Journal of Universal Computer Science, vol. 14, no. 16 (2008), 2720-2736
submitted: 9/7/07, accepted: 15/10/07, appeared: 28/8/08 © J.UCS

Intelligent Decision Support in Medicine: back to Bayes?

Gitte Lindgaard, Catherine Pyper, Monique Frize
(Carleton University, Ottawa, Canada
gitte_lindgaard@carleton.ca, {cpyper, mfrize}connect.carleton.ca)

Robin Walker
(IWK Health Centre, Halifax, Canada
Robin.Walker@iwk.nshealth.ca)

Craig Boutilier, Bowen Hui
(University of Toronto, Toronto, Canada
{chair, bowen}@cs.toronto.edu)

Sheila Narasimhan
(Carleton University, Ottawa, Canada
snarasim@connect.carleton.ca)

Janette Folkens
(Carleton University, Ottawa, Canada
lu-lu@rogers.com)

Bill Winogron, Peter Egan, Colin Jones
(S4Potential, Almonte, Canada
{bwinogron, pegan, cjones}@s4potential.com)

Abstract: Decision Support Systems are proliferating rapidly in many areas of human endeavour including clinical medicine and psychology. While these are typically based on rule-based systems, decision trees, or Artificial Neural Networks, this paper argues that Bayes' Theorem can be applied fruitfully to support expert decisions both in dynamically changing situations requiring the system progressively to adapt, and when this is not the case. One example of each of these two types is given. One provides diagnostic support for human decision makers; the other, an e-health mental intervention system provides decision rules enabling it to respond and provide the most appropriate training modules to input from clients with changing needs. The contributions of psychological research underlying both systems is summarized.

Keywords: Bayes' Theorem, Decision Support Systems (DSS), diagnostic error, individuating information, base rates, e-health intervention
Categories: H.1.2, H.4.2, I.2.1

1 Introduction

Decision Support Systems (DSSs) are proliferating in many areas of human endeavor involving complex problem solving including medical diagnosis. Expert system

algorithms range from rule-based systems [(Ramnarayan, Britto, 2002]; [Seidel et al, 2003]), via Artificial Neural Networks (ANNs) (Burnside, 2005), Bayesian decision theory (Tung, Quek, 2005), decision trees (Dudoit, Fridlyand & Speed, 2000) to support vector machines (Furey et al., 2000; Yeoh et al., 2002), and fuzzy logic (Tung, Quek, 2005). The fact that many of these systems attempt to imitate human decision processes (e.g. Borra et al., 2007; Goggin et al., 2007) underscores the often misguided assumption that human performance is inevitably optimal and hence worth imitating. This approach is puzzling to researchers studying human decision processes, as a whole family of serious shortcomings to human reasoning, choice, and decision making are well documented in that research domain (Kahneman et al., 1982). One major objective of research in human judgment and decision making is to understand the reasons underlying suboptimal decision processes in order to identify strategies for improving human performance. One of the two cases presented in this paper discusses problems specifically associated with medical diagnostic decision making. It introduces a DSS that is based on lessons learned from empirical research into certain judgmental biases that were found systematically to impact diagnostic decisions in a series of controlled laboratory experiments. That DSS is non-adaptive in the sense that it relies on a finite, static database. It aims to support a particular class of medical diagnostic decisions for resident pediatricians.

Clinical psychology is another important area in which DSSs could be usefully applied. The second case presented here concerns the design of an adaptive DSS that provides psychological e-health services to clients in a variety of clinical settings and skill-training programs. It relies on ongoing information about a client for selecting the most suitable learning modules for that particular client.

The contribution of psychology varies between the two systems. In the non-adaptive medical diagnostic DSS, it supports human decisions; in the adaptive system, the DSS provides the relevant variables and decision rules enabling the computer to select learning modules for clients whose needs change dynamically throughout the skill-training course. In both cases, the main emphasis is on the contribution of psychology to the background justification, design, and evaluation of DSSs.

As both the case studies apply Bayesian decision models, the main purpose of this paper is to show that Bayes' Theorem can offer a viable and flexible approach to the design of DSSs. One of the strengths of Bayesian models is precisely that they are adaptive in the sense that they are able to 'learn' iteratively from 'experience' and without having to change the core model. This enables such models to become customized to individual users' needs, which is very important in the context of clinical psychology. In addition, we argue that Bayesian models can be usefully applied to support human decision making in non-adaptive situations as well as in more traditional AI situations in which the DSS adapts to human users whose needs change over time. The two examples of DSSs are currently under development.

Bayes' Theorem and the notion of probability are discussed in the next section, in which the concept of diagnosticity is also introduced. It is followed by a discussion of the magnitude of problems in diagnostic medicine and then by a brief outline of the potential causes of flaws in the information integration stage of decision making. The experiments forming the basis for the diagnostic DSS are then summarized, and the creation of the diagnostic DSS is described. The second DSS, an e-health intervention

system, is then introduced, followed by an outline of the research underlying the design of the user model for the e-healthDSS. Finally, a set of conclusions are drawn.

2 Bayes' Theorem and the notion of probability

The notion of probability is relates to the degree of belief warranted by evidence (epistemic probabilities) on the one hand, and with the tendency, displayed by some chance devices, to produce stable relative frequencies (aleatory probabilities) on the other. Whereas the statistical probability concerns the way evidence from various sources is combined into a numeric statement irrespective of the judge's belief, the epistemic probability incorporates an assessment of the judge's personal belief as well, generated from autobiographical experience and state of knowledge about the evidence. The human-generated probability reflects both arithmetic calculations and degree of belief - it is an epistemic probability. By contrast, a computer-generated probability is an arithmetic computation of given numeric values – it is a statistical probability. Consequently, it is unrealistic to expect the two to be identical. Not surprisingly, computer-generated probabilistic judgments are most accurate: subjective beliefs are more likely to attenuate than to increase judgmental accuracy because beliefs are derived from a judge's understanding of her autobiographical experience.

