Revealing the Dynamics of Medical Diagnostic Reasoning as Step-by-Step Cognitive Process Trajectories

Dominik Battefeld[1], Sigrid Mues[2], Tim Wehner[2], Patrick House[3], Christoph Kellinghaus[4], Jörg Wellmer[2], Stefan Kopp[1]

[1]Social Cognitive Systems Group, Faculty of Technology, CITEC, Bielefeld University, Bielefeld, Germany
[2]Ruhr-Epileptology, University Hospital Knappschaftskrankenhaus Bochum, Ruhr-University, Bochum, Germany
[3]Epileptologicum Hamburg, Hamburg, Germany
[4]Department of Neurology and Epilepsy Center Klinikum Osnabrück, Osnabrück, Germany

Abstract

A detailed understanding of the cognitive process underlying diagnostic reasoning in medical experts is currently lacking. While high-level theories like hypothetico-deductive reasoning were proposed long ago, the inner workings of the step-by-step dynamics within the mind remain unknown. We present a fully automated approach to elicit, monitor, and record diagnostic reasoning processes at a fine-grained level. A web-based user interface enables physicians to carry out a full diagnosis process on a simulated patient, given as a pre-defined clinical vignette. By collecting the physician's information queries and hypothesis revisions, highly detailed diagnostic reasoning trajectories are captured leading to a diagnosis and its justification. Four expert epileptologists with a mean experience of 19 years were recruited to evaluate the system and share their impressions in semi-structured interviews. We find that the recorded trajectories validate proposed theories on broader diagnostic reasoning, while also providing valuable additional details extending previous findings.

Keywords: Differential Diagnosis; Diagnostic Reasoning; Reasoning Process Analysis; Seizure; Epilepsy

Introduction

Medical diagnoses are key to managing and curing diseases (Donner-Banzhoff, 2022), and hinge on a solid categorization of a patient's complaints in mind. Xu et al. (2016) have identified major cases of misdiagnosis of transient loss of consciousness, where patients not suffering from epilepsy were mismanaged with anti-epileptic drugs, lost their driving license, or dropped into unemployment. Indeed, the estimated rate for seizure misdiagnosis lies between 4.6% and 30% (Chowdhury, Nashef, & Elwes, 2008). Medical experts are aware of these risks but are cognitively challenged by a highly uncertain and incomplete problem. The presence or absence of specific semiological features cannot warrant a decision towards or against any differential diagnosis and information is mainly obtained through subjective, personal dialogue rather than objective test results (Malmgren, Reuber, & Appleton, 2012). Up to today, the most promising diagnostic approach is rigorous anamnesis, critical eyewitness report analysis, and conservative examination report interpretation (Plug & Reuber, 2009). As a result, collected evidence suffers from inaccurate memory retrieval by patients or simply miscommunication. In addition, cognitive biases like favoring one hypothesis over another due to experiential familiarity (Availability bias) or missing important information during symptom exploration (Premature closure) are common (Saposnik, Redelmeier, Ruff, & Tobler, 2016).

A lot of research in Cognitive Science has looked at how expert physicians approach this problem and how they are still able to act efficiently given the myriad of possible questions to ask, diagnostic tests to conduct, and hypotheses to consider (Croskerry, Campbell, & Petrie, 2023; L. Cheng & Senathirajah, 2022; Kumar, Ferguson, Swee, & Suneja, 2021; Koufidis, Manninen, Nieminen, Wohlin, & Silén, 2021; Gupta et al., 2021; Scholz, Krems, & Jahn, 2017; Brush, Sherbino, & Norman, 2017; Banda, 2010; Elstein, 2009; Coderre, Mandin, Harasym, & Fick, 2003). Still, as Sox, Higgins, and Owens (2013) state, we "know more about how clinicians *should* reason than about how they *do* reason" (p. 8). In this paper, we introduce an approach to unravel and quantify the medical diagnostic reasoning process at a fine-grained level. To that end, we propose a patient simulator along with a web-based monitoring tool where experts can visually explore information, request examinations, and hypothesize about differential diagnoses. The tool tracks all actions during information exploration and hypothesis revision, resulting in a step-by-step cognitive process trajectory that leads to a final diagnosis and its justification. We argue that this approach provides valuable insights into the cognitive reasoning process behind diagnostic exploration in a controlled environment, while still being scalable to many study participants and various simulated patients.

To ensure the scientific validity and practical usability of our approach we report on a proof-of-concept study in which we collected 40 diagnostic reasoning trajectories of 4 medical experts from 3 different medical facilities on 10 artificial clinical vignettes. Additional semi-structured post-interviews were conducted with each participant to evaluate the realism of the simulation and monitoring tool. We find that the monitoring tool can replicate previous findings from empirical work on diagnostic reasoning while additionally providing an in-depth look into the cognitive dynamics of information exploration and hypothesis revision. The conducted interviews suggest that the tool is easy to understand, intuitive to use, and judged as closely related to a real diagnostic process, while also raising some improvement requests. In the following, we first elaborate on how medical diagnostic reasoning has been studied and modeled in related work. Then we introduce our approach to monitor medical diagnostic reasoning and present the evaluation study run with medical experts in the domain of seizure diagnosis. We conclude with a discussion of the results and limitations.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Related Work

Medical diagnostic reasoning has been studied in both Cognitive Science/Psychology and Medicine. Studies in the former realm are concerned with the mechanisms underlying the reasoning process and its dynamics, as well as in providing explanations for cognitive biases observed in this. The most prominent cognitive model of diagnostic reasoning is the hypothetico-deductive reasoning model (Elstein, Shulman, & Sprafka, 1978) according to which physicians generate leading hypotheses early on and then strive to decrease the number of differential diagnoses by directed exploration of possible information.

