Contributions of Psychology to the Design of Diagnostic Decision Support Systems

Gitte Lindgaard, Janette Folkens, Catherine Pyper, Monique Frize, Robin Walker

Contributions of psychology to the design of diagnostic decision support systems

Gitte Lindgaard[1], Janette Folkens[1], Catherine Pyper[1],Monique Frize[1] and Robin Walker*
[1]Carleton University, Ottawa, ON K1S 5B6, Canada; *IWK Health Centre, 5850/5980 University Avenue, Halifax, NS B3K 6R8, Canada

Abstract: This paper describes how psychological research can contribute to the requirements engineering, the design and usefulness of a Diagnostic Decision Support System (DDSS) intended to support pediatric residents' diagnostic decisions. Research on cognitive biases in Bayesian decision tasks is discussed. The design of the DDSS is briefly outlined, and a formative usefulness test is reported. Under the assumption that a particular cognitive bias could be overcome by showing it to participants, pediatric residents were given a set of Bayesian decision tasks. One half was given an opportunity to interact with NeoPeDDS and the other half was not. Results showed that NeoPedDDS usage improved the accuracy of the diagnostic decisions, but that formal training in Bayesian statistics appears to be necessary for residents to evaluate ambiguous information in a normatively correct manner.

Keywords: Bayes' Theorem, human decision making, probabilities, Diagnostic Decision Support Systems, diagnosticity

1. Introduction

Numerous tools and techniques have evolved to support usability engineers in the user requirements gathering process [1,2,3]. One assumption in most of these is that the target users are already known along with their background, tasks and goals, the tools they currently use, and the context in which they will use the new application. It is also often assumed that these techniques suffice to derive the detailed understanding needed to design an effective application. Similarly, many knowledge elicitation techniques exist in Artificial Intelligence [4] to elucidate experts' mental models underlying their tacit problem solving and decision processes in complex, often highly dynamic, situations. Consequently, the design of many expert systems rests on the assumption that human performance is inevitably optimal and hence worth emulating [5,6]. This paper describes how understanding derived from fundamental psychological decision research was applied to the initial design of a prototype for a diagnostic decision support system in neonatology.

The next section introduces decision issues in medical diagnosis, highlighting the need for decision support systems to assist physicians processing ambiguous diagnostic information. It is followed by a brief discussion of the most popular medical decision support systems, and then by an outline of how probabilities relate to medical diagnosis in Bayesian decision tasks. The notion of diagnosticity is then discussed, leading into a short description of the creation of the decision support system called NeoPeDDS, and subsequently, the usefulness test of NeoPeDDS. Finally, the next steps in this research program are outlined, and concluding remarks are made.

2. Decision issues in medical diagnosis

As the queues of patients needing urgent medical attention are growing rapidly world-wide, medical practitioners are under increasing pressure quickly to formulate a diagnosis and initiate treatment. Yet, diagnostic decisions are very complex. Patients rarely present with a clear, unambiguous clinical picture as described in medical textbooks. Many patients suffer from multiple diseases, and they may not know which symptoms are most important to report. In addition, many of the symptoms patients display may be ambiguous, equally indicative of different diseases. For neonatologists who deal with newborn babies, this complexity is further exacerbated because sick infants cannot say how they feel or where it hurts. Neonatology is therefore an attractive domain in which to provide effective computer-based Diagnostic Decision Support Systems (DDSSs).

Given this complexity, it is not surprising that misdiagnosis is a recognized problem in medicine. Figures released by the Institute of Medicine (IOM) on the annual number of deaths of hospitalized patients due to some kind of medical error are estimated to lie between 98,000 [7] and 115,000 [8]. It is believed that many of those deaths could be prevented [9,10,11]. However, figures vary widely. One recent review [12] found diagnosis-related errors to account for 10-30% of all errors recorded. Others estimate the proportion to be up to 76% [13]. Figures from autopsies have consistently yielded a misdiagnosis rate of 40% over the past 65years [6], but autopsies are no longer conducted routinely. The accuracy of recent figures is therefore debatable as are both the definition and the calculation of "preventable error" [12]. However, regardless of how these are counted, the number of misdiagnoses is high. The reasons are said to range from macro-level health system related problems [13] to micro-level cognitive errors [9]. Empirical evidence suggests that medical diagnosticians find it difficult to deal with ambiguous data [14]. This was the problem addressed in the design of the prototype introduced here, called NeoPeDDS. It was informed by research highlighting some of the cognitive biases to which human decision making, including medical diagnostic decisions, are prone.

3. Decision support systems in pediatrics

Several DDSSs supporting pediatric decision making have emerged. The most popular of these are ePocrates and Isabel. ePocrates is an impressive DSS based on a comprehensive database of diseases, treatments, drugs, and much more [5]. It runs on virtually any mobile device including iPods and the BlackBerry. The online version provides instant access to a wide range of information and services. The database is updated regularly. It enables the diagnostician to compare the occurrence of symptoms in different diseases. However, it does not provide frequencies of occurrence of symptoms, making it impossible for the physician to use the information to weigh the probabilities of different diseases against one another. Isabel is a web-based natural-language DDSS that aims to reduce diagnostic errors. It applies word-matching searches through unformatted medical texts to arrive at a list of diagnoses. It provides a list of possible diagnoses in response to symptoms and other user-entered clinical findings [15,16]. Several studies, e.g. [17] and [18] have showed that Isabel can lead pediatricians to diagnoses they would otherwise not have considered. However, the output can be overwhelming, as Isabel provides up to 10 diagnostic categories, each of which may point to up to 35 different diseases. Each result links to a Knowledge Mobilizing System that allows perusal of medical texts about a particular disease. This can distract from the purpose of arriving at a final diagnosis quickly. Although it is popular, Isabel does not exactly yield quick diagnostic aid in a high-pressure clinical setting. Rather, it provides supplemental reading material for a more leisurely approach to diagnostic decision making where timing is not critical. One major drawback is that Isabel's suggested diagnoses appear implicitly to be equiprobable because it lacks quantified information about symptom diagnosticity. Its database is drawn from a cross-section of medical texts which do not quantify symptoms. Therefore, there is no way to calculate the frequency of occurrence of symptoms or the probability associated with different diseases in the light of the clinical picture a given sick infant presents. Such an approach continues to force clinicians to rely on personal experience and various decision heuristics when diagnosing ambiguous cases.

4. Probabilities and Bayesian decision tasks

The notion of probability is connected with the degree of belief warranted by evidence (epistemic probabilities) and with the tendency to produce stable relative frequencies (aleatory probabilities). Statistical probabilities concern the way evidence from various sources is combined into a numeric statement irrespective of the judge's belief. Epistemic probabilities incorporate an assessment of the judge's personal belief, generated from autobiographical experience and state of knowledge about the evidence. The human-generated epistemic probability reflects both arithmetic calculations and degree of belief. By contrast, a computer-generated statistical probability is an arithmetic computation of given numeric values. It is therefore unrealistic to expect the two to be identical.

Subjective beliefs are more likely to attenuate than to increase judgmental accuracy because beliefs are derived from the judge's own experience.

The output of a Bayesian analysis is a distribution of probabilities over a set of hypotheses. The model is normative in the sense that it specifies certain internally consistent relationships among probabilistic opinions that prescribe how opinions should be revised with new incoming information. Existing knowledge is summarized in prior (aleatory) probabilities, the base rates, and incoming case-specific evidence provided through individuating information. The outcome of a Bayesian analysis, the posterior probability, is calculated by combining the base rates and the individuating information. Two hypotheses, H and $\hat{H}$, assessed against one another, are expressed in the base rates such that $P(H) + P(\hat{H}) = 1.0$. The model demands that the individuating information be considered in terms of its support for both hypotheses, the weighting of which leads to the posterior probability. This weighting results in a revision of the opinion contained in the original base rates. When the evidence supports both hypotheses H and $\hat{H}$ to an equal extent, no revision should occur. The resulting posterior probability is therefore identical to the base rate representing the hypothesis in terms of which the judgment is made.

Numerous early studies in Bayesian decision making led researchers to conclude that people are, by and large, good Bayesians [19], except that they tend to revise their judgments less than demanded by the model upon receiving additional case-specific information. They were "conservative" [20]. Numerous subsequent findings refuted that early belief, showing instead that people did not behave in a Bayesian manner at all [21,22]. People were found to ignore the base rates and instead rely exclusively on the individuating information, even when that information was completely nondiagnostic. That is, it equally supported both hypotheses or none of these. Objectively, such information should be ignored; judgments should rely exclusively on the base rates.

5. The notion of diagnosticity

Diagnosticity refers to "how much potential impact a datum should have in revising one's opinion without regard to what the prior odds are" [22, p.778]. In order to determine the informativeness (diagnosticity) of the individuating information in cases where it consists of several items, a value must be assigned to each item. Early studies revealed a robust tendency of people to rely on the degree to which the individuating information was representative of the to-be-judged hypothesis. Fischhoff and Beyth-Marom's [23] and Beyth-Marom and Fischhoff's [24] explanation for this reliance on representativeness is that people judge the evidence solely by the degree to which the individuating information supports the hypothesis being entertained. They argue that people simply do not understand that diagnosticity is a measure of the relative support for both hypotheses, H and $\hat{H}$. The concept of diagnosticity is touched upon in some recent studies in the medical domain [7,24,25]. However, its importance for leading to a more accurate estimation of the posterior probabilities in the face of an ambiguous clinical picture is not strongly

emphasized. Studies involving nondiagnostic individuating information have primarily focused on posterior probability estimates. By contrast, numerous studies in our lab focusing on nondiagnosticity in occupationally relevant Bayesian tasks have repeatedly confirmed the tendency for probability estimates to resemble judgments of representativeness [46-50]. Using very carefully constructed case-specific information in which the symptoms displayed by fictitious patients suffering from one of two competing diseases, our experiments have been conducted with both nurses and physicians and a range of different diseases. Medical texts provide lists of diseases and symptoms that are likely to be observed in each disease, but they do not quantify symptoms. The fact that many symptoms may occur in different diseases is not obvious. Therefore, it was necessary to pre-test symptoms on different samples of participants to ensure that highly diagnostic symptoms were perceived to occur very frequently in one, and very infrequently in the other of the two diseases to be exposed in Bayesian tasks.

Our results have consistently revealed a primacy effect. That is, the symptom shown first in a vignette was invariably found disproportionally to affect the probability estimate. Thus, a high-diagnostic symptom supporting the to-be-judged disease presented first resulted in a significantly higher probability estimate when the same symptom was presented later. Conversely, a high-diagnostic symptom supporting the alternative disease presented first suppressed the probability estimate, with a decreasing effect when presented later. The primacy effect suggests a tendency to focus on a particular diagnosis very early in the process. Graber [27] claims that "knowledge deficits are rarely the cause of cognitive errors in medicine; these errors more commonly involve defective synthesis of the available data" (pp. 1-2). This concurs with the suggestion that diagnosticians select a single, very salient symptom right away and use it as a pivot around which they collect additional information. Such a strategy could bias the integration of information in ambiguous cases, leading to "premature closure" whereby other possible diagnoses are not considered once a hypothesis is entertained [27]. For example, the judge may ignore available data that conflict with the current hypothesis; the fact that the selected pivot may point to different diseases may not be detected if only one hypothesis is entertained. A more detailed description of these experiments may be found in Lindgaard et al. [28].

6. Creating NeoPeDDS

In terms of supporting the task of diagnosing, we assumed that a display of the most likely diseases along with their respective probabilities would raise awareness of the possibility that more than one disease could account for the constellation of a patient's symptoms. Thus, upon entering at least one symptom into NeoPeDDS and telling the system to 'diagnose', it generates a list the five most probable diagnoses, complete with the relevant probabilities. Five was chosen as the maximum to keep the diagnostic task manageable. The diagnoses are based on the World Health Organization's International Classification of Diseases (ICD10) and on a Bayesian analysis of epidemiological data.

The database comprises 97 complete records collected from some 1,200 infants admitted to the neonatal intensive care unit at the Children's Hospital of Eastern Ontario and diagnosed with respiratory distress. Respiratory distress was selected as the target condition because it occurs relatively frequently and because the signs and symptoms are ambiguous, pointing to different possible causes. The dataset enabled accurate quantification of the relative diagnosticity of each sign and symptom associated with every causal condition upon which $P(D|H)$ and $P(D|\hat{H})$ were calculated.

NeoPeDSS was developed by using Object Oriented Software Development (OOSD) and Usage Centered Design [1]. OOSD discusses requirements in terms of a use-case model, which consists of actors and use-cases. These were developed to define the possible sequences of system-actor interactions; use-case diagrams modeled the system requirements and boundaries using actors and use-cases to improve the breakdown of the system according to the user requirements. An abstract representation of the user interface was then designed. A navigation map tied the use-case narratives to a flow between interaction contexts. A content model as well as a navigation map was created from the essential use-cases and their relationships before designing the GUI prototype which was used to assess our assumption that a mere display of a set of diseases would suffice for pediatric residents to eliminate the primacy effect observed in our earlier experiments.

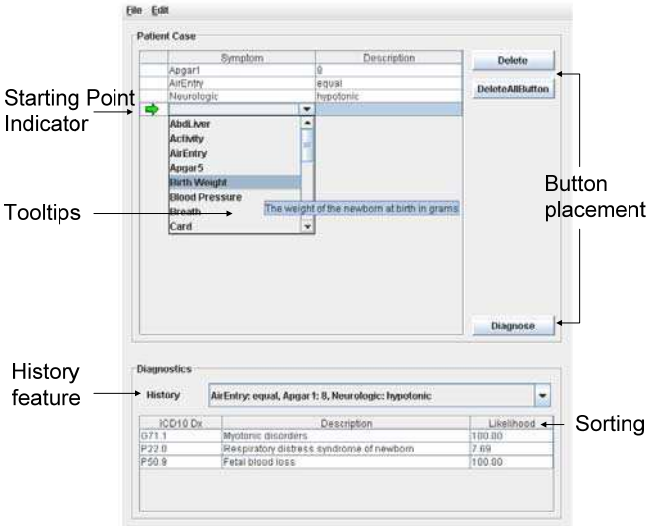

Fig. 1: NeoPeDDS GUI

Usability was assessed of the early prototype before recruiting pediatric residents in the formative usefulness test. Thus, a heuristic evaluation was conducted by two usability experts, and two empirical usability tests were performed using HCI students who were naïve with respect to both the system and pediatrics. The task scenario exposed the core

system tasks such as entering a patient case and retrieving a list of possible diagnoses as well as modifying the patient case. Problems revealed resulted in four relatively minor modifications to the prototype: (1) the layout of button locations was modified to separate the 'Diagnose' function from the 'Delete' functions to prevent accidental deletion of cases still being worked on; (2) tool tips were added; (3) a visual cue was added to indicate where to enter data, and (4) a history feature was added enabling users to retrieve an earlier case and compare it with the current results. The GUI is shown in Figure 1.

Users are able to add more symptoms to refine the probabilities even after the 'Diagnose' button has been pressed. As more information is added, the relevant probabilities are adjusted accordingly. The 'Diagnose' button may be pressed as many times as the diagnostician likes.

7. Formative usefulness test

Three major issues were addressed in the formative test in which resident pediatricians took part. First, we needed to demonstrate that NeoPeDDS could improve diagnostic accuracy. To test that, five test cases were prepared for which participants proposed a preliminary diagnosis before, and a final diagnosis after, using NeoPeDDS. Accordingly, Hypothesis 1 predicted that more correct final diagnoses would be found after than before NeoPeDDS. Second, it tested the assumption that the display of the five most probable diagnoses would increase awareness of the possibility that more than one disease should be considered. Hypothesis thus 2 predicted that base rates would be used after, but not before, exposure to NeoPeDDS. Third, to recognize the relative worthlessness of the individuating information, participants must consider both diseases. Doing so should eliminate the primacy effect found in previous studies. Hypothesis 3 therefore predicted that a primacy effect would be found in probability estimates made before, but not in those made after, exposure to NeoPeDDS.

8.1 Method

Participants: Some 40 senior resident pediatricians were recruited from various university hospitals in Canada and the United States. NeoPeDDS was presented online, enabling the participants to complete the study in their own time and in several steps if they chose. Once a case had been evaluated, they were unable to go back over it. Upon completion of the test, they received a $100 gift certificate by email.

Materials: A fictitious cover story was created to provide a plausible explanation for the limited information available about each infant to be assessed. The 24 vignettes, each describing a sick infant and containing three symptoms, were constructed such that they were nondiagnostic or near-nondiagnostic. Each vignette contained three signs or symptoms: one, either high- or low-diagnostic, supporting Respiratory Distress Syndrome

(RDS-H), another one supporting the alternative disease Ĥ, Transient Tachypnoe, (TTN-H) and the third was nondiagnostic (e.g. runny nose). In nondiagnostic vignettes, both the diagnostic symptoms were either high or low in diagnosticity. In the near-nondiagnostic vignettes, one was high, and the other low in diagnosticity, yielding HL and LH vignettes. The three symptoms were combined factorially. Participants estimated the probability of RDS-H, disease H. Another five cases were described in more detail. These required the pediatricians to propose a preliminary diagnosis before using NeoPeDDS, and a final diagnosis afterwards. NeoPeDDS was not available for their preliminary diagnosis.

Design: One half of the participants were assigned at random to the Low Base Rate group (LBR) where the RDS-H base rate was low (28/100 fictitious cases) and that of TTN-H was high (72/100). This was reversed for the other half, assigned to the High Base Rate group (HBR). The test comprised three phases as well as some pre- and post-test questions. In Phase 1 participants estimated the probability of RDS-H in each of the 24 vignettes. In Phase 2 one half (half LBR and half HBR) were shown NeoPeDDS and how it worked as well as being given an opportunity to play with it before seeing the five cases to be diagnosed. The other half read an article on diagnostic error. In Phase 3, all participants again assessed the same 24 vignettes, presented in a different random order, with different, randomly assigned names. Finally, they indicated the perceived frequency of occurrence of each of the signs and symptoms exposed in the vignettes.

Procedure: Once a pediatrician had agreed to participate, a unique login code was emailed to them. Upon logging into the private and secure site, they completed an informed consent form before gaining access to the test. They then answered the pre-experimental questions and proceeded to Phase 1. They were told that they could log out at any time during the test. If they chose to complete the test in stages, they were told that the program would remember where exactly they had been before logging out and it would take them back to the next task. Upon completion of the entire test, they were presented with a debriefing form, which they could print out, thanked for their time, and advised that a gift certificate would be sent to the email address to which their login code had been sent. All results were sent automatically to a server at Carleton University.

8. Results and Discussion

A comparison of the number of correct preliminary and final diagnoses proposed in Phase 2 was significant ($t(19) = 2.93$, $p < .008$), thereby supporting Hypothesis 1. To test Hypothesis 2 predicting that base rates would be used in judgments of the 24 vignettes after (Phase 3), but not before exposure to NeoPeDDS (Phase 1), a 2 x 2 x (2) mixed-design ANOVA was conducted for base rate groups (HBR, LBR), exposure to NeoPeDDS (exposure, no exposure), and experimental Phase (1, 3). This analysis should ideally have yielded a two-way interaction of Phase and exposure as well as a main effect for exposure and Phase respectively. None of these predicted effects occurred ($F < 2$), thereby refuting Hypothesis 2. However, contrary to the expectation that the base rates would not affect

probability estimates before exposure to NeoPeDDS, there were no interactions involving the base rate groups, and the main effect for base rate group was highly significant ($F(1,36) = 15.85$, $p < .000$). This suggests that base rates were used to some extent even before exposure to NeoPeDDS, but that this usage did not increase after exposure. The result is puzzling, as our previous research has consistently shown base rate neglect in many similar experiments [26]. Since roughly 75% of the participants were not familiar with formal decision models, this finding cannot be attributed to prior knowledge of Bayes' Theorem. Other researchers have [26] shown that people relied exclusively on the base rates when the nondiagnosticity of the individuating information was palpably worthless. As the pediatricians adjusted their probability estimates less than Bayes' Theorem demands, the evidence was apparently not deemed palpably worthless.

To test Hypothesis 3, predicting that a primacy effect would be found in the probability estimates made before, but not after, exposure to NeoPeDDS, a repeated-measures ANOVA was conducted separately for phases 1 and 3. This was justified by a non-significant t-test comparing estimates of participants who had been exposed to NeoPeDDS with those who had not ($t < 1$). The main effect for serial position was significant both for Phase 1 ($F (1, 39) = 5.45$, $p < .05$) and Phase 3 ($F (1, 39) = 10.86$, $p < .01$), as were the interactions of symptom and serial position ($F (1, 39) = 10.86$, $p < .01$ Phase 1; $F (1, 39) = 28.89$, $p < .001$ Phase 3). Figure 2 shows that estimates were higher for RDS-H in serial position 1 than in serial position 3, and that the reverse was true for TTN-H estimates, resulting in a clear convergence of estimates across serial position. Both of these findings indicate the presence of a primacy effect. As such an effect had been predicted for Phase 1 but not for Phase 3, the results partially supported Hypothesis 3.

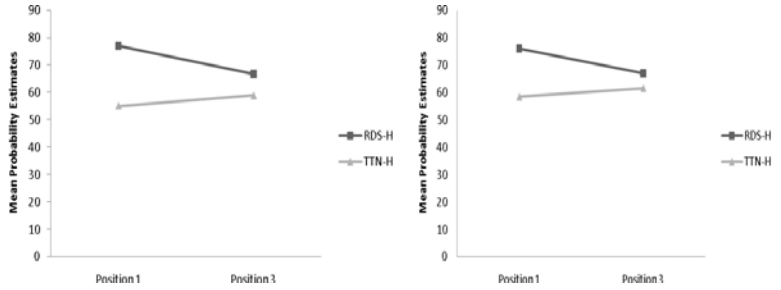

Fig. 2. The serial position effect in Phase 1 (left panel) and in Phase 3 (right panel)

One explanation of the persistent primacy effect is that individuating information is considered only in terms of the nominator, here RDS-H, as Beyth-Marom and Fischhoff [29] claimed, because they do not understand the concept of diagnosticity. Another possibility is that participants simply weighed information confirming RDS-H more heavily than information disconfirming it, perhaps because they did not know how to deal with conflicting data in the Bayesian framework. The data are insufficient unequivocally

to discern which of these possibilities may account for the results. In order to determine if they were able to assess the frequency of occurrence of the different symptoms, they were given two lists at the end of the test. For one list they were asked to indicate in how many infants out of 100, all diagnosed with RDS-H, they would expect to find each symptom. For the other list they were asked the same question, but this time the infants were said to have been diagnosed with TTN-H. The lists were identical, and they both displayed all the symptoms that featured in the short cases, albeit shown in different random orders in the two lists. The findings showed that participants were fully aware of the relevant frequencies of occurrence. Thus, for example, the H-diagnostic symptom supporting RDS-H was seen to occur very frequently in RDS-H and very infrequently in disease TTN-H, and vice versa for the H-diagnostic symptom supporting TTN-H. Apparently, they were sensitive to symptom diagnosticity, but they did not know how to combine symptoms pointing to the two diseases into a normatively correct judgment. It would therefore appear that at least some training is necessary for a DDSS based on a Bayesian algorithm such as NeoPeDDS to provide optimal assistance to physicians. One most encouraging feedback was that all but a single participant said that they would use a DDSS such as NeoPeDDS if it were made available to them.

9. Conclusions and next steps

The above findings suggest that NeoPeDDS did facilitate the task of diagnosing to some extent. However, awareness of the possibility that several diseases may account for a highly ambiguous clinical picture, did not suffice for participants to utilize the base rates optimally in their probability estimates. People may generally have a poor understanding of the concept of diagnosticity because they do not understand the relevance of the denominator term, $P(D|\hat{H})$ for the posterior probability, $P(H|D)$. Eddy [29] has shown that physicians have difficulties distinguishing between the terms $P(H|D)$ and $P(D|H)$. It is conceivable that this difficulty extends to the necessity of estimating $P(D|\hat{H})$ even when clinicians are capable of estimating the frequency of occurrence of individual symptoms as was the case here. The above data are insufficient to determine participants' understanding of diagnosticity, as they could have relied either on the absolute frequency of occurrence of the symptoms under RDS-H (H), or on the relative difference in frequency of occurrence under both competing hypotheses, RDS-H and TTN-H (H and $\hat{H}$). Either approach would affect the estimates in a similar manner because the H-and L-diagnostic symptoms differed along both dimensions. A H-diagnostic symptom was high in absolute frequency of occurrence under the hypothesis it supported as well as in the difference in frequency of occurrence under both hypotheses. Similarly, a L-diagnostic symptom was low in both absolute and relative frequency of occurrence. One way to test people's understanding of the concept of diagnosticity could be to present problems where

$$P(D|H) = 0.85 \text{ and } P(D|H_1) = 0.83$$

$$P(D|H) = 0.04 \text{ and } P(D|H_1) = 0.02$$

$$P(D|H) = 0.85 \text{ and } P(D|H_1) = 0.20$$

If the concept is not understood correctly and people rely only on the absolute frequency of D under Hypothesis H, the resulting P(H|D) should be approximately equal for (a) and (c) but lower for (b). If people rely on the difference in frequency of occurrence of D under both hypotheses H and $\hat{H}$, the resulting P(H|D) should be approximately equal for (a) and (b) but higher for (c). If P(H|D) is calculated in a normatively correct manner, taking both the absolute and the relative frequency of occurrence into account, then (c) should be highest, followed by (b) and (a). This will be tested in a future experiment. Finally, we will add a short training module showing how Bayes' Theorem works, and add more practice examples. This will be tested independently.

Acknowledgements

We thank Dr. Satid Thammasitboon for reviewing and adjusting the cases and for giving us access to participants. and all the student and pediatric resident participants who so willingly took part in the usefulness test.

References

1. Constantine, L.L. & Lockwood, L.A.D. (1999). Software for use: A practical guide to the models and methods of usage-centred design, Addison-Wesley, Reading, MA.
2. Mayhew, D. (2003). *The usability engineering lifecycle: a practitioner's handbook for user interface design,* Morgan Kauffman, San Francisco, CA.
3. Preece, J., Rogers, Y. & Sharp, H. (2007). *Interaction design: Beyond human-computer interaction*, John Wiley & Sons Ltd. Hoboken NJ, 2nd. edition
4. Ford, D.N. & Sterman, J.D. (1998). Expert knowledge to improve formal and mental models, *System Dynamics Review*, 14, 309-340.
5. Borra, R.C., Andrade, P.D. Corrêa, L. and Novelli, M.D. Development of an open case-based decision-support systemd for diagnosis in oral pathology, *Europran Journal of Dental Education 11*, (2007), 87-92.
6. Goggin, L.S., Eikelboom, R.H. and Atlas, M.D. Clinical decision support systems and computer-aided diagnosis in otology, *Otolaryngology 136*, (2007), 521-526.
7. Hughes, C.M., Phillips, J., and Woodcock, J. How Many Deaths Are Due to Medical Errors? *JAMA*, *284* (2000), 2187-2189.

8. Miller R.M., Elixhauser A., and Zhan, C. Patient safety events during pediatric hospitalizations. *Pediatrics. 111* (2003), 1358–66.
9. Berner, E.S, and Graber, M.L. Overconfidence as a cause of diagnostic error in medicine, *American journal of medicine*, *121(SA)*, (2008), 2-23.
10. Kohn, L.T., Corrigan, J., and Donaldson, M.S. *To err is human: building a better health system* National Academic Press, Washington, DC, 2000, 1-2.
11. Leape, L.L. Institute of Medicine Medical Error Figures Are Not Exaggerated. *JAMA 284* (2000)*,* 95-97.
12. Hayward, R.A., and Hofer, T.P. Estimating hospital deaths due to medical errors: Preventability is in the eye of the reviewer. *JAMA 286,* (2001), 415-420.
13. Schiff, G.D., Kim, S., Abrams, R., Cosby, K., Lambert, B., Elstein, A.S., Hasler, S., Krosnjar, N., Odwazny, R., Wisniewski, M.A., and McNutt, R.A. Diagnosing diagnosis errors: Lessons from a multi-institutional collaborative project, *Advances in patient safety 2*, (2005), 255-278.
14. Croskerry, P. Critical thinking and decision making: Avoiding the perils of thin-slicing. *Annals of Emergency Medicine 48*, *6* (2006)*,* 720-722.
15. Bavdekar, S.B. and Pawar, M. Evaluation of an internet-delivered pediatric diagnosis support system (ISABEL®) in a tertiary care center in India, *Indian pediatrics*, *42*, November (2005), 1086-1091.
16. Larissa, A., Lyman, J., and Borowitz, S. *Impact of a web-based diagnosis reminder system on errors of diagnosis.* Poster session presented at American Medical Informatics Association Annual Conference, (2006).
17. Maffei, F., Nazarian, E., Ramnarayan, P., Thomas, N., and Rubenstein, J. Use of a web-based tool to enhance medical student learning in the pediatric intensive care unit and inpatient wards. *In Proc. 15th. Ann. Ped. Criti. Care Coll., Interactions in Pediatric Critical Care, 6*, 1, (2004) September 30-October 2.
18. Peterson, C.R., and Beach, L.R. Man as an intuitive statistician, *Psychological bulletin*, *68*, 1 (1967), 29-46.
19. Edwards, W. Conservatism in human information processing, in D. Kahneman, P. Slovic & A. Tversky (eds), *Judgment under uncertainty: Heuristics and biases*, Cambridge university press, Cambridge, UK, 2002.
20. Kahneman, D., and Tversky, A. The simulation heuristic"; in D. Kahneman, P. Slovic & A. Tversky (eds) *Judgment under uncertainty: Heuristics and biases*; Cambridge University press, Boston, MA 2002.
21. Koehler, J.J. The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges, *Behavioral and brain sciences*, *19,* 1, (1996), 1-53.
22. Fischhoff, B., and Beyth-Marom, R. Hypothesis evaluation from a Bayesian analysis, *Psychological review*, *90*, 3 (1983), 239-260.
23. Beyth-Marom, R., and Fischhoff, B. Diagnosticity and pseudodiagnosticity, *Journal of personality and social psychology 45,* (1983), 1185-1195.
24. Sonnenberg, A. We only see what we already know – a modified Bayes' formula to explain inherent limitations of diagnostic tests, *Medical hypotheses 63*, (2004), 759-763.
25. Alvarez, S.M., Poelstra, B.S. and Burd, R.S. Evaluation of a Bayesian decision network for diagnosing pyloric stenosis, *Journal of pediatric surgery*, *41*, (2006), 155-161.
26. Graber, M.L. Diagnostic errors in medicine: What do doctors and umpires have in common?, *Morbidity & mortality 2*, (2007), 1-6.
27. Lindgaard, G. Pyper, C., Frize, M. & Walker, R. (2008). Does Bayes have it? Decision support systems in diagnostic medicine, International journal of industrial ergonomics, 39(3), 524-532.