In Bayes' Theorem, knowledge is represented via hypotheses, H_i, each of which is characterized by a subjective probability $p(H_i)$, representing one's confidence in its truth [De Finetti, 1976]. The output of a Bayesian analysis is a distribution of probabilities over a set of hypotheses. These probabilities can be used in combination with information about payoffs associated with various decision possibilities and 'states of the world' to implement any number of decision rules. 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 so-called the base rates, and incoming case-specific evidence is provided through individuating information. The outcome of a Bayesian analysis, the posterior probability, is calculated by combining the base rates and the individuating information. The model is iterative in the sense that a posterior probability resulting from one calculation becomes the prior probability in the next as more individuating information becomes available. Two hypotheses, H and $\hat{H}$, are assessed against one another, 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, which in turn 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 of opinion should occur. The resulting posterior probability is therefore identical to the base rate representing the hypothesis in terms of which the judgment is made.

In the case of the medical diagnostic DSS discussed here, the Bayesian algorithm calculates the probability associated with the possible diseases that could account for the clinical picture via the signs, symptoms, and laboratory findings entered by the diagnostician. It provides the five highest probabilities of the list of possible diseases.

In the case of the e-health intervention DSS, it relies on feedback by the client for deciding upon the best possible learning module.

2.1 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" ([Wells, Lindsay, 1980], p. 778). A component that should have no impact on the judgment is thus, by this definition, nondiagnostic. In order to determine the informativeness (diagnosticity) of the entire individuating information in cases where it consists of several items, a value must be assigned to each item. The combined value of items then determines the degree of importance and hence its support, for the hypothesis being entertained.

Early studies of subjective probability assessments in Bayesian decision tasks found consistently that judges adjusted their probabilities less than Bayes' Theorem demands. By failing to adjust their probabilities as much as the objectively should have, the judges were thus 'conservative' ([Phillips, Edwards, 1966]; [Edwards, 1982]). The debate on conservatism led to Peterson and Beach's (1967) optimistic claim that "man is an intuitive Bayesian". As a consequence of Peterson and Beach's thorough literature review, researchers' interest in conservatism waned quickly and quietly [Fischhoff, Beyth-Marom, 1983], especially with the rise of research into the so-called "base-rate fallacy" phenomenon [Bar-Hillel, 1980].

This line of research, inspired by Kahneman and Tversky (1972), provoked the bold claim that "man is apparently not a conservative Bayesian; he is not a Bayesian at all" (p. 450). One of the most famous cases tested by Kahneman and Tversky was the lawyer-engineer problem in which subjects were provided with several descriptions of people under high (70/100 lawyers in a particular sample of lawyers and engineers) or low (30/100 lawyers) base rate conditions. Each vignette presented as a coherent paragraph describing a set of personal characteristics of an individual and presented as the individuating information. The paragraph was intended to be nondiagnostic, that is, equally descriptive of a lawyer or an engineer. Thus, the posterior probabilities should have been identical with the prior probability of lawyers, either .70 or .30. Yet, the individuating information was consistently assessed in terms of the degree to which it appeared to represent the stereotype of a lawyer. Kahneman and Tversky (1982) called this the 'representativeness bias. Hundreds of laboratory studies conducted since then have çonfirmed the robustness of that bias but have failed to shed much light on the phenomenon other than to support the claim that "base rates are universally ignored" ([Koehler, 1996], p. 2).

Studies of nondiagnosticity of individuating information and its impact on base rate usage have primarily focused on outcome probability estimates rather than on the fate of individual items in the individuating information in the final judgment. The psychological research underlying the justification for the (Resident Diagnostic Decision System) ReDDS, the non-adaptive DSS described later, focused on the integration of items in the individuating information instead.

Diagnosticity of the individuating information is equally important in the e-health intervention DSS. It uses the client's input to update its 'opinion' and select learning modules of increasing complexity. Client input is provided in the form of answers to questions and small tests presented at the end of a training module, as well as on other

forms of client-computer interactions during the playing of the previous training module.

3 Diagnostic Decision Support in medicine: is there a problem?

According to certain patient safety research reports, between 98,000 [Hughes et al., 2000] and 115,000 [Miller et al., 2003] hospitalized people die every year in the United States due to some kind of medical error. There is some dispute about the accuracy of the figures ([McDonald et al., 2000]; [Leape, 2000]) and about the definition and calculation of "preventable error" [Hayward, Hofer, 2001]. One recent review of some 14 studies of general medical errors published between 1991 and 2004 [Schiff et al., 2006] found diagnosis-related errors to account for 10-30% of all errors recorded. Others estimate that these errors account for up to 76% of all medical errors [Amy et al., 2006]. Apparently, the 'gold standard' of misdiagnosis obtained from autopsies has consistently yielded a misdiagnosis rate of 40% over the past 65 years [Croskerry, 2006]. Unfortunately, autopsies are not performed routinely any more. Yet, with increasing pressure on medical personnel to attend to more patients in shorter time frames while also working extremely long hours, it is safe to predict that the problem of misdiagnosis is likely to increase.

Online resources providing both more general medical information, for example, PubMed (http://www.pubmedcentral.nih.gov/), eMedicine (www.emedicine.com) and SearchMedica (www.SearchMedica.co.uk), as well as specialized applications such as DermConsult (www.dermconsult.com.au) are proliferating; some of these are enjoying heavy usage by medical practitioners (Chamberlain, 2006), but it is as yet unclear just how much they benefit the physicians and their patients. One popular diagnostic web-based DSS, Isabel, provides information in the form of additional diagnoses which the practitioner may or may not have considered in the assessment of a given patient (e.g. [Ramnarayan, 2005]; [Ramnarayan et al., 2006]; [Bavdekar, Pawar, 2005]. Isabel draws on cross-references from a range of medical textbooks; it has been found to perform quite well in terms of "including [73%] all key diagnoses" [Ramnarayan, 2005] as well as including the "single expected" diagnosis in over 90% of cases in various validation studies [Ramnarayian, Cronje, 2005]. It can parse both keyboard entries and unformatted, spoken natural language; its output is a list of 10-15 possible clinical conditions [Bavdekar, Pawar, 2005] listed in 15 clinical categories, including Gastrointestinal Disorders, Nervous System Disorders, Shock States, Urologic Disorders, Infection Diseases, and others. When patients present with ambiguous symptoms that may point to several possible diseases, the output can be overwhelming and very time-consuming to process. Furthermore, since Isabel's data are derived from medical texts, it cannot provide probabilities associated with each of the possible conditions; it merely aims to "remind" the medical practitioner of alternative possibilities. Thus, while the expansion of the problem space is a desirable feature, Isabel also runs the risk of causing information overload. ReDDS takes a different approach; it provides probabilities based on the relative frequency of occurrence of signs and symptoms in different diseases contained in its database. ReDDS is extremely limited in scope; this is because it is intended as a 'proof of concept' in its present form.

3.1 Potential causes of flawed information integration

The objective of the research underlying the design of ReDDS was to trace the effect of specific items in the individuating information in an occupation-related context rather than relying on social stereotypes. Therefore, serial position effects, primacy and recency effects, were of specific interest. Much research into primacy effects has supported the attention decrement hypothesis, in which the weighting of later presented items decreases due to a progressive reduction in attention over a number of items [Anderson, 1982]. In recency effects this process is reversed: later items receive more attention than earlier presented items. When judges fall prey to either of these, the final judgment depends on the serial position of items of information deemed most important. When the individuating information is nondiagnostic and the serial position of the items contained in it is varied systematically, the effect of each item on the ensuing judgment can be accurately traced and distinguished from a so-called anchoring effect. In the presence of an anchoring effect, the item weighted most heavily and thus serving as an anchor, may be presented in any serial position in the stimulus array [Lopes, 1983]. Anchoring may result if diagnosticians fail adequately to adjust their opinion in the light of other items of information. If a certain item is deemed particularly important and is selected as an anchor, the resulting judgments should be very similar regardless of the serial position in which that item is presented.

A confirmation bias ([Wason, 1960]; [Wason, 1968]; [Klayman, 1995]) exists when subjects systematically display inappropriately high confidence in one hypothesis [McKenzie, 2004]. Confirmation biases can be inferred from a failure to change one's opinion in the face of nonsupporting or contradictory evidence, or selection of data favoring one's hypothesis while ignoring data that would contradict it [Klayman, Ha, 1987]. If diagnosticians systematically weight symptoms confirming a disease in terms of which a subjective probability is made in a Bayesian decision task when the overall case information is carefully balanced to be nondiagnostic, the estimate should consistently be higher than justified by the evidence.

It is difficult to distinguish between reliance on a confirmation bias or use of an anchor in judgments containing nondiagnostic individuating information. If the item selected as an anchor also confirms hypothesis H, the two strategies would be indistinguishable: in either case, the individuating information, $P(D \mid H)$, would affect the judgment more than warranted by the evidence. If however, the judgment is dominated by the evidence in support of the alternative hypothesis, $\hat{H}$, the denominator term, $P(D \mid \hat{H})$, would suggest anchoring.

Graber (2007) 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 Eddy and Clanton's (1982) suggestion that medical diagnosticians select a single, very salient symptom 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 possible diagnoses are not considered once a hypothesis has been identified [Graber, 2007]. This possibility was pursued in the psychological studies summarized very briefly below in which information integration strategies employed by nurses and physicians was studied.

3.2 Summary of studies informing the design of ReDDS

A series of five experiments were conducted to learn more about the information integration strategies of medical diagnosticians when faced with a minimum of individuating information which, on balance, is nondiagnostic. Experiment 1 explored the use of nondiagnostic individuating information in a set of different medical conditions presented to a sample of 80 expert clinical nurses. Each condition was causally related to two diseases, each with unique symptoms, and also sharing certain symptoms. Four symptoms were presented in vignettes each describing an individual patient. One symptom confirmed the hypothesis in terms of which the overall judgment was made, one disconfirmed it, one was neutral (it could occur in either of the two diseases), and one was irrelevant. The diseases were presented as base rates – for example Acute abdomen caused either by diffuse peritonitis (29/100 cases) or bowel obstruction (71/100 cases). The results yielded no serial order effects although weightings of each symptom were always highest for the confirmatory, and lowest for the disconfirmatory symptom, suggesting that the attempted nondiagnosticity of the individuating information as a whole was not recognized.

Experiment 2 aimed to quantify the perceived diagnosticity of each symptom in the vignettes and to select a subset of symptoms such that diagnosticity could be balanced in subsequent experiments still aiming to produce truly nondiagnostic cases. Lists of symptoms were prepared for each medical condition such that some confirmed the disease in terms of which the expected frequency rating was made, some disconfirmed it while simultaneously confirming the opposite disease, and some were neutral. Each list was presented twice - once for each disease belonging to a medical condition. Thus for the 'Acute abdomen' condition, judgments were made in terms of 'diffuse peritonitis' in one, and 'bowel obstruction' in the other list. Symptoms were arranged in different random orders for each disease, and always presented such that the same list was not seen twice in a row. Subjects were asked to assess the frequency with which they expected each symptom to be present in a sample of 100 patients, all diagnosed with the disease in question.

While the perceived frequency of occurrence varied greatly within and between diseases, suggesting that the four symptoms indicative of one of the diseases were not perceived to be equally diagnostic, the procedure enabled us to identify high- and low-diagnostic symptoms characteristic for each disease. The high-diagnostic symptom received the highest mean weighting under the disease for which it was characteristic and a very low rating for the alternative disease, whereas the mean rating for the low-diagnostic symptom differed slightly when presented in the two contexts.

Experiment 3 presented vignettes for one disease-pair to a sample of 44 expert nurses and each requiring a single probability estimate. Each vignette contained one high-or low-diagnostic confirmatory symptom, one high-or low-diagnostic disconfirmatory symptom, and one neutral symptom, presented in a completely factorial manner, with each symptom appearing in each of the three serial positions and in all possible combinations. Judgments were made in terms of the probability of angina pectoris, one of the two diseases, and base rates varied between the two groups of subjects (28/100 or 72/100 patients with angina pectoris respectively). The results showed a clear primacy effect: a high-diagnostic confirmatory symptom presented first in a vignette resulted in significantly higher probability estimates compared with

later presentation of the same symptom. The high-diagnostic disconfirmatory symptom resulted in lower probability estimates when presented first than when presented later in the sequence. As predicted, base rates did not affect the probability ratings at all.

Experiment 4 replicated Experiment 3 using another disease-pair and a new sample of 44 expert nurses. The results resembled those obtained in Experiment 3 very closely, again showing a very clear primacy effect. Assuming that nurses generally do not propose medical diagnoses in their clinical work, the vignettes used in Experiment 3 were again tested in Experiment 5 on a sample of 32 medical practitioners. Again, the results yielded a highly significant primacy effect consistent with the earlier results and thus supporting the robustness of the primacy effect in these kinds of clinical, occupationally relevant, cases.

Nondiagnosticity of conflicting information which, when summed, equally supports both competing hypotheses can only be detected if both of these are considered. The persistent presence of a primacy effect in three experiments suggests that the symptoms were not weighted according to both hypotheses: subjects apparently failed to detect the nondiagnosticity of the cases. The primacy effect also excludes the possibility that subjects selected an anchor in any of the three experiments. Had they done so, all the probability should have been virtually identical because the estimates were based on the same symptoms. The results did not support a tendency to favor either hypothesis; the presence of a confirmation bias cannot therefore be entirely refuted. However, all results suggest that items were processed in the order in which they were seen: the mean estimates were consistently closer together in the third than in the first serial position due to a progressive increase in disconfirmatory estimates and a progressive decrease in confirmatory estimates. This is consistent with the attention decrement hypothesis.

The extent to which subjects' understanding of the diagnosticity concept may have been correct is not entirely clear from the above results. They could have relied either on the absolute frequency of occurrence of the symptoms in the to-be-evaluated disease, or on the relative difference in frequencies of occurrence of symptoms under H and $\hat{H}$. Either approach could affect the estimates in a similar manner because high- and low-diagnostic symptoms differed along both dimensions: a high-diagnostic symptom was high in absolute frequency of occurrence under hypothesis H as well as in the difference in frequency of occurrence under both hypotheses H and $\hat{H}$.

To the extent that the above experimental findings may be indicative of the way diagnosticians process information in practice, a DSS based on such research should support diagnostic decision making. However, such a DSS would only be worthwhile if it can be shown they address an existing problem. That is explored next.

3.3 Creating ReDDS (Resident Decision Support System)

The above experiments suggest that the task of diagnosing may proceed in the manner described by Eddy and Clanton (1982). They suggest that the diagnostician generates a hypothesis from a single salient, easy-to-observe, symptom, and weighs additional information according to its support for that hypothesis. A DSS may thus improve diagnostic decision making simply by providing alternative possible hypotheses, which would encourage the diagnostician to widen the range of hypotheses to be pursued. That is precisely what Isabel does, but it does not provide probability

associated with each potential output condition. Therefore, differentiation between the proposed hypothetical conditions is not really facilitated, and nor does it promote a better general understanding of the concept of diagnosticity. In order to achieve this, the diagnosticity of each sign, symptom, and laboratory finding must be quantified in relation to each possible outcome.

Ideally, a DSS should be able to calculate posterior probabilities, $P(H|D)$. However, because the base rates, $P(H)$ and $P(\hat{H})$, will vary depending on the reference groups chosen for comparison, this is not possible. Say, for example, a male infant is admitted with symptoms X, Y, and Z. Should the reference groups be male ($P(H)$) versus female ($P(\hat{H})$) infants, all infants displaying symptoms X, Y, and Z versus all infants ($P(H)$) displaying symptoms X, A, and Y ($P(\hat{H})$), or yet other groups? Each calculation would clearly result in different posterior probabilities, thereby confusing rather than assisting the diagnostician.

ReDDS is based on a subset of data from a database of 1200 infants admitted to the neonatal intensive care unit at the Children's Hospital of Eastern Ontario. It aims both to facilitate the diagnosis of respiratory distress in infants and to teach residents the concept of diagnosticity. Respiratory distress was selected as the target condition in this because it occurs relatively frequently and because the signs and symptoms are ambiguous, generally pointing to different possible causes.

Upon the advice by our expert physician partner, an experienced medical records librarian worked systematically through the patient database to identify infants diagnosed with respiratory distress. Some 97 cases fulfilled the selection criteria. As many of the records in the database were incomplete, the librarian consulted the original patient records to yield a complete dataset. This dataset enabled accurate quantification of the relative diagnosticity of each sign and symptom associated with every causal condition which, in turn, enabled the calculation of $P(D|H)$ and $P(D|\hat{H})$ for any combination of symptoms. The ReDDS GUI interface rank orders and presents the five most probable conditions based on the signs and symptoms entered by the physician. By contrast, Isabel's suggested conditions appear implicitly to be equiprobable because it lacks quantified information about the diagnosticity of each datum in the database relative to each condition.

Basically, the diagnostician enters a set of signs, symptoms, and laboratory values into ReDDS and presses a button labeled 'Diagnose' when she is ready. In response, ReDDS provides the list of five diseases and their associated probabilities. More symptoms can be entered to refine the diagnosis, and the 'Diagnose' button may be pressed as many times as the diagnostician likes, even after entering only a single symptom.

3.4 Validation of ReDDS

The most important purpose of ReDDS is to demonstrate that diagnostic performance can improve by encouraging the diagnostician to widen the range of hypotheses to be pursued when diagnosing ambiguous cases. This is consistent with the aim of Isabel, except that it may provide a much longer list of possible conditions and no probabilities. We also aim to demonstrate that appreciation of quantified diagnosticity is generalizable to diseases not covered by the DSS. Let us assume that Graber's (2007) assertion is correct, that defective synthesis of available data rather than knowledge deficits are the underlying cause of cognitive errors in diagnosis. The

sheer realization that one is prone to fall victim to primacy effects when generating and testing diagnostic hypotheses should help physicians to consider possible alternative diseases and also the likelihood associated with each. On the one hand, the initial search should thus be widened to include alternatives not otherwise considered, and on the other hand, it should also be narrowed only to consider the most likely alternatives.

ReDDS is currently undergoing validation via a controlled laboratory experiment. Two groups of subjects comprising senior pediatric residents and expert neonatologists are first shown a set of very brief vignettes designed to be nondiagnostic; that is, the individuating information has been carefully balances so as to point equally to the two hypotheses being tested. The participants' task is to estimate the probability that the infant described in the vignette suffers from one of the diseases. This design resembles that described in experiments 3-5 reported earlier. In the second phase, one half of the participants are trained on ReDDS, and the other half is given an article on medical diagnosis to read. Both groups are then asked to diagnose another set of vignettes describing infants suffering from a range of diseases. In the third phase, participants are again given a set of nondiagnostic vignettes as in the first phase. We predict that participants in the group trained on ReDDS will utilize the base rates in the third, but not in the first, phase, and that the untrained group will not. In addition, and consistent with the earlier results, we predict a primacy effect for all participants in the first phase, but only for the untrained group in the third phase. We are reasonably confident that this validation study will lead to a better understanding of diagnosticity as well as to better diagnostic performance.

4 Psychologist in the box: an e-health mental intervention system

Clinical psychologists spend a large amount of time teaching their clients simple, routine social and interpersonal skills. The lessons needed to achieve this are repetitive, and the time devoted to these reduces the amount of face-to-face time that the clinician could otherwise spend dealing with more challenging issues. The motivation behind the project is therefore to eliminate the routine component from the precious one-on-one consultation time and provide the lessons on a computer system instead. The Bayesian algorithm must be designed such that it can be applied to a relatively wide variety of training courses. Currently, these range from courses in anger management, substance abuse, and an information course for parents dealing with their autistic children's severe behavioural problems. Consequently, the target audiences also vary widely. One of the main design challenges is that at least some of these target audiences have very limited education, a short attention span, and most likely also deficiencies in short-term, working memory. For example, the anger management course is currently being used throughout Canada's correction system; incarcerated young offenders are required to complete the course. These clients are generally not highly motivated, and certainly not intrinsically motivated, to complete the course, let alone to learn the lessons in it. Others, such as parents of autistic children, can only devote a very limited period of time to each session, or they need a quick reminder for ways to deal with an acute situation. Likewise, people, for example, who are hospitalized with substance abuse, may not recognize the need for behaviour modification at the time they are encouraged to take the course.

In order to keep the client's attention focused on the course content and the lessons it seeks to convey, one important requirement for the training programs is to ensure that clients can readily identify with that content and that it is entertaining. All the courses comprise short scenarios, film clips, in which professional actors demonstrate correct and incorrect behaviour. A background character, or a brief text message shown on the screen, may ask for the client's opinion on the behaviour of the actors. The correctness of the client's answers, the outcome of short quizzes, and client preferences together determine if and when the next module may be attempted. Alternatively, the client may choose to repeat the same module, or to see another one containing the same message but shown in a slightly different way. Clearly, for a course module to appeal equally to a, say, 17-year old incarcerated youth as well as to, say, a 50-year old woman, both of whom need anger management training, it is necessary to create many variations on the theme. The central message is the same, but the social context, the actors, the language they use, the way they move, dress, and so forth, vary from one version to the next. The computer continually and actively engages with the client, responding to the client's input, and adjusting as far as possible to the client's preferences while also providing instructive feedback in a manner that is acceptable to clients and that will keep them interested and engaged.
Perhaps the biggest challenge for the Bayesian algorithm is to determine 'who' the client is, so as to start off the 'relationship' on an agreeable basis; one wrong choice at the very beginning is likely to lose the client's interest right away. Of course, it is also very important to continue to keep the client's attention once the course is going, but it is easier to correct for a small mistake if a 'relationship' has already been established with the client.

In order to start the process effectively with a new client who may resent having to do the course, we need to understand what and how much the DSS needs to 'know' about the client and also about the client's learning style preferences. Our background in psychology has taught us that personality is stable. A literature review suggested that knowledge of certain generic personality traits in a population may help to infer an individual's preferred learning style and thus assist the DSS in the wise selection of learning modules. One major challenge was therefore to design a generic user model in order to learn how to equip the DSS with the necessary initial knowledge about a new client. This issue is addressed next.

4.1 Designing a user model based on personality traits and learning style

Personality traits influence many aspects of individual behaviour such as attitude and motivation, including attitude towards learning [Komarraju, Karau, 2005]. It would therefore seem advantageous for training systems to adapt to a learner's personality. One goal was therefore to develop a model that can recognize a client's personality. Since some of our target populations can be assumed to approach the e-health intervention DSS with a poor attitude and low motivation, requesting them to fill in a complete personality assessment instrument before starting the course per se was not viable. A second goal was therefore to assess if the necessary personality knowledge could be gleaned reliably via a short set of questions presented electronically and without violating the copyrights of existing instruments. The third goal was to assess the degree to which the claim in the literature that knowledge of certain personality traits can be mapped reliably onto preferred learning style.

There are many theories of personality and many ways to assess it (see e.g. [Larsen, Buss, 2002] for a recent review), but there is general agreement that personality is biologically based [Eysenck, 1981]. The literature also suggests that the most relevant personality trait to an e-learning environment is the extroversion-introversion continuum [Gill, Oberlander, 2002; Mairesse, Walker, 2006]. Introverts appear to have a higher level of arousal in the autonomous nervous system than extroverts. This difference can account for different behaviours and preferences among introverts and extroverts: extroverts are said to be under-aroused - introverts are over-aroused. As people work best at a moderate level of cortical arousal, extroverts will tend to look for external stimulation to reach an optimal arousal level; introverts will try to avoid highly arousing situations [Dewaele, Furham, 1999]. Eysenck and Eysenck [1964] notes that "the typical extrovert is sociable, likes parties, has many friends, needs to have people to talk to, and does not like reading or studying by himself. He craves excitement, takes chances, often sticks his neck out, acts on the spur of the moment, and is generally an impulsive individual" (p. 8). Eysenck goes on to argue that an extrovert prefers to keep moving and doing things, tends to be aggressive and lose their temper quickly. By contrast, the typical introvert is a quiet retiring person, introspective, fond of books rather than people; he is reserved and distant except to intimate friends. He tends to plan ahead, 'looks before he leaps' and distrusts the impulse of the moment. Introverts do not like excitement. Extroverts also talk more, make fewer pauses and hesitations; they have higher speech rates, shorter silences and higher verbal output, but not as broad a vocabulary as introverts. For the purpose of designing the user models to be included in the DSS, it is important to note that extroverts like action, movement, fast pace, and working with others. By contrast, introverts prefer working alone, quiet solitude, and a slower pace that gives them more time to think about and digest the learning material. However, it is also worth noting that this same knowledge can and should be applied when designing interactive, adaptive e-learning modules. For various reasons that space limitations prevent us from reporting here, we selected the Eysenck Personlaity Questionnaire Revised (EPQ-R) [Eysenck, Eysenck, 1975] personality instrument for our experiment. In addition, we designed and tested a short version comprising some eight questions in an attempt to capture the elements of extroversion-introversion essential for the DSS to make a reasonable first decision about a new client's degree of extroversion and use this for selecting the most suitable learning module.

As is true for personality, there are many models of learning style and numerous instruments for assessing people's preferred learning styles ([Kolb, Kolb, 2005] [Honey, Mumford, 1992] [Herrmann, 1989]). The models of Kolb, Honey and Mumford, and Herman all acknowledge the possibility that learners may have more than one preferred learning style, depending on the learning context and the type of material to be learned. This is not necessarily a disadvantage for adaptive learning environments, as it gives more options to present learning material that will suit a particular learner. Again, for various reasons we selected the Honey and Mumford [1992] scale in attempt to correlate learning style with the introversion-extroversion continuum of personality in our experiment. Briefly, the scale distinguishes between activist, reflectors, theorists, and pragmatists. Activists usually embrace novel concepts enthusiastically but tend to lose patience quickly. They learn best from competitive activities and respond well to challenges. Reflectors are cautious people

who consider their actions carefully before making a decision. They tend to learn best when given time to prepare in advance. Theorists consider all alternatives and draw conclusions from their experiences. They attempt to fit their observations into a logical model or theory and learn best when required to understand complex problems. Pragmatists get impatient with too much reflection; they like to experiment with new plans, acting immediately without too much discussion. They learn best when the link between the subject matter and the desired outcome is apparent or there are obvious advantages to learning a given task.

4.1.1 Assessing the relationship between extroversion and learning style

The EPQ-R scale, the abbreviated version, and the Honey and Mumford scales were given to a random sample of 40 individuals ranging in age from 18 to 50 years and recruited from personal contacts. Gender was balanced, and the administration of the questionnaires was counter-balanced to avoid serial position effects. Before filling in the questionnaire, participants were asked to describe an experienced situation in which they felt they had been successful, and one in which they felt they had been unsuccessful. The sequence of these was also counter-balanced between participants. This was done to assess the degree to which participants' range, type, and volume of words would map to their degree of extroversion. One half of the participants described their episodes verbally to a researcher who recorded the session; the other half typed theirs into the computer as we assumed that people would speak more than they would write. This was borne out by the results. The balance of extroverts ($n = 34$) to introverts ($n = 6$) was highly skewed in the original sample, an additional six introverts were recruited to yield a balanced sample of 24 participants for the data analysis. Consistent with the predictions in the literature, the results revealed a significant positive correlation between extroversion and the activist learning style ($r = .563$, $p < .01$), a negative correlation with the reflector learning style ($r = -.323$, $p < .05$) and no correlation with the pragmatist or the theorist learning style. Analysis of the word count showed that extroverts used more words overall ($t(22) = 1.04$, $p < .01$) and more optimistic words ($t(22) = 1.48$, $p < .01$) than introverts. Word types were determined by the scheme provided by Pennebaker and King [1999] which is based on over 1200 essays and profiles from letters, essays, and other text samples. It thus appears that there is some link between the personality extroversion-introversion personality trait and learning style that can be usefully exploited in the design of our DSS. In terms of the short version of the personality questionnaire, the analysis showed that the distribution of extroverts and introverts closely resembled that obtained from the EPQ-R instrument. Thus, for the present purposes, it was concluded that the essence of extroversion can be captured in very few questions that will be asked at the very outset of the client-computer interaction.

Now, in order to select appropriate scenarios, the expert system applies this 'knowledge' of the distribution of the distribution of extraversion, (P(H), and introversion, ($P(\hat{H})$, in the client population. The client's responses to the eight personality-related questions serve to determine that client's degree of extroversion. The higher the degree of extroversion, the more likely it is that the client's preferred learning style 'activist' rather than 'reflector', and conversely, the lower the extroversion score, the higher the likelihood that the client tends towards the 'reflector' end learning style. Additional and ongoing informal conversation, together

with feedback on scenarios already shown, scores on quizzes, and answers to questions about the scenarios, the DSS continues to assess and adjust its 'opinion' of the client's degree of extroversion to guide further learning module selection. Through this continual evaluation of the client's feedback, the expert system progressively adjusts its knowledge about the client simply by calculating the posterior probability, $P(H \mid D)$ of extroversion. It uses each such outcome as the base rate, P(H), for the next calculation, based on further client input, comprising $P(P(D \mid H)$ and $P(D \mid \hat{H})$. Since $P(H) + P(\hat{H}) = 1.0$, adjustments to P(H) automatically updates the value of $P(\hat{H})$ as well. This way, the system 'learns' to fine-tune its selections to the needs of the client. Thus, the customized user model evolves during continued client-computer interaction.

5 Conclusion

In this paper we have attempted to show that Bayes' Theorem can be applied successfully to DSSs, even in circumstances requiring no adaptation once the basic database and decision rules are in place such as in ReDDS. We also attempted to show that Bayes' Theorem can be successfully applied to machine-based learning in a dynamically changing environment. The psychological contribution to ReDDS is an understanding to information integration, whereas to the 'psychologist in the box' project, it provides the initial knowledge of the distribution of base rates of the important variables comprising the individuating information, $P(D \mid H)$ and $P(D \mid \hat{H})$, as well as providing the selection rules based on evaluation of the different kinds of human input. We thus maintain that Bayesian models can be applied usefully to different kinds of decision problems, and that psychological research can successfully contribute to the background justification, the definition, design, and evaluation of DSSs in the medical and psychological arenas, regardless of whether the system is intended to support human- or machine learning.

Acknowledgements

This work was partially supported by a Canadian National Science & Engineering Research Council (NSERC)/Cognos Industry Research Chair grant, IRCSA 234087-05, and also by OCE/Precarn.

References

[Amy et al., 2006] Amy, L.R., Lyman, J., & Borowitz, S.: "Impact of a web-diagnosis reminder system on errors of diagnosis"; Proceedings American Medical Information Association (AMIA) annual conference, 11-15 (2006) November, Bethesda, MD, 2-6.

[Anderson, 1982] Anderson, N.H.: „Methods of information integration theory"; Academic Press / New York, NY (1982).

[Bar-Hillel, 1980] Bar-Hillel, M. I.: „On the subjective probability of compound events"; Organizational behavior and human performance, 9, (1980), 396-406.

[Bavdekar, Pawar, 2005] Bavdekar, S.B., 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.

[Borra, Andrade, Borra, Corrêa, Novelli, 2007] Borra, R.C., Andrade, P.D., Corrêa, L., Novelli, M.D.: “Development of an open case-based decision-support system for diagnosis in oral pathology”; European Journal of Dental Education, (2007), 11, 87-92.

[Burnside, 2005] Burnside, E.S.: “Bayesian networks: Computer-assisted diagnosis support in radiology”; Academic Radiology, (2005), 12, 422-430.

[Chamberlain, 2006] Chamberlain, A.J.: „The internet as a clinical aid for the dermatologist“; British Medical Journal (BMJ), 333, (2006), 1143-1145.

[Crosskerry, 2006] Crosskerry, P.: „Diagnostic failure: A cognitive and affective approach“; Advances in patient safety, 2, (2006), 241-254.

[Dewaele, Furnham, 1999] Dewaele, J., Furnham, A.: “Extraversion: The unloved variable in applied linguistic research”; Language Learning, (1999), 49(3), 509-544.

[De Finetti, 1976] De Finetti, B.: „Probability: Beware of falsifications!“ Scientia, 3, (1976), 283-303.

[Dudoit, Fridyand, Spleed, 2000] Dudoit, S., Fridyand, J & Spleed, T.P.: “Comparison of discrimination methods for the classification of tumors using gene expression data”; Technical report 576, Department of statistics, University of California, Berkeley, (2000).

[Eddy, Clanton, 1982] Eddy, D.M., Clanton, C.H.: „The art of diagnosis: solving the clinicopathological exercise“; New England journal of medicine, 306, (1982), 1263-1268.

[Edwards, 1982] Edwards, W.: „Conservatism in human information processing“; in B. Kleinmuntz (Ed). Formal representation of human judgment, John Wiley & sons, New York, NY (1982)

[Ennett, 2003] Ennett, C.: „Imputation of missing values by integrating artificial neural networks and case-based reasoning“; Unpublished PhD Thesis, Carleton University (2003)

[Eysenck, Eysenck, 1964] Eysenck, H.J., Eysenck, S.B.G.: “Manual of the Eysenck Personality Inventory”; Hodder & Stroughton, London (1964).

[Eysenck, Eysenck, 1975] Eysenck, H.J., Eysenck, S.B.G.: “Manual of the Eysenck Personality Questionnaire. London: Hodder & Stoughton (1975).

[Eysenck, 1981] Eysenck, M.W.: “Learning, memory and personality”, in H.J. Eysenck (Ed), A model for personality, Springer Verlag, Berlin, (1981), 169-209.

[Fischhoff, Beyth-Marom, 1983] Fischhoff, B. & Beyth-Marom, R.: „Hypothesis evaluation from a Bayesian analysis“; Psychological review, 90, 3, (1983), 239-260.

[Furey, Cristianini, Duffy, Bedmarski, Schummer, Hassler, 2000] Furey, T.S., Cristianini, N., Duffy, N., Bedmarski, D.W., Schummer, M., Hassler, D.: “Support vector machine classification and verification of cancer tissue samples using microarray expression data”; Bioinformatics, (2000), 16 (10), 906-914.

[Goggin, Eikelboom, Atlas, 2007] Goggin, L.S., Eikelboom, R.H., Atlas, M.D.: “Clinical decision support systems and computer-aided diagnosis in otology”, Otolaringo head neck surgery, (2007), 136(4), 521-526.

[Gill, Oberlander, 2002] Gill, A.J., Oberlander, J.: „Taking care of the linguistic features of extraversion“; Proceedings 24th. Annual conference cognitive science society, (2002), 363-368.

[Graber, 2007] Graber, M.L.: „Diagnostic errors in medicine: What do doctors and umpires have in common?"; Morbidity & mortality,(2007), 1-6, (2007).

[Hayward, Hofer, 2001] Hayward, R.A., Hofer, T.P.: „Estimating hospital deaths due to medical errors: Preventability is in the eye of the reviewer"; Journal of the American Medical Association, 286, (2001), 415-420.

[Herrmann, 1989] Herrmann, N: „The creative brain"; Brain Books, The Ned Hermann Group, North Carolina, (1989).

[Honey Mumford, 1992] Honey, P., Mumford, A.: "The Manual of Learning Styles", Ardingly House, Berkshire, (1992).

[Hughes et al., 2000] Hughes, C.M., Phillips, J., Woodcock, J.: "How Many Deaths Are Due to Medical Errors?"; Journal of the American Medical Association, 284, (2000), 2187-2189.

[Kahneman, Tversky, 1982] Kahneman, D. & Tversky, A.: „The simulation heuristic"; in D. Kahneman, P. Slovic & A. Tversky (eds)," Judgment under uncertainty: Heuristics and biases"; Cambridge university press, Boston, MA (1982)

[Kahneman, Tversky, 1972] Kahneman, D. & Tversky, A.: „Subjective probability: A judgment of representativeness"; Cognitive psychology, 3, (1972), 430-454.

[Klayman, 1995] Klayman, J.: Varieties of confirmation bias"; Psychology of learning and motivation, (1995), 32, 385-418.

[Klayman, Ha, 1987] Klayman, J. & Ha, Y-W.:„Confirmation, disconfirmation, and information in hypothesis testing, Psychological Review, (1987) 94 (2), 211-228.

[Koehler, 1996] Koehler, J.J.: „The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges"; Behavioral and brain sciences, 19, 1, (1996), 1-53.

[Kolb, Kolb, 2005] Kolb, A. Y., Kolb, D. A.: "The Kolb Learning Style Inventory- Version 3.1 2005 Technical Specifications"; Experience Based Learning Systems, Inc., Case Western Reserve University (2005).

[Komarraju, Karau, 2005] Komarraju, M., Karau, S.J.: „The relationship between the big five personality traits and academic motivation"; Personality and individual differences, 39 (2005), 557-567.

[Larsen, Buss, 2002] Larsen, R., Buss, D.: "Personality psychology: Domains of knowledge about human nature"; McGraw Hill, New York, (2002).

[Leape, 2000] Leape, L.L.: "Institute of Medicine Medical Error Figures Are Not Exaggerated"; Journal of the American Medical Association,(2000), 284, 95-97.

[Lopes, 1983] Lopes, L.L.: „Toward a procedural theory of judgment"; Psychological documents, 13, 2, (1983), MS no. 2590.

[Mairesse, Walker, 2006] Mairesse, F., Walker, M.A.: "Words mark the nerds: Computational models of personality recognition through language", Proceedings 28th. Annual conference cognitive science society, (2006), 543-548.

[McDonald et al., 2000] McDonald, C.J., Weiner, M., Hui, S.L.: „Deaths Due to Medical Errors Are Exaggerated in Institute of Medicine Report"; Journal of the American Medical Association 284, (2000), 93-95.

[McKenzie, 2004] McKenzