segments that represent approximately one concept-relation-concept triplet). Propositions are considered as the atomic unit of meaning in human memory in cognitive theories of text comprehension [3], and propositional analysis has been used extensively as a method for the study of clinical comprehension [24].

These segments were then coded using a medical knowledge hierarchy framework developed by Evans and Gadd [16]. Two authors (VVD, TC) with training in clinical psychiatry performed the coding. This framework has been applied to the study of diagnostic reasoning by Patel and her colleagues in a number of previous studies [4, 25–27]. This hierarchical framework consists of four levels of clinical knowledge organization, beginning with the *observation level* that consists of all the perceived information related to the case presentation. The next level is the *finding level*, which contains the interpretation of the observations or facts that are of clinical significance. Clusters of findings grouped into categories that are diagnostically relevant are categorized as belonging to the *facet level*. Therefore, facets can be considered as a type of intermediate construct. The next level is the *diagnostic level*, which is formed by a set of concepts that are the basis for management [16]. The hierarchy also includes a *system complexes* level that links diagnoses that tend to occur in the same patient, and relates to both risk factors and comorbid conditions. We did not consider the system complexes level of the hierarchy in our analysis, as our chief concerns were with the generation and justification of intermediate constructs, which occur at the facet level.

In the domain of psychiatry, relevant findings, finding-facet and facet-diagnosis relationships are encoded in the diagnostic and statistical manual of mental disorders (DSM). For the research described in this paper we used the fourth edition (the DSM-IV) [28] as a reference point, as both the generation of our case scenarios and the collection of our data preceded the release of the fifth edition (the DSM-V).

We focused our analysis at the level of facets. In addition, we identified two management-related clusters that are not strictly "facets" in the Evans and Gadd sense as they do not suggest specific diagnoses (one might make this argument about the "danger" facet also), but nonetheless constitute clusters of case information that are relevant to diagnosis and treatment. These were termed "psycho-social/family" (findings concerning the home environment of the patient) and "management" (description of a therapeutic plan).

For both clinical cases, the facets were identified and coded by one of our research team members (VVD). Facets were assigned to relevant propositional segments, when the segment concerned described clinical features related to one of the aforementioned facets. This assignment was based on the clinical knowledge of the researchers, with reference to the DSM-IV. For a subset of the propositions (206 segments, 12.5% of the total), the coding was independently conducted by another team member [14]. The discrepancies between the two members (13/206 = 6.3%) were resolved through collaborative discussion. In addition, this process resulted in the addition of 2 new coding categories ("borderline traits" and "post trauma"), which were applied by recoding the rest of the data set. The following facets and management-related clusters (MRCs) were identified (Table 1):

4.5.5 Analysis of summary data: analysis of facet occurrence—Facets that were identified through the coding process were evaluated based on their usage in verbal summaries. In other words, we investigated whether the interface influenced participants' organization of the clinical summary and the accuracy of their proposed diagnosis. We identified how frequently the facets were mentioned in the summary, when content pertinent to the facets was articulated in the summary and the characterization of consecutive mentions of facet-related content in the verbal summary. This analysis was conducted to evaluate the hypotheses that the interface may affect the content, order and coherence of the clinical case summaries.

4.5.5.1 Frequency of facet occurrence: In order to evaluate the extent to which the content relevant to each of the coded facets was emphasized by the participants, we assessed the number of propositions in the clinical summary that were annotated as related to each of the facets.

4.5.5.2 Relative position of facet occurrence: One of the hypotheses of this research was that the order in which facet-specific content is presented by the interface might affect the order in which this content is represented mentally, and that this would be reflected in the order with which it appears in the verbal summary of the case. The relative position of content related to each facet provides relevant insight into (a) the order of the appearance of facet-relevant content in a participant's summary and (b) how the order of the appearance of facets possibly influenced the clinical comprehension and problem-solving process of the

participants. The relative position of facet-relevant content was determined by counting where and how many times content coded as related to a particular facet appeared over the length of a transcript. The average value of a position per facet was then computed based on the point in the sequence of propositions where a facet appeared over the length of the transcript. The average value was then divided by the total number of propositions per participant's summary to isolate the relative position of each facet's appearance over the length of the transcript.

For example, consider a case summary that consisted of 71 propositions. Of the 71 propositions, there were 8 propositions that were clinically classified as "psychosis". These 8 propositions occurred at the following positions over the length of transcript: 8, 16, 22, 23, 25, 44, 52 and 54. The average position of "psychosis" facet was computed to be 30.5. This average value was then divided by the length of transcript (i.e., 71) to compute relative average position of psychosis facet to be 0.43 (30.5/71= 0.429). This analysis was conducted for both the simple and complex cases across both interface conditions.

4.5.5.3 Consecutive appearance of facets: Consecutive appearances of facet-relevant content were defined as a set of facets that appeared consecutive to each other in a participant's coded transcript. A further hypothesis was that there would be a trend toward greater coherence in summaries from the IC group, and that coherence can be estimated by considering the number of propositions pertaining to the same facet-level hypothesis that occur in sequence. In addition, this analysis provides insight into the extent to which content related to different facets appears together. The basic features of the consecutive appearances of facets (including the position of these sets of facets and the length of the sets of facets) will reveal (a) which facet-specific content tended to cluster together in a summary, (b) where these facet-related clusters appear over the length of the transcript, and (c) how these facet-related clusters potentially influenced clinical comprehension, problem solving and diagnostic accuracy. The relative position of each consecutive set of facets per participant was also evaluated.

The average position of a set of consecutive facet-related propositions over the total length of the transcript was computed. This procedure was repeated for all transcribed summaries of the participants across both cases and interface conditions. We also determined the average length of the consecutive set of propositions related to the same facet per case across both interface conditions. Such consecutive appearances reveal an extended discussion of elements related to a particular facet, suggesting an underlying mental representation in which case data are coherently organized in accordance with a particular facet model (See Table 3).

4.5.6 Evaluating relevance using LSA

We used LSA [17] to evaluate the relatedness between transcribed protocols from each participant and a reference standard created during the course of previous research [13]. The reference standard was derived from the think-aloud protocols gathered from two domain experts as they reasoned through the cases. The text included in the reference standard was the union of the sets of propositional segments recalled by each expert, constructed with the

assumption that the union of these recall protocols would include the majority of the clinically relevant components of the cases concerned (see [13] for further details).

LSA is a method of distributional semantics [29] that derives a human-like measure of semantic relatedness between terms, and larger units of text such as paragraphs. These measures of similarity have been used to approximate human performance on a number of cognitive tasks [17], and LSA is by now well established as a method within the cognitive science and information retrieval communities. The measures of semantic relatedness were derived from the same LSA-based semantic space we have utilized in previous experiments [6]. This space was constructed specifically for the content domain of clinical psychiatry, and details of the document set upon which it was trained can be found in [6][1].

For each participant on each case, we compared a LSA vector derived from the entire contents of their verbal think aloud protocol to a LSA vector derived from the reference standard for this case. LSA vectors were generated by superposing the LSA term vector for each term occurring in the transcript (or reference standard) and normalizing the resulting vector. The metric of comparison was the cosine metric, as is standard in LSA experiments.

5 Results

5.1 General characteristics: time spent, workload and usability

There were no significant differences between the usability of the two interfaces measured using the SUS scoring approach ($t(11)=1.55$, $p>0.05$). Based on a two-way analysis of variance (interface type x case complexity), we evaluated the differences in time spent and workload across both interfaces. We found no significant main effects on the time spent (M_{IC} = 474.1*s*, $M_{No\text{-}IC}$= 462.2*s*) across both interfaces ($F(1)=0.032$, $p > 0.05$) or cases ($F(1)=0.009$, $p> 0.05$; *C.V.* = 38.3). There was also no significant interaction ($F(1) =0.002$, $p> 0.05$) Additionally, there was also no significant main effects of workload across interfaces ($F(1)=0.04$, $p>0.05$) or cases ($F(1)=0.04$, $p>0.05$) or interaction ($F(1)=0.01$, $p > 0.05$). Additionally, based on Chi-square independence tests, there was no significant association between the accuracy of diagnosis for either cases for IC ($\chi^2(1)=0.001$, $p>0.05$) or No-IC ($\chi^2(1)=0.34$, $p>0.05$) interfaces.

5.2 Divergent recall

Table 2 presents our analysis of the think-aloud protocols of our participants, as they relate to the nine points of divergent recall identified by Sharda and his colleagues (five from Case 1: 1A–1E, and three from case 2: 2G–2I). For this analysis, we considered “recall” to constitute mention in the think-aloud protocol of all clinically relevant elements of the text segment highlighted by Sharda and his colleagues. For example, for segment 1A, both “dizziness” and “trouble sleeping” would need to be mentioned for recall to be acknowledged, and for segment 2H both the “flashback” and the fact that this involved a

[1]Although this space was also used to draw associations between related terms for the CSS, we would not expect this to bias our relevance results, as the trained system that maps between intermediate constructs and terms in the space was not utilized for the relevance analysis.

"past sexual assault" would need to be mentioned. However, mentioning modifiers of degree such as "frequently" and "prominent" was not considered essential.

The performance of certain individual participants is immediately evident from this table. In particular, participant 3 (NO-IC) did not produce any verbalization for case 2. Similarly, participants 6 and 8 (both IC) produced very sparse (or empty) protocols, and failed to mention any of the points of divergent recall. Aside from these two participants, the IC group displayed a greater tendency to recall these points, particularly when highlighted by the system[2]. In Case 1 all of the IC participants aside from the two with unusually sparse protocols mentioned all of the points of divergent recall that were highlighted by the system, which was not the case for the NO-IC group. For case 2, it is apparent that more of the IC group attended to the three points of divergent recall, all of which were highlighted by the system. These findings support the hypothesis that the system exerts effects on the process of clinical comprehension. In this case, the effects appear to be positive, in that the IC group appeared to better attend to clinically relevant points that were neglected by non-experts in a previous study.

5.3 Patterns of navigation

Figure 3 depicts the pattern of navigation of one of the participants (PS2), using the IC interface for the simple case (case 1). While reading the text, *"the patient was separated from her husband"*, the participant hypothesized that an abusive relationship with her husband may have precipitated this separation. To investigate this, the "danger" facet was selected (at time 13:23 minutes during the case), revealing information that confirmed this hypothesis. This illustrates a pattern of use in which the IC interface features were used to evaluate a hypothesis that was generated during the course of reading the narrative text summary.

Following the preliminary review of the "danger" facet, the participant focused on additional information that was organized under this facet. In addition to the case subject's abusive relationship with her husband, other evidence of tumultuous relationships led the participant to generate a diagnostic hypothesis of borderline personality disorder. This illustrates a pattern of use in which the sequential organization of information associated with a facet-level interface element directed attention toward related elements of the case, leading to the generation of a new diagnostic hypothesis. While this hypothesis is not in fact accurate, it follows logically from the information attended to previously.

Next, the participant read the highlighted text in the case narrative, and the text surrounding it. From this text, the participant identified a history of psychotic symptoms (including potentially dangerous command hallucinations) and antidepressant therapy, which led the participant to consider the diagnostic hypothesis of a mood disorder with psychosis, his second differential diagnosis. In this case the hypothesis was accurate, although it was not

[2]While it is not surprising that the system missed the possible paranoid undertones of the patient's suspicions of her teacher, it is interesting to note that it failed to recognize the textbook manic symptom, "shopping frequently". As the system was trained on annotated discharge summaries, it did not learn the significance of this finding, which may be more commonly encountered in textbooks than in clinical practice.

specific as it could refer to a number of diagnoses with both mood and psychotic components[3].

Following the first pattern of use in which the interface tabs were used to explore a hypothesis derived from the text, the participant then clicked on the "mood" facet (at time 16:07 minutes) to search for any depressive episodes in the past. When reviewing the findings organized under this facet, the participant found that the patient was diagnosed with psychotic depression – a fact that supported the generated hypotheses. After reading all the highlighted text under this facet, the participant concluded that the patient initially developed depression, followed by anxiety, agitation and psychotic symptoms. These findings support the participant's hypothesis that the patient developed a mood disorder with psychotic features. So this illustrates a pattern of use in which the highlighted text (rather than the information organized under a facet model) is reviewed rapidly to search for history consistent with the facet-level diagnostic hypothesis currently under consideration.

Then, the "substance abuse" facet was clicked (at time 19:40 minutes) and the absence of substance abuse history was confirmed[4]. This action represents another pattern of use in which the interface is used to rule out an alternative diagnostic hypothesis for the sake of completeness. Similarly, while reading through the remaining part of the narrative summary, the participant again utilized the facets on the interface to confirm his thoughts and recollections by reviewing, and at times clicking upon the findings organized at facet-level.

Figure 4 represents the patterns of navigation of the same participant, using the IC interface for the complex case (case 2). In this case, the participant started exploring the facets immediately, and focused on the issue of potential dangerousness after encountering information pertaining to self-harm in the information organized under the "danger" facet. The participant then clicked on this facet (at time 24:53 minutes) to explore causes underlying an alleged suicide attempt. When exploring the highlighted text, the participant read juxtaposed sentences describing multiple past diagnoses and admissions, and the current presentation involving command auditory hallucinations with suicidal content. Additional information regarding childhood sexual abuse was also read. In the past, the patient had presented with symptoms of both depression and hyperactivity along with some self-destructive behavior. These findings led to the generation of borderline personality disorder and bipolar disorder as initial diagnostic hypotheses.

Next, the participant clicked on the "mood" facet (at time 27:03 minutes in the video) to see if there were any clinically significant findings related to bipolar disorder, following a pattern of use that was also observed in Case 1. Under the "mood" facet the participant observed that the patient had a past diagnosis of bipolar disorder (supporting his previous hypothesis), a history of past depressive episodes along with flashbacks of sexual abuse and racing thoughts. Based on these findings, another diagnostic hypothesis was generated - the patient may suffer from an anxiety disorder. After reviewing all the mood-related findings, the participant moved to the social history in the text of the summary, in order to seek

[3]It consists of the intersection of two facet-level pre-diagnostic hypotheses, and as such exemplifies diagnostic reasoning at a level of abstraction consistent with that observed to be used by domain experts.

[4]Substance abuse must be excluded as an alternative explanation to diagnose most mood and psychotic disorders.

further support for his hypothesis of borderline personality disorder. In the social history, a description of the patient's instability in maintaining relationships reinforced the suspicion of borderline personality disorder. The participant then moved to the "substance abuse" facet (at time 31:03 in the video), to rule out this diagnostic alternative. He did not find evidence that substance abuse was a significant contributor to the present condition.

While reading through the remaining part of the case, the participant clicked on the facets several times to confirm the previously encountered findings, and look for any that may have been missed. Since he suspected bipolar disorder, he reviewed the "mood" facet to look for a history of manic episodes. Similarly, he clicked on the "psychosis" facet to confirm the nature of the patient's command auditory hallucinations. This pattern of use conforms to the previously identified pattern in which the system is used as a final step to review the case for completeness. This review was more extensive than in the first case, as one might anticipate given the greater cognitive demands of case 2, where the past diagnoses assigned to the patient cover almost the entire range of psychiatric disorders. Furthermore, the findings of the case are difficult to organize at a facet level, as the patient is malingering in order to obtain supplemental security income (SSI). As the symptoms were confabulated by the patient, they lack congruity that one would expect from a patient with a legitimate psychiatric disorder. This incongruity in itself is a diagnostic cue that suggests the patient may be malingering.

In summary, we identified the following five patterns of navigation:

Hypothesis evaluation: The interface was used to evaluate hypotheses generated while reading the narrative text summary, by reviewing related information organized at facet level or by reviewing the highlighted facet-relevant components of the narrative summary.

Leveraging text juxtaposition: Sequential organization of information associated with interface elements at facet level led to the generation of new diagnostic hypotheses. A similar strategy occurred when text highlighted by the interface, and text juxtaposed with this text, contained narrative rich in diagnostically useful information, leading to the generation of facet-level diagnostic hypotheses.

Review to Exclude: The interface was used to rule out alternative diagnostic hypotheses for the sake of completeness, by reviewing elements organized at facet level.

Review to Confirm: The interface was used to confirm thoughts and recollections by reviewing the findings organized at facet level (both in the interface and highlighted in the text).

Facet-level Preview: The facet-level elements were reviewed before the narrative text was read.

The examples described above illustrate how the interface mediated reasoning at facet level. Furthermore, it is clear that the interaction with the interface influenced the evolution of diagnostic hypotheses (both at facet-level and at the diagnostic level), in turn guiding further interaction with the system. We also observed that participants using the IC interface were more efficient in recollecting certain clinically significant findings while summarizing the

case. For example, in their summaries six out of seven participants who used the IC interface recollected that the patient (in Case 1) had command auditory hallucinations to kill herself and her husband. In contrast, only two out of eight participants in the NO-IC mentioned this finding, which is of clinical importance for the assessment of the risk of potential dangerousness (a similar imbalance is evident in the think-aloud protocols, as shown in Table 2).

Participants in the No-IC group tended to read through the entire case narrative from top to bottom, while thinking aloud. They subsequently reviewed the narrative to either confirm or to exclude the previously made diagnoses. However, they did not exhibit the flexible hypothesis- and data- driven navigation patterns observed in the IC group.

5.4 Relative position of facets

Figure 5 shows the distribution of the content related to the various facets averaged over all participants across interface type and case complexity. While there were no overall significant differences in the content across the four groups, there were certain nuances in the use of content. For example, the No-IC group generated more content related to psychosocial aspects in both the simple and complex cases, and more content related to mania in the simple case.

In addition to the distribution of facet-related content, we also evaluated the relative position of such content within a clinical summary. The relative position of facet-related propositions represents when the specific facet was verbalized during the clinical summary. The primary assumption is that the interface affects clinical comprehension, which in turn would influence the order of appearance of content in clinical summary. Facets appeared to differ in their position of appearance between the IC and No-IC conditions – the difference was more pronounced in the complex case than in the simple case. For the *simple* case, “mood” and “psychosocial” appeared to differ in their position of appearance between IC and No-IC conditions, with “mood” appearing much earlier in the No-IC condition and “psychosocial” appearing later in the No-IC condition (no statistically significant differences were observed).

For the *complex* case, “depression” appeared later in the clinical summary in the No-IC case than the IC case. In contrast, facets related to “mood”, “psychosis” and “danger” appeared earlier in the No-IC case than the IC case (no statistically significant differences were observed).

5.5 Consecutive appearance of facets

In addition to this analysis, we characterized the consecutive appearance of facet-related propositions. Such consecutive appearances reveal an extended discussion of elements related to a particular facet, suggesting an underlying mental representation in which case data are coherently organized in accordance with a particular facet model (See Table 3). The length of these sequences of facets varied by participant: the smallest sequence was two segments in length, and the largest sequence was nine segments in length.

The frequency with which propositions related to a particular facet appeared in sequences of two or more was computed for each participant, and then averaged across all participants for each case and interface type. For example, for the “mania” facet element in the IC interface (for the simple case), the average consecutive occurrence was 1.8. This can be explained as follows: on average propositions relevant to the “mania” facet element occurred about twice in sequence per verbal summary (in the IC-simple condition). A higher value for consecutive appearances shows clustering of similar facet elements, suggesting a more coherently organized mental representation of the case.

In both simple and complex cases, the IC interface appeared to support the clustering of elements related to facets for “psychosis” and “management”. In comparison to the No-IC group, the IC group participants appeared to organize facets sequentially more frequently across both case types. In other words, these findings suggest that the IC interface may assist participants’ organization of information from the case into meaningful, clinically relevant clusters.

For the *complex* case, the computed lengths of the set of facet-related propositions were statistically significant for “psychosocial” (df=35, p= .02) and approached statistical significance for “psychosis” (t = 1.9, df = 35, p-value = 0.064). For “psychosis”, the length of the set of facet elements was longer for the IC group of participants than for No IC group of participants. In contrast, the length of the set of facet-related elements for “psychosocial” was longer for the No IC group of participants than for IC group of participants. These findings suggest that the IC interface supported the clustering and coherence of elements related to the facet “psychosis”, but hindered this aggregation for elements related to the “psychosocial” facet. For the *simple* case, differences in the computed length of the sequence of facet-related elements were statistically significant for “mania” ($t = -2.1$, df = 28, p-value = 0.045), where the length of the set of facet elements was longer for the No IC group of participants than for IC group of participants. These findings suggest that the IC interface may differentially affect participants’ organization of information into meaningful clinically relevant clusters in cases where these clusters are explicitly represented.

5.6 Diagnostic accuracy and management

Table 4 provides a summary of the diagnostic accuracy of each participant. Interestingly, the IC group tended to produce longer case summaries (median word count=504.5) than the No-IC group (median word count=315), though the difference in mean word count is not statistically significant on account of within-group variation.

With respect to the simple case, where the correct diagnosis was schizoaffective disorder, the majority (6/8) of the No-IC group mentioned this as either a primary or differential diagnosis. Exactly half of the IC group did also, although no participant in this group mentioned it as their primary diagnosis. However, in both groups all participants that provided diagnoses that accounted for both the mood and the psychotic features of the case. The system did not support the assembly of these facet-level elements into a higher-level diagnosis (i.e., schizoaffective disorder), so the fact that more participants in the No-IC group made this connection may not be related to interface-based support. An alternative hypothesis might be that the temporal relationships between the psychotic and mood-related

symptoms are best presented in narrative form, and that breaking the flow of this narrative may distract attention from these aspects of the case. However, a review of the think-aloud protocols did not support this hypothesis, as there were more examples of explicit temporal reasoning in the protocols of the IC group than the No-IC group.

With respect to the second case, it is somewhat surprising that the accurate diagnosis of malingering was suggested by more participants in the IC group (3/8) than the No-IC group (1/8), though the No-IC participant did select this as their primary diagnosis. Accurate diagnosis here rests upon two observations: (1) the symptoms do not fit any particular diagnosis comfortably, and (2) the patient is receiving Supplemental Security Income (SSI) on account of her apparent disability, and hence has a so-called “secondary gain”, an advantage deriving from her alleged symptoms. While cognitive support may make it easier to make the first of these observations, the fact the patient is on SSI was not highlighted by the system, and was noted by more of the No-IC group during the process of interpreting the case. One possibility is that the cognitive support provided by the system preserved cognitive resources for the higher-level reasoning that would be required to make this connection. However, it is difficult to draw firm conclusions from this finding on account of the variable extent to which the IC group utilized the system. For example, the participant that made the most extensive use of the system neither observed the SSI nor made the correct diagnosis.

Regarding management, as it would be appropriate in both cases for initial management to be symptomatic, recommendations from both groups were largely consistent. For the first case, those participants who provided management plans almost invariably described a medication regimen including a mood stabilizer and anti-psychotic agent. Some participants mentioned other measures including a pregnancy test (which would be a good idea in this case as several mood stabilizing agents are known to be teratogenic), hypnotics, supportive therapy and, in some cases, an involuntary admission if necessary. For the second case, participants also tended to propose the use of mood stabilizers and antipsychotic agents. Other treatment options included various forms of psychotherapy. A pattern delineating the management plans of each group (IC vs. No-IC) was not apparent.

5.7 Relevance

In order to estimate the extent to which participants emphasized relevant content, we compared their verbal protocols to a reference standard derived from expert verbal protocols, using LSA. The LSA similarities for all participants, as well as the means and medians per case for each group are shown in Table 5. On account of the variance in performance within the groups, the median is a more robust descriptive statistic than the mean when comparing these groups to one another. The median scores are higher with cognitive support in both cases, but this difference is statistically significant for case 2 only. The only statistically significant difference in means occurred between the performance of the No-IC group across cases [$t(7) = 2.7229$, $p = .0296$]. This indicates a significant drop in the similarity between participants and the reference standard in the No-IC group between the simple and complex case. These findings indicate a general trend toward protocols with greater similarity to the expert-derived reference standard in the IC group. They also suggest

that cognitive support may have a greater effect in more complex cases, as one might anticipate given the greater cognitive load involved.

In interpreting these results, it is important to note that individual performances were strongly correlated across cases (Pearson's r=0.9250). In other words, individuals whose protocols tended to approximate the reference standard in the simple case also tended to do so in the complex case, across all participants regardless of the IC vs. No-IC condition. So the characteristics of individual participants are an important factor here, which should be taken into account when interpreting these results.

6 Discussion

While wide differences in individual performance limited our ability to identify definitive quantitative differences in the organization of information in the summaries produced by groups using the two interfaces, qualitative analysis revealed nuances in the nature of use of the interface. For example, one of the key differences we observed was how the IC-interface supported participants' clustering of case-relevant information. Such organization of key diagnosis-related information can have significant impact on supporting clinicians in their reasoning and diagnostic activities. It seems intuitive that offloading this task to the interface can free additional cognitive resources for high-level problem solving. This is a potential explanation for the IC group participants' more detailed recollection and better diagnostic accuracy in the more complex case. Additionally, the off-loading of cognitive load to external resources (such as a support system), can be potentially advantageous for patient safety, as more cognitive resources are available for error detection and recovery [30, 31].

Another aspect that was revealed by the qualitative analysis is the role of annotated/highlighted text in facilitating easy identification of important concepts. As reported in prior research (See e.g., [1]) augmenting the presentation of electronic documentation that affords easy identification and readability can increase usability of narrative text. This concept, often referred to as "enrichment" [32], increases the amenability of electronic narrative to rapid reading. For example, reading the results from a graph is easier than from a table. Our results point to the fact that clinicians focus on the highlighted text (often also reading preceding and following text) to capture key concepts that are relevant to the case. In our analysis of divergent recall, participants in the IC group more reliably attended to clinically relevant elements of the case that had previously been neglected by non-expert participants, particularly when those elements were highlighted on the interface. These elements included indicators of potential dangerousness pertinent to near-term management that could have dire consequences if ignored. Pioneering research in cognitive psychology has illustrated "pop-out effects" that allows people to focus on text (or other content) that has additional features (e.g., color) [33]. We believe such effects are at play when the IC-based interface is employed.

These findings are reminiscent of those encountered in previous research, in which the structure of an EHR interface was found to influence the clinical data emphasized by physicians [27]. These effects may be helpful or harmful. Though it is true that more participants in our IC group identified certain important clinical findings that have bearing

on acute management, it is also the case that the summaries of this group placed less emphasis on psychosocial aspects of these cases, which should not be neglected in a holistic case assessment. This aside, the observed effects of cognitive support appear to have been generally positive. In addition, the LSA findings suggest a trend toward expert-like emphasis with cognitive support, and that this cognitive support may exert a greater effect on clinical comprehension in more complex cases, where cognitive support is potentially needed, even for expert physicians.

Qualitative analysis also revealed patterns of use through which the facet-level perspectives provided by the system were leveraged during the course of clinical problem solving, providing clear evidence that, when used effectively, the system can mediate cognition at the facet level.

We would like to acknowledge the following limitations of our study. First, the study was conducted with a small sample of participants (n=16) at a single site. However, considering this is an exploratory evaluation of a new interface, and that the number of participants represents nearly 80% of the qualified participants (PGY 3 or 4), we believe that we had a representative sample for this study. Second, we did not ascertain the response times during generation of the verbal summaries from memory. Other memory-related metrics of cognitive performance, such as response time latency, may reveal additional insights when applied to these data. Third, the degree to which the system was used varied widely within the IC group – some participants frequently reverted to reading the narrative text without explicit reference to the facet-level elements. While it may be the case that the highlighting of text related to these elements exerted a subliminal effect, these participants did not make optimal use of system capabilities. This is a limitation of this study, and a hazard of permitting users the freedom to use the interface as they please. As with any new (experimental) system that veers significantly from the conventional approaches, users may take a while to acclimatize to use of the system. The adaptive learning process that drives effective system adoption takes time and repetitive usage. While we attempted to control these confounding variables - through training videos, reading the documentation and a practice task - we believe, retrospectively, that a longitudinal study design (with some interface improvements) may be necessary to capture some of the effects of the IC interface. Further studies with a more restrictive experimental protocol are underway to address this limitation. In addition, we note that as our analysis was based on verbal protocols, non-verbal aspects of decision making, such as mental images, were not addressed in this study. It is also the case that our experiments were conducted under laboratory conditions. As such we have yet to evaluate the impact of a busy clinical environment on use of the interface.

7 Conclusion

This paper describes the design and evaluation of a theoretically driven CSS for psychiatry. Comparison between participants with and without the use of this interface suggests that the interface exerts an influence on the content and coherence of spoken summaries of the case, and that these effects, which include appropriately emphasizing diagnostic and therapeutically elements of the case, appear to be more pronounced in complex cases with greater cognitive load. This shows that knowledge-based organization may provide better

support as complexity of clinical cases increases. Qualitative analysis suggests highlighted aspects of the cases are more likely to receive attention, and reveals several patterns of use through which the system mediates facet-level diagnostic reasoning. Through such mediation, some of the cognitive work of clinical comprehension is redistributed from clinician to machine.

Acknowledgments

This research project was partially supported by Grant No. 220020152 by James S McDonnell Foundation (JSMF) for Cognitive Complexity and Error in Critical Care and Grant R01 LM07894 from the National Library of Medicine to Vimla L. Patel; and by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects Program (SHARP) from the Office of the National Coordinator for Health Information Technology to Jiajie Zhang.

References

1. Kannampallil TG, Franklin A, Mishra R, Almoosa KF, Cohen T, Patel VL. Understanding the Nature of Information Seeking Behavior in Critical Care: Implications for the Design of Health Information Technology. Artificial Intelligence in Medicine. 2012; 57(1):21–29. [PubMed: 23194923]
2. Kannampallil TG, Schauer GF, Cohen T, Patel VL. Considering complexity in healthcare systems. Journal of Biomedical Informatics. 2011; 44(6):943–7. [PubMed: 21763459]
3. Kintsch W, Greeno JG. Understanding and solving word arithmetic problems. Psychological Review. 1985; 92:109–129. [PubMed: 3983303]
4. Patel VL, Arocha JF, Kaufman DR. Diagnostic Reasoning and Medical Expertise. The Psychology of Learning and Motivation: Advances in Research and Theory. 1994; 31:187–252.
5. Arocha, JF.; Patel, VL. Construction-Integration Theory and Clinical Reasoning. In: Weaver, CA.; Mannes, S.; Fletcher, CR., editors. Discourse Comprehension: Essays in Honor of Walter Kintsch. Lawrence Erlbaum Associates; Hillsdale, NJ: 1994. p. 359-381.
6. Cohen T, Blatter B, Patel VL. Simulating Expert Clinical Comprehension: Adapting Latent Semantic Analysis to Accurately Extract Clinical Concepts from Psychiatric Narrative. Journal of Biomedical Informatics. 2008; 41(6):1070–1087. [PubMed: 18455483]
7. Chi, MTH.; Glaser, R.; Farr, MJ. The Nature of Expertise. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
8. de Groot, AD. Thought and Choice in Chess. The Hague, The Netherlands: Mouton; 1965.
9. Lesgold, A.; Rubinson, H.; Feltovich, P.; Glaser, R.; Klopfer, D.; Wang, Y. The nature of expertise. Lawrence Erlbaum Associates; Hillsdale, NJ: 1988. Expertise in a complex skill: Diagnosing x-ray pictures; p. 311-342.
10. Norman GR, Rosenthal D, Brooks LR, Allen SW, Muzzin LJ. The development of expertise in dermatology. Archives of Dermatology. 1989; 125:1063–1068. [PubMed: 2757402]
11. Patel, VL.; Groen, GJ. The General and Specific Nature of Medical Expertise: A Critical Look. In: Ericsson, A.; Smith, J., editors. Towards a General Theory of Expertise: Prospects and Limits. Cambridge University Press; Cambridge, UK: 1991. p. 93-125.
12. Patel, VL.; Evans, DA.; Kaufman, DR. Cognitive Framework for Doctor-Patient Interaction. In: Evans, DA.; Patel, VL., editors. Cognitive Science in Medicine: Biomedical Modeling. MIT Press; Cambridge, MA: 1989. p. 253-308.
13. Sharda P, Das AK, Cohen T, Patel VL. Customizing Clinical Narratives for the Electronic Medical Record Interface Using Cognitive Methods. International Journal of Medical Informatics. 2006; 75:346–368. [PubMed: 16125455]
14. Hutchins, E. Cognition in the Wild. Cambridge, MA: MIT Press; 1995.
15. Cohen, T. Augmenting Expertise: Toward computer-enhanced clinical comprehension. New York, NY: Columbia University; 2007.

16. Evans, DA.; Gadd, CS. Managing Coherence and Context in Medical Problem-Solving Discourse. In: Evans, DA.; Patel, VL., editors. Cognitive Science in Medicine: Biomedical Modeling. MIT Press; Cambridge, MA: 1989. p. 211-255.
17. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 1997; 104:211–240.
18. Gärdenfors, P. Conceptual spaces: The geometry of thought. Cambridge, MA: MIT Press; 2004.
19. Giles, JT.; Wo, L.; Berry, MWSDMKD. Software for text mining, in statistical data mining and knowledge discovery. 2001. GTP (General Text Parser) Software for Text Mining; p. 455-471.
20. Ericsson, KA.; Simon, HA. Protocol Analysis: Verbal Reports as Data. Cambridge, MA: Harvard University Press; 1993.
21. Patel VL, Arocha JF, Kaufman DR. A Primer on Aspects of Cognition for Medical Informatics. Journal of the American Medical Informatics Association. 2001; 8(4):324–343. [PubMed: 11418539]
22. Hart, SG.; Stavenland, LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P.; Meshkati, N., editors. Human Mental Workload. Elsevier; Amsterdam, The Netherlands: 1988. p. 139-183.
23. Brooke, J. SUS: a 'quick and dirty' usabiliy scale. In: Jordan, PW.; Thomas, B.; McClell, IL., editors. Usability Evaluation in Industry. Taylor & Francis; London, UK: 1996. p. 189-195.
24. Arocha JF, Wang D, Patel VL. Identifying Reasoning Strategies in Medical Decision Making: A Methodological Guide. Journal of Biomedical Informatics. 2005; 38:154–171. [PubMed: 15797004]
25. Patel VL, Groen GJ, Scott HS. Biomedical Knowledge in Explanations of Clinical Problems by Medical Students. Medical Education. 1988; 22(5):398–406. [PubMed: 3205191]
26. Patel VL, Kaufman DR, Arocha JF. Emerging Paradigms of Cognition in Medical Decision Making. Journal of Biomedical Informatics. 2002; 35:52–75. [PubMed: 12415726]
27. Patel VL, Kushniruk AW, Yang S, Yale J-F. Impact of a Computer-based Patient Record System on Data Collection, Knowledge Organization, and Reasoning. Journal of the American Medical Informatics Association. 2000; 7(6):569–585. [PubMed: 11062231]
28. Spitzer, RL.; Gibbon, ME.; Skodol, AE.; Williams, JBW. DSM-IV casebook: A learning companion to the Diagnostic and Statistical Manual of Mental Disorders. 4. American Psychiatric Association; 1994.
29. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. Journal of Biomedical Informatics. 2009; 42(2):390–405. [PubMed: 19232399]
30. Amalberti, R.; Wioland, L. Human error in aviation. In: Soekkha, H., editor. Aviation safety: human factors, system engineering, flight operations, economics, strategies, management. VSP; Utrecht: 1997.
31. Patel, VL.; Kaufman, DR.; Cohen, T. Cognitive Informatics in Health and Biomedicine: Case Studies on Critical Care, Complexity and Errors. London, UK: Springer; 2014.
32. Pirolli P, Card S. Information foraging. Psychological Review. 1999; 106:643–675.
33. Treisman AM, Gelade G. A feature integration theory of attention. Cognitive psychology. 1980; 12:97–136. [PubMed: 7351125]

Highlights

- We evaluate the effects of cognitive support on psychiatric clinical comprehension
- Users selectively attended to clinically relevant points highlighted by the system
- Organization of information facilitated hypothesis generation and evaluation
- Users focused on information pertinent to acute care

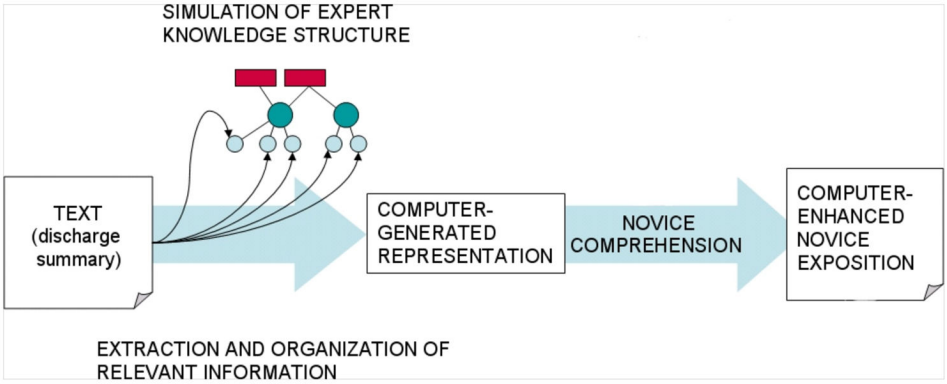

Figure 1.

Redistribution of aspects of expert comprehension from human to computer in order to support expert-like comprehension by the clinician-computer dyad.

Figure 2.

Three areas of the IC user interface: (A) Overview and summary of the diagnostic categories, (B) detailed patient note and (C) the details related to the selected diagnostic category from (A). The arrows in region B show the elements of this frame that can be resized with a mouse drag.

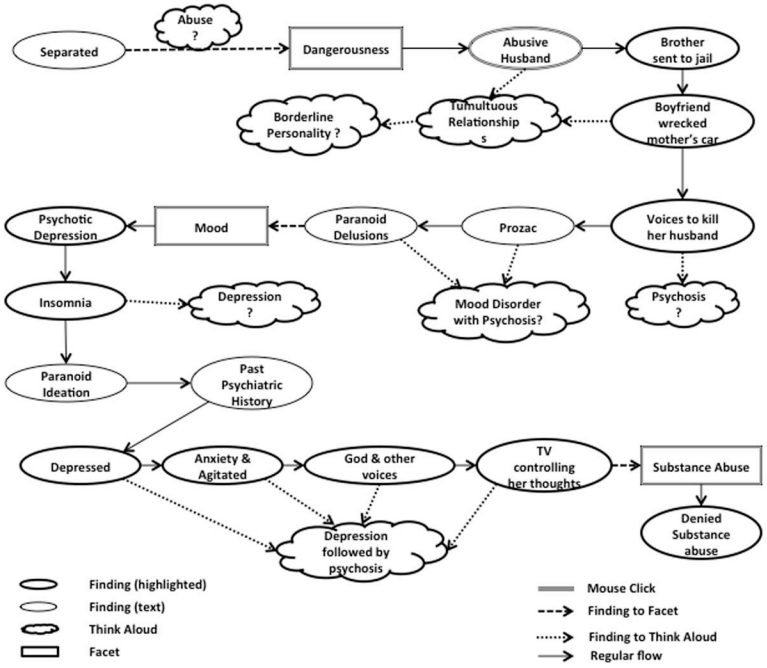

Figure 3.

Schematic figure showing the progression of case interpretation, comprehension and navigation for the *simple* case for the IC condition.

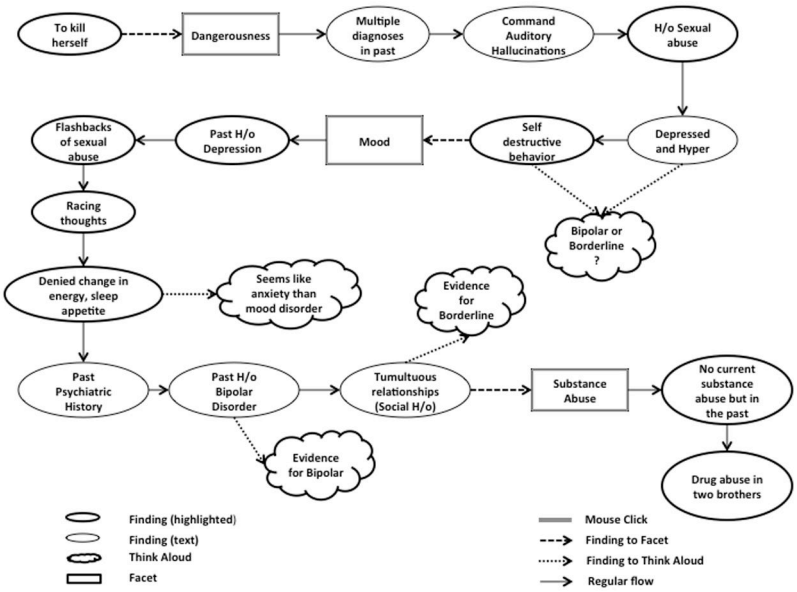

Figure 4.

Schematic figure showing the progression of case interpretation, comprehension and navigation for the *complex* case for the IC condition.

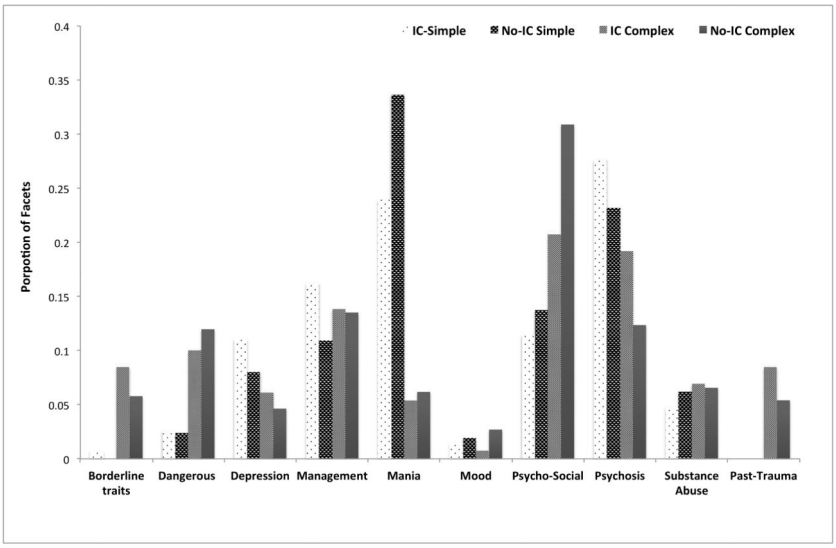

Figure 5.

Distribution of content related to the various facets across the verbal summaries (between IC, No-IC in both simple and complex cases)

Table 1

Identified facets and patient management related aspects with relevant examples

Facet/MRC	**Examples of relevant findings from cases**
Borderline Traits	Disruptive behavior, Unstable relationships, Self-hurting behavior
Dangerousness	Suicidal and homicidal tendencies or attempts
Depression	Isolation, Not taking care of herself, Insomnia, Dizziness
Management	Psychotherapy, Psychopharmacology (Haldol, Risperdal)
Mania	Agitated, Pressured speech, Shopping spree, Talking to herself, “Driven by a motor”, Hyper religiosity, Hyper sexuality.
Mood	Labile mood, Mood symptoms
Post trauma	Childhood sexual abuse, flashbacks, past trauma
Psycho-Social/Family	Divorce, On SSI (Supplemental Security Income), Family history of psychosis, Stressful family situations.
Psychosis	Hallucinations, Paranoid delusions, Command auditory hallucinations.
Substance abuse	Prescription drug abuse. Alcohol abuse, Opiate abuse.

Table 2

Comparison across points of "divergent recall" characterized by Sharda *et al* 2005. The top half of the table is derived from the think-aloud protocol captured during exploration of the case. SUBJ=subject number. WC1=word count of protocol for case 1 and so forth. ✔ indicates recall of the proposition(s) concerned. Grey cells indicate no think aloud data was produced by this participant during interpretation of the case. 1A=proposition(s) A for case 1 and so forth. White text on a black background indicates this information was highlighted by the system. The bottom half of the table describes the propositions highlighted by Sharda et al, as well as their clinical significance and which facet, if any, the system grouped them under.

	IC								NO-IC							
SUBJ	2	4	6	8	10	12	14	16	1	3	5	7	9	11	13	15
WC1	613	761	65	88	1296	1373	1058	600	274	296	215	264	995	1336	1115	356
1A	✔	✔			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
1B	✔	✔			✔	✔	✔	✔			✔		✔	✔	✔	
1C		✔			✔	✔	✔		✔	✔			✔	✔	✔	
1D		✔			✔	✔		✔	✔	✔			✔	✔	✔	
1E	✔	✔			✔	✔	✔	✔					✔	✔		
1F	✔	✔			✔	✔	✔	✔					✔	✔	✔	
WC2	371	925	0	174	1434	827	1002	820	193	0	232	591	614	1799	1104	128
2G		✔			✔	✔	✔	✔						✔	✔	
2H	✔				✔	✔	✔				✔			✔	✔	
2I	✔	✔			✔	✔	✔						✔	✔	✔	

	Proposition(s)	Facet	Clinical significance
1A	Dizziness, trouble sleeping	Mood	Differential – somatic disease
1B	Patient irritable	Mood	Irritable mood (mania/depression)
1C	Preoccupied that college teacher was trying to ruin her grades	None	Paranoid delusion: symptom of schizophrenia and related d/o
1D	Shopping frequently	None	Textbook manic symptom
1E	Prominent thought disorder	Psychosis	'Class A' symptom of schizophrenia and related d/o
1F	Command hallucinations to kill herself and her husband	Danger	Indicates potential dangerousness
2G	Content of command auditory hallucinations: command "to kill herself by cutting her wrists"	Danger	Indicates potential dangerousness
2H	Flashback to a past sexual assault	Danger	Symptom of post-traumatic stress disorder
2I	Racing thoughts	Mood	Textbook manic symptom

Table 3

Frequency of consecutive appearance of facet contents. Each number represents the average number of consecutive appearances (of length >= 2) of facet elements for each interface type and case complexity. NA refers to instances where the particular facet element did not appear in the verbal summary (e.g., Borderline traits for Simple-IC condition).

	Simple-IC	Simple No-IC	Complex IC	Complex No-IC
Borderline traits	NA	NA	0.1	0.1
Danger	0	0.1	1.0	0.4
Depression	0.9	0.6	0.4	0.3
Management	1.6	0.6	1.4	0.9
Mania	1.9	1.8	0.3	0.5
Mood	NA	NA	NA	NA
Past trauma	NA	NA	0.9	0.5
Psychosocial	0.9	0.9	2.1	1.9
Psychosis	3	1	1.6	0.5
Substance abuse	0.4	0.5	1.0	0.8

Table 4

Diagnostic accuracy. SA = schizoaffective disorder. M = malingering. SSI = mention of social security income (or disability-related compensation) in the think-aloud protocol. ★ = definitive diagnosis, which subsumes✔= differential diagnosis, or in the case of SSI,✔= mentioned SSI in think-aloud protocol. ND = no diagnosis provided. NTA= no think aloud. WC1 = summary word count for case 1. WC2 = summary word count for case 2.

	IC								NO-IC							
SUBJ	2	4	6	8	10	12	14	16	1	3	5	7	9	11	13	15
WC1	156	603	402	50	1067	765	604	632	287	323	188	289	624	423	516	288
SA		✔		ND	✔	✔	✔		★	ND	✔	★	✔	★	✔	
WC2	220	321	332	51	1310	1403	546	463	307	214	277	591	614	443	696	272
SSI		✔	NTA		✔	✔	✔	✔	✔	NTA	✔	✔	✔	✔	✔	✔
M			✔	ND	✔	✔	ND			ND	★					

Artif Intell Med. Author manuscript; available in PMC 2015 October 01.

Table 5

Similarity between LSA vectors for each participant and the reference standards.

IC	PS2	PS4	PS6	PS8	PS10	PS12	PS14	PS16	MEAN	MED
Simple	0.778	0.899	0.785	0.676	0.944	0.952	0.958	0.902	0.862	0.901
Complex	0.785	0.855	0.745	0.676	0.927	0.936	0.976	0.881	0.848	0.868[+]

NO-IC	PS1	PS3	PS5	PS7	PS9	PS11	PS13	PS15	MEAN	MED
Simple	0.872	0.830	0.842	0.847	0.967	0.956	0.951	0.844	0.889[*]	0.859
Complex	0.858	0.730	0.852	0.748	0.960	0.919	0.936	0.778	0.848[*]	0.855[+]

[*] significant difference across cases by paired t-test.

[+] significant difference across conditions (IC vs. NO-IC) by Wilcoxon's signed rank test.

Artif Intell Med. Author manuscript; available in PMC 2015 October 01.