As this model is unable to explain the rapid, unconscious reasoning processes of highly experienced practitioners, Barrows, Norman, Neufeld, and Feightner (1982) introduced the pattern recognition model. Here, physicians rely on gradually acquired and highly compiled knowledge structures to match observed disease patterns to disease categories in memory (Brush et al., 2017). This model goes hand in hand with the illness script theory (Schmidt & Volder, 1984; Custers, Boshuizen, & Schmidt, 1998) that tries to describe the organization of memory structures. Over time, medical knowledge is said to compile into illness scripts: list-like structures that capture the enabling conditions (age, sex, medication, occupation, etc.) of facilitated pathophysiological malfunctions that cause observable symptoms.

To unify the dualism between rapid, unconscious reasoning and hypothesis-driven analysis, the dual-process theory has been proclaimed as a "universal model of diagnostic reasoning" (Croskerry, 2009) in which the pattern recognition model acts as the fast, non-analytical route (System I) and the hypothetico-deductive reasoning model as deliberate, analytical reasoning (System II). More recent approaches propose the cognitive zipper model (Yazdani & Hoseini Abardeh, 2020) or the predictive brain model (Lim, 2021). The former explains interpersonal variance in diagnostic behavior by the degree to which medical knowledge and the evolving problem representation of the patient can be merged to infer and validate diagnostic hypotheses. The latter reduces clinical reasoning to predictive error processing as iterative matching of top-down expectations based on illness scripts and bottom-up observed cues similar to inductive foraging (Donner-Banzhoff & Hertwig, 2014).

Diagnostic errors are often studied in conjunction with cognitive biases. Graber, Franklin, and Gordon (2005) reviewed cases of misdiagnosis for system-related and cognitive factors and apart from a high prevalence of cognitive misconceptions (5.9 per case on average), errors mainly originated from faulty information processing and hypothesis verification. Croskerry (2003) and Saposnik et al. (2016) highlight the relevance of cognitive biases by enumerating commonly observed candidates like anchoring, availability bias, confirmation bias, premature closure, and overconfidence. In an fMRI study, Melo et al. (2017) found that monitoring activity in the frontoparietal attention network decreased once information reduced uncertainty about the diagnosis and thus may be one cause for premature closure. Another assumption is that the non-analytical route of reasoning may be prone to error and facilitate biased decisions (Croskerry, 2013).

Studies in the field of Medicine are mainly concerned with evaluating diagnostic reasoning or diagnoses. A basic approach is a case presentation in written or visual form with a subsequent multiple choice test (Dekhtyar et al., 2022; Y. Cheng, Yen, Chen, Chen, & Hiniker, 2018). These approaches omit tracking of the reasoning process and aim to score outcome performance. Other settings use explanation tasks (Arocha, Wang, & Patel, 2005) in which physicians are presented with a full case report and then instructed to enumerate all remembered findings in a free recall task, to provide a pathophysiological explanation of the observed findings, and to commit to one or multiple differential diagnoses. Others track reasoning trajectories with think-aloud protocols either during case presentation at specific points of partial information (Gupta et al., 2021) or after the diagnosis has been made (Coderre et al., 2003; Soh et al., 2020). These protocols are collected via in-person interviews in which the physician are instructed to report on their reasoning trajectory. Charlin et al. (2012) tried to unravel the complexity of the process by self-assessments of experienced practitioners during recurring discussion rounds to derive consensus on essential cognitive states and actions during reasoning.

Methods like eye-tracking were used to gain insights into the sequence in which information about symptoms is processed (Scholz et al., 2017). The ability to identify and interpret salient features within a case presentation was measured by asking medical students to infer - based on the information given - the two most likely differential diagnoses and to list features that justify them (Groves, Scott, & Alexander, 2002). Case presentation is mostly given as written reports, as video recordings, or as standardized patients mimicked by actors (Fürstenberg et al., 2020; Soh et al., 2020). Most similar to ours are studies utilizing virtual patients in e-learning settings for medical students. Hege, Kononowicz, Kiesewetter, and Foster-Johnson (2018) use a web-based tool to display a complete patient summary with text descriptions, tables, images, and videos. Participants can freely add identified findings, differential diagnoses, tests/examinations, and treatment suggestions to a concept map.

In sum, many approaches emphasize some fixed opinion or "product" at the end of reasoning, e.g. differential diagnosis, concept maps, or pathophysiological explanations. Gupta et al. (2021) stated that their work "is the only study to evaluate clinical decision making *during* the evolution of a case presentation in hospital medicine physicians." (p. 10). Think-aloud protocols may be used to unravel opinion development, but are costly to analyze and may suffer from post-hoc rationalization (Summers, 2017). Our approach aims to facilitate a comprehensive and efficient analysis of this evolution by capturing information exploration and hypothesis revision online and fully automated.

Monitoring Tool

The monitoring tool consists of two building blocks. First, a collection of artificial clinical vignettes has been defined based on electronic health records and case reports. Second, these patients are made queryable through a web-based monitoring tool.

Clinical Vignettes and Simulated Patients

Cases are given as patients defined in terms of artificial clinical vignettes. A clinical vignette is a set of variable-value pairs for each queryable information ranging from biographic information and medical history to current complaints and results of medical examinations. Each piece of information is associated with a certainty on a 4-point Likert scale and a patient response that formulates the value of the variable in natural language. All vignettes consist of the same 192 variables with values chosen to fit the particular medical case. In total, we constructed 10 artificial patients suffering from either an epileptic seizure, a psychogenic seizure, a syncope, a non-epileptic sleep disorder, or a metabolic disorder (2 patients for each disease).

Artificial (instead of actual) vignettes were used as they provide a valid tool to assess reasoning in a standardized manner (Veloski, Tai, Evans, & Nash, 2005; Hege et al., 2007). We deliberately built complex patients for the vignettes to challenge the experts and induce analytical thinking. All vignettes are based on actual medical case reports (Haji Seyed Javadi, Hajiali, & Nassiri Asl, 2014; Hellmich, 2020; Gerlach & Bickel, 2021) and confidential electronic health records of patients at the Ruhr-Epileptology in Bochum. They have been validated for internal soundness by medical experts before the study.

Interactive Diagnosis and Monitoring

The diagnosis of a single patient is carried out in a dedicated user interface (see Figure 1). It starts with a self-report of the patient about their age, sex, and initial complaints. Then, the physician can prompt the patient for biographic information, medical history, current complaints, and other typical anamnesis questions, conduct sophisticated examinations like a tilt table test, and request lab reports like a blood gas analysis by entering the name of a specific variable in the clinical vignette (e.g. seizure duration). All of this information can be requested at any time and in any order. Due to this active search, we induce exploration behavior of potentially important information. Given the large number of possibilities, physicians need to reason about which information to acquire next.

The user interface understands 522 aliases to account for potential synonyms or slight differences between names of the same variable (e.g. seizure duration vs. seizure length) and displays suggestions based on the text entered to ensure that users find what they are looking for. Suggestions for detected aliases are restricted to close matches to not suggest a variable the physician did not have in mind initially. By entering written text, physicians have to actively think about possible information queries instead of browsing through possibilities as in (Kiesewetter et al., 2020). All queried information is immediately displayed with attached patient responses or examination outcomes.

After each information query, the physician is pinged to update the list of hypotheses or to confirm to leave it unchanged. This step ensures that the currently recorded hypothesis state remains up to date over the course of the trajectory. Additionally, physicians can freely take written notes and refine their differential diagnoses by adding, removing, or revising active hypotheses via drag & drop at any time. The list of possible diagnostic options was pre-defined and included all medically viable options for seizures (Benbadis, 2009). A fixed set was defined because the differential diagnoses within seizure diagnosis are well-known among practitioners. The hard task is to think about the right hypothesis at the right time rather than memorizing all possibilities (Brush et al., 2017).

Once a physician feels confident to commit to one diagnosis, they trigger the end of the process, select the diagnosis from the currently active hypotheses, and enter a free-text justification for their decision. This ensures that participants can freely decide how much information they need. Ending information exploration is thus also an action (Wilson, Bonawitz, Costa, & Ebitz, 2021). During this whole process, the server tracks every action performed by the physician. Recorded data includes information about the physician, initially given information, the true diagnosis, all actions with timestamps, the predicted diagnosis, and its justification.

Evaluation and Validation Study

To validate our approach to elicit and analyze diagnostic reasoning trajectories, we conducted a small-scale study from November 2023 to January 2024 in cooperation with staff members of the Ruhr-Epileptology in Bochum, the Epileptologicum in Hamburg, and the Department of Neurology and Epilepsy Center at Klinikum Osnabrück. 4 expert epileptologists participated in the study (3 male, 1 female, 49.5 ± 6.1 years old) with a reported 19.0 ± 3.5 work years of experience in seizure diagnosis. Participant names were pseudonymized (WA: Specialist Physician, 13 years experience; BG: Senior Physician, 22 years experience; EJ: Senior Physician, 21 years experience; RB: Chief Physician, 20 years experience). Study participation was completely voluntary and no expenses were offered beforehand or provided afterward.

Procedure

The study was carried out online. Each participant joined a Zoom meeting and received a pseudonymous account to log into the server. All participants watched the same explanation video of the diagnostic user interface exemplified by a tutorial patient with either a broken leg or arm. They entered personal information about their age, sex, position within their respective medical facility, major work area, years of diagnostic expertise, experience with epileptic seizures, psychogenic seizures, syncopes, and technology in general. Then, each

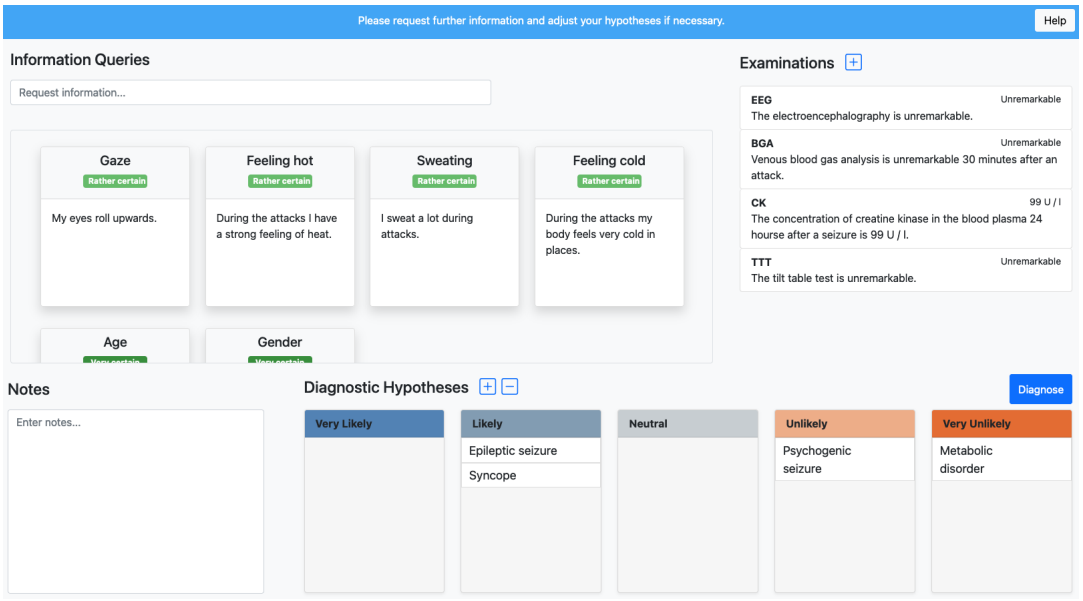

Figure 1: Diagnosing is carried out in a responsive UI, where physicians can enter anamnestic questions and receive corresponding patient responses (top left), conduct specific examinations (top right), take notes on the medical case (bottom left) and hypothesize about the diagnosis by adding or removing hypotheses and revisioning their likelihood (bottom right).

participant diagnosed the same tutorial patient that was shown in the explanation video to gain hands-on experience with the monitoring tool, followed by the 10 case vignettes described above in a unique, randomized order. After each diagnosis, technical feedback was gathered on whether information or a hypothesis was missing or any failure in the program occurred. Finally, a performance report was displayed indicating correct and incorrect diagnoses for each clinical vignette. Evaluation at this point targets the internal validity of the monitoring tool and its ability to elicit and record exploration behavior and opinion revision.

The study ended with a video-recorded semi-structured interview to gather the impression experts had during the interaction with the tool. Questions targeted the general impression, correctness of the information, excess or missing information, patient consistency, missing diagnoses, usability of the user interface, technical problems, and realism of the simulation. We used this open-ended data collection method rather than strict survey responses to be able to freely and exhaustively identify strengths and weaknesses from the viewpoint of an experienced practitioner. This evaluation targets the external validity of the monitoring tool and its ability to generalize to medical diagnostic reasoning over seizure-like events. The study procedure was approved before data collection by the Ethics Review Board of Bielefeld University.

Results

The study yielded 40 diagnostic trajectories (Fig. 2) and 4 feedback interviews. We will first present the collected trajectories, analyze exploration behavior and hypothesis revision, and then summarize the feedback interviews.

Information Exploration

115 out of the 192 variables in each clinical vignette were queried in at least one trajectory, with a remarkable difference between physicians (BG: 46, WA: 55, RB: 62, EJ: 95) and diseases (epileptic seizure: 47, metabolic disorder: 68, non-epileptic sleep disorder: 79, psychogenic seizure: 70, syncope: 82). Concerning the exploration of specific patients, two physicians agreed on the relevance of information (i.e. the decision of seeking or ignoring it) during one case in on average 87% of all variables (WA-BG: 88.7%, WA-RB: 88.8%, WA-EJ: 84.5%, BG-RB: 89.3%, BG-EJ: 86.3%, RB-EJ: 84.6%). Additionally, we observed a personal preference during information exploration in the proportion of queried anamnestic information as opposed to examination reports (WA: 58.6%, BG: 84.2%, RB: 77.5%, EJ: 83.0%).

Hypothesis Revision and Diagnostic Accuracy

During a diagnostic trajectory, physicians revisioned their hypotheses 4 times on average (BG: 3.0, EJ: 6.6, RB: 2.8, WA: 3.7). The targeted diseases were most often the most common causes epileptic seizure, psychogenic seizure, or syncope (BG: 83.3%, EJ: 97.0%, RB: 96.4%, WA: 75.7%). The list of hypotheses comprised 2.4 differential diagnoses on average (BG: 3.0 ± 1.8, EJ: 2.5 ± 0.5, RB: 1.5 ± 0.7, WA: 2.5 ± 1.1) and once under consideration, no hypothesis was ever deleted from this list. The time at which revisions are performed along a trajectory is subject to major interpersonal variance. While EJ generates hypotheses almost immediately after hearing the initial complaints of the patient, BG explores information without a leading hypothesis in mind and constructs their opinion close to the end of the trajectory. RB

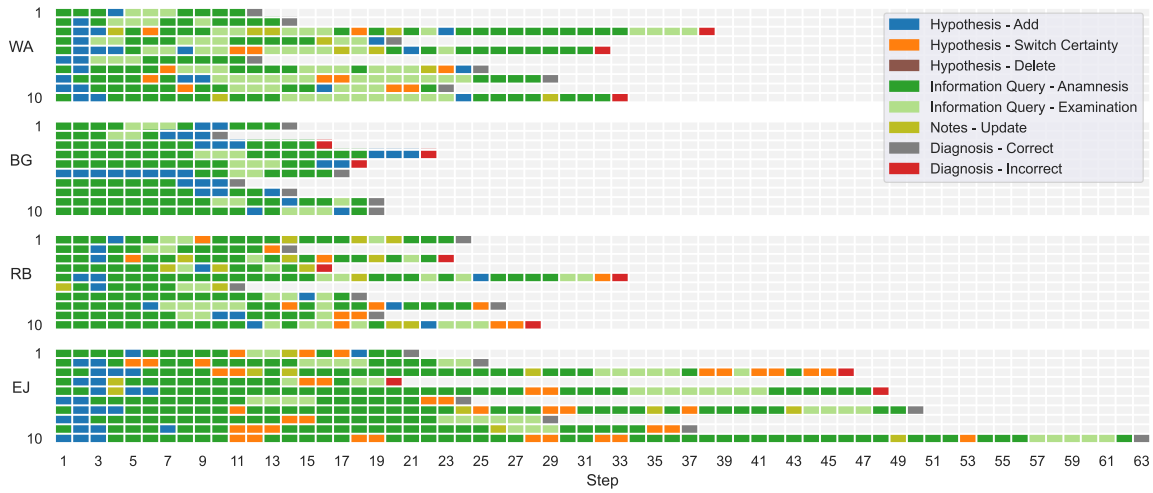

Figure 2: Action sequences of all 40 collected trajectories. Each row corresponds to one diagnostic process trajectory, each column to one step within a trajectory, and the color of each cell indicates the performed action. The four blocks correspond to the four participants marked with their pseudonyms. Within each block, the 10 patients are ordered equally, i.e. row one in block WA and row one in block BG are two diagnostic trajectories for the same patient from different physicians.

adds hypotheses during exploration and refines the differential at the end, while WA extends the list of hypotheses early and refines their opinion mostly at an intermediate stage of diagnosis. The overall accuracy is 67.5% (WA: 70%, BG: 70%, EJ: 70%, RB: 60%). The most prevalent diseases are identified most successfully (epileptic seizures: 100%, psychogenic seizures: 100%, syncopes: 75%) while all physicians struggle with less common causes (metabolic disorders: 12.5%, non-epileptic sleep disorders: 50%). Metabolic disorders and syncopes are misdiagnosed as epileptic or psychogenic seizures. Mimics for non-epileptic sleep disorders comprise psychogenic seizures, syncopes, and paroxysmal kinesigenic dyskinesia.

Exemplary Reasoning Comparison

As an example, we compare trajectories of WA and EJ diagnosing patient 10 suffering from syncopes (see Fig 3). Both physicians generate initial hypotheses at the start of diagnosing. EJ adds the most common diseases epileptic seizure, psychogenic seizure, and syncope on neutral certainty, while WA adds epileptic seizure and syncope with the former being likely and the latter unlikely. EJ starts exploring the information in small-sized chunks of 5 to 8 anamnestic variables while updating two hypotheses after each chunk and thus revises their opinion regularly. The trajectory ends with a longer "reassurance" phase, where information on 27 variables is gathered and only one hypothesis is updated by downgrading its certainty. Examinations are only queried in the last steps of the trajectory. EJ commits to the correct diagnosis and justifies it by enumerating three salient variables. WA on the other hand pauses their exploration of 26 variables once to add psychogenic seizure as a new hypothesis right after receiving 10 examination reports and then commits to the recently added hypothesis as (incorrect) diagnosis. The justification frames it as a suspected diagnosis in the absence of truly predictive semiological elements and claims that the long-term EEG should carry more information when taking the high frequency of seizures into account.

Semi-Structured Interviews

Video transcription followed by a semantic approach to inductive thematic analysis (Braun & Clarke, 2006) revealed two main themes within the interviews: *program usability* and *clinical realism*. The usability was rated as easy to understand and impressive. No participants experienced technical errors but three participants mentioned the tediousness of querying information step by step instead of receiving bundled information like anamnestic reports or when "letting the patient tell their story". The clinical vignettes were judged as a realistic reflection of patients where responses are often fuzzy and examination reports rarely warrant an immediate diagnostic decision. No information within each vignette was seen as redundant or unnecessary. Critique mainly targeted the missing detail in eyewitness reports and the fuzzy formulation of EEG reports. Three participants judged the options for diagnostic hypotheses as sufficient up to even too detailed, while another one was missing a vestibular organ disorder as an option. Two participants expressed skepticism towards the immediacy of hypothesis updates and said the diagnosis was complicated due to an active search for information. Other critiques revolved around the inability to issue follow-up questions for queries and to phrase queries in natural language. Nonetheless, all participants judged the program as a realistic formalization of a clinical diagnosis considering the aforementioned improvements and the methodical difference of searching for information via keyboard rather than talking to an actual patient.

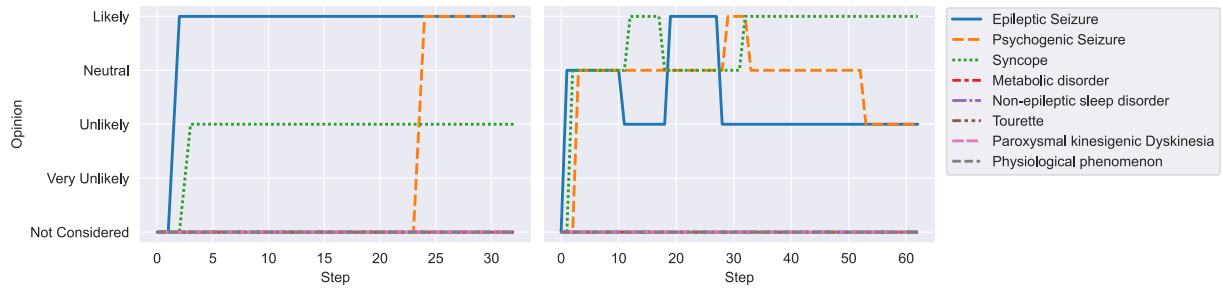

Figure 3: Two reasoning trajectories by physician WA (left) and EJ (right) for patient 10 suffering from syncope.

Discussion

The results from the evaluation study demonstrate the feasibility and validity of our tool for monitoring the diagnostic reasoning of medical experts and for revealing interesting findings about it. For one, although all participants share similar expertise and we see a tendency towards common diseases as diagnostic hypotheses, a high interpersonal variance is evident concerning which information is queried, when and how hypotheses are updated, and how much information is needed to commit to a final diagnosis. Participants WA and EJ show major similarities to the hypothetico-deductive reasoning model (Elstein et al., 1978). RB and BG share a longer exploration phase before adding diagnostic candidates, with BG showcasing a more extreme form.

In terms of diagnostic accuracy, BG performs most efficiently on the information available. Whether there is a causal connection between this reasoning approach and diagnostic performance should be clarified in subsequent studies, especially taking the pattern recognition model (Barrows & Feltovich, 1987) and the differences in interpersonal knowledge structures into account. With an overall accuracy of 67.5%, the performance of physicians is lower than expected (Chowdhury et al., 2008), but this artifact may be generated by the skewed disease prevalence in our study. Participants performed most reliably on the three differential diagnoses most common in their medical subfield while the patient population in our study was uniform. Rather than extrapolating to the performance of epileptologists, we thus acknowledge that decision support has a high potential in ameliorating the identification of rare causes apart from day-to-day business.

Interesting about the reasoning comparison is that both physicians generated a correct hypothesis right at the start - but then diverged in their reasoning. EJ followed the classic schema of hypothetico-deductive reasoning and challenged two concurring differential diagnoses at each time to iteratively filter out the best option. This procedure is standard and promoted by course books on medical decision-making (Sox et al., 2013, p. 17f). WA failed to update previous hypotheses and instead added a new option that eventually became the final diagnosis although both have been evaluated as equally likely at this step. This diagnostic error may therefore be rooted in biased hypothesis evaluation.

This is but one example where biased behavior can be localized within the reasoning trajectory. Other prominent biases reported by Saposnik et al. (2016) can be detected as well: Adding a new hypothesis and clinging to its likelihood despite gradually increasing contradictory evidence (anchoring bias), expressing a focus on the most prevalent diseases while neglecting viable but less common differential diagnoses (availability bias), only querying information that is highly associated with the lead hypothesis (confirmation bias), committing to a hypothesis after a short exploration phase that does not warrant a diagnostic decision (premature closure) or higher subjective certainties than the given partial information would support (overconfidence).

A strong limitation of our study is the small number of participants. Gathering a significantly larger pool of trajectories would enable in-depth quantitative analysis. However, recruiting participants within the small subset of eligible physicians with matching medical background and proficiency is no easy task but necessary to strengthen our conclusions. Another limitation lies in the deliberate, methodically motivated distortion of the interaction setting between physician and patient. Seizure diagnosis is a deeply personal conversation touching on a broad range of anamnestic fields as well as personal issues like drug addiction and psychological trauma with people that have been suffering from their disease for decades potentially. These contextual factors were purposely excluded to focus on the isolated cognitive assessment of reasoning abilities given partial information.

Conclusion

This paper investigated whether information about cognitive processes in diagnostic reasoning can be elicited, gathered, and analyzed in a standardized, scalable, and efficient manner. By implementing a computer-based monitoring tool that enables physicians to carry out diagnoses on pre-defined clinical vignettes, we were able to collect highly detailed diagnostic reasoning trajectories that can be analyzed quantitatively and qualitatively. Results from a first evaluation study suggest that the collected trajectories can capture the fine-grained dynamics of the reasoning process and, e.g., reveal interpersonal variance and biased decisions. The data replicates previous findings and theories of diagnostic reasoning, while additionally helping to shed light on the step-by-step cognitive dynamics within the mind of expert physicians.

Acknowledgments

Funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

We would like to thank Jan-Malte Giannikos, Jo Henning Ermshaus and Ilja Abramov for their reliable assistance during the implementation of the server.

References

Arocha, J. F., Wang, D., & Patel, V. L. (2005). Identifying reasoning strategies in medical decision making: A methodological guide. *Journal of Biomedical Informatics*, *38*(2), 154–171. doi: 10.1016/j.jbi.2005.02.001

Banda, S. (2010). Overview of diagnostic reasoning: Hypothetical-deductive strategy, problem representation, semantic qualifiers, illness scripts, pattern recognition and prototypes. *Medical Journal of Zambia*, *36*(3). doi: 10.4314/mjz.v36i3.56074

Barrows, H. S., & Feltovich, P. J. (1987). The clinical reasoning process. *Medical Education*, *21*(2), 86–91. doi: 10.1111/j.1365-2923.1987.tb00671.x

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and Investigative Medicine. Medecine Clinique Et Experimentale*, *5*(1), 49–55.

Benbadis, S. (2009). The differential diagnosis of epilepsy: A critical review. *Epilepsy & Behavior*, *15*(1), 15–21. doi: 10.1016/j.yebeh.2009.02.024

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi: 10.1191/1478088706qp063oa

Brush, J. E., Sherbino, J., & Norman, G. R. (2017). How Expert Clinicians Intuitively Recognize a Medical Diagnosis. *The American Journal of Medicine*, *130*(6), 629–634. doi: 10.1016/j.amjmed.2017.01.045

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., … Bourdy, C. (2012). Clinical reasoning processes: unravelling complexity through graphical representation: Clinical reasoning: graphical representation. *Medical Education*, *46*(5), 454–463. doi: 10.1111/j.1365-2923.2012.04242.x

Cheng, L., & Senathirajah, Y. (2022). Testing Medical Student Diagnostic Reasoning Using Clinical Data Visualizations. In B. Séroussi et al. (Eds.), *Studies in Health Technology and Informatics.* IOS Press. doi: 10.3233/SHTI220596

Cheng, Y., Yen, K., Chen, Y., Chen, S., & Hiniker, A. (2018). Why doesn't it work? Voice-driven interfaces and young children's communication repair strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children (IDC'18)* (pp. 337–348). Trondheim, Norway: ACM. doi: 10.1145/3202185.3202749

Chowdhury, F. A., Nashef, L., & Elwes, R. D. C. (2008). Misdiagnosis in epilepsy: a review and recognition of diagnostic uncertainty. *European Journal of Neurology*, *15*(10), 1034–1042. doi: 10.1111/j.1468-1331.2008.02260.x

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, *37*(8), 695–703. doi: 10.1046/j.1365-2923.2003.01577.x

Croskerry, P. (2003). The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. *Academic Medicine*, *78*(8), 775–780. doi: 10.1097/00001888-200308000-00003

Croskerry, P. (2009). A Universal Model of Diagnostic Reasoning. *Academic Medicine*, *84*(8), 1022–1028. doi: 10.1097/ACM.0b013e3181ace703

Croskerry, P. (2013). From Mindless to Mindful Practice — Cognitive Bias and Clinical Decision Making. *New England Journal of Medicine*, *368*(26), 2445–2448. doi: 10.1056/NEJMp1303712

Croskerry, P., Campbell, S. G., & Petrie, D. A. (2023). The challenge of cognitive science for medical diagnosis. *Cognitive Research: Principles and Implications*, *8*(1), 13. doi: 10.1186/s41235-022-00460-z

Custers, E. J., Boshuizen, H. P., & Schmidt, H. G. (1998). The Role of Illness Scripts in the Development of Medical Diagnostic Expertise: Results From an Interview Study. *Cognition and Instruction*, *16*(4), 367–398. doi: $10.1207/s1532690xci1604_1$

Dekhtyar, M., Park, Y. S., Kalinyak, J., Chudgar, S. M., Fedoriw, K. B., Johnson, K. J., … Stern, S. (2022). Use of a structured approach and virtual simulation practice to improve diagnostic reasoning. *Diagnosis*, *9*(1), 69–76. doi: 10.1515/dx-2020-0160

Donner-Banzhoff, N. (2022). *Die ärztliche Diagnose: Erfahrung - Evidenz - Ritual* (1. Auflage ed.). Bern: Hogrefe. doi: 10.1024/86194-000

Donner-Banzhoff, N., & Hertwig, R. (2014). Inductive foraging: Improving the diagnostic yield of primary care consultations. *European Journal of General Practice*, *20*(1), 69–73. doi: 10.3109/13814788.2013.805197

Elstein, A. S. (2009). Thinking about diagnostic thinking: a 30-year perspective. *Advances in Health Sciences Education*, *14*(S1), 7–18. doi: 10.1007/s10459-009-9184-0

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, Mass: Harvard University Press. (OCLC: 1165494938)

Fürstenberg, S., Helm, T., Prediger, S., Kadmon, M., Berberat, P. O., & Harendza, S. (2020). Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. *BMC Medical Education*, *20*(1), 368. doi: 10.1186/s12909-020-02260-9

Gerlach, R., & Bickel, A. (2021). *Fallbuch Neurologie* (5., unveränderte Auflage ed.). Stuttgart New York: Georg Thieme Verlag. doi: 10.1055/b000000427

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic

Error in Internal Medicine. *Archives of Internal Medicine*, *165*(13), 1493. doi: 10.1001/archinte.165.13.1493

Groves, M., Scott, I., & Alexander, H. (2002). Assessing clinical reasoning: a method to monitor its development in a PBL curriculum. *Medical Teacher*, *24*(5), 507–515. doi: 10.1080/01421590220145743

Gupta, A., Quinn, M., Saint, S., Lewis, R., Fowler, K. E., Winter, S., & Chopra, V. (2021). The variability in how physicians think: a casebased diagnostic simulation exercise. *Diagnosis*, *8*(2), 167–175. doi: 10.1515/dx-2020-0010

Haji Seyed Javadi, S. A., Hajiali, F., & Nassiri Asl, M. (2014). Zolpidem Dependency and Withdrawal Seizure: A Case Report Study. *Iranian Red Crescent Medical Journal*, *16*(11). doi: 10.5812/ircmj.19926

Hege, I., Kononowicz, A. A., Kiesewetter, J., & Foster-Johnson, L. (2018). Uncovering the relation between clinical reasoning and diagnostic accuracy – An analysis of learner's clinical reasoning processes in virtual patients. *PLOS ONE*, *13*(10), e0204900. doi: 10.1371/journal.pone.0204900

Hege, I., Ropp, V., Adler, M., Radon, K., Mäsch, G., Lyon, H., & Fischer, M. R. (2007). Experiences with different integration strategies of case-based e-learning. *Medical Teacher*, *29*(8), 791–797. doi: 10.1080/01421590701589193

Hellmich, B. (Ed.). (2020). *Fallbuch Innere Medizin* (6th ed.). Stuttgart: Georg Thieme Verlag. (Pages: b-007-170975) doi: 10.1055/b-007-170975

Kiesewetter, J., Sailer, M., Jung, V. M., Schönberger, R., Bauer, E., Zottmann, J. M., … Fischer, M. R. (2020). Learning clinical reasoning: how virtual patient case format and prior knowledge interact. *BMC Medical Education*, *20*(1), 73. doi: 10.1186/s12909-020-1987-y

Koufidis, C., Manninen, K., Nieminen, J., Wohlin, M., & Silén, C. (2021). Unravelling the polyphony in clinical reasoning research in medical education. *Journal of Evaluation in Clinical Practice*, *27*(2), 438–450. doi: 10.1111/jep.13432

Kumar, B., Ferguson, K., Swee, M., & Suneja, M. (2021). Diagnostic Reasoning by Expert Clinicians: What Distinguishes Them From Their Peers? *Cureus*. doi: 10.7759/cureus.19722

Lim, T. K. (2021). The predictive brain model in diagnostic reasoning. *The Asia Pacific Scholar*, *6*(2), 1–8. doi: 10.29060/TAPS.2021-6-2/RA2370

Malmgren, K., Reuber, M., & Appleton, R. (2012). Differential diagnosis of epilepsy. *Oxford textbook of epilepsy and epileptic seizures*, 81–94. doi: 10.1093/med/9780199659043.003.0008

Melo, M., Gusso, G. D. F., Levites, M., Amaro, E., Massad, E., Lotufo, P. A., … Friston, K. J. (2017). How doctors diagnose diseases and prescribe treatments: an fMRI study of diagnostic salience. *Scientific Reports*, *7*(1), 1304. doi: 10.1038/s41598-017-01482-0

Plug, L., & Reuber, M. (2009). Making the diagnosis in patients with blackouts: it's all in the history. *Practical Neurology*, *9*(1), 4–15. doi: 10.1136/jnnp.2008.161984

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*, *16*(1), 138. doi: 10.1186/s12911-016-0377-1

Schmidt, H. G., & Volder, M. L. d. (1984). *Tutorials in problem-based learning: new directions in training for the health professions*. Assen [Netherlands]: Van Gorcum. (OCLC: 12721622)

Scholz, A., Krems, J. F., & Jahn, G. (2017). Watching diagnoses develop: Eye movements reveal symptom processing during diagnostic reasoning. *Psychonomic Bulletin & Review*, *24*(5), 1398–1412. doi: 10.3758/s13423-017-1294-8

Soh, M., Konopasky, A., Durning, S. J., Ramani, D., McBee, E., Ratcliffe, T., & Merkebu, J. (2020). Sequence matters: patterns in task-based clinical reasoning. *Diagnosis*, *7*(3), 281–289. doi: 10.1515/dx-2019-0095

Sox, H. C., Higgins, M. C., & Owens, D. K. (2013). *Medical decision making* (2nd ed ed.). Chichester, West Sussex, UK : Hoboken, New Jersey: John Wiley & Sons. (Medium: electronic resource)

Summers, J. S. (2017). *Post hoc ergo propter hoc* : some benefits of rationalization. *Philosophical Explorations*, *20*(sup1), 21–36. doi: 10.1080/13869795.2017.1287292

Veloski, J., Tai, S., Evans, A. S., & Nash, D. B. (2005). Clinical Vignette-Based Surveys: A Tool for Assessing Physician Practice Variation. *American Journal of Medical Quality*, *20*(3), 151–157. (_eprint: https://doi.org/10.1177/1062860605274520) doi: 10.1177/1062860605274520

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. doi: 10.1016/j.cobeha.2020.10.001

Xu, Y., Nguyen, D., Mohamed, A., Carcel, C., Li, Q., Kutlubaev, M. A., … Hackett, M. L. (2016). Frequency of a false positive diagnosis of epilepsy: A systematic review of observational studies. *Seizure*, *41*, 167–174. doi: 10.1016/j.seizure.2016.08.005

Yazdani, S., & Hoseini Abardeh, M. (2020). A novel model of clinical reasoning: Cognitive zipper model. *Journal of Advances in Medical Education & Professionalism*, *8*(2). doi: 10.30476/jamp.2020.82230.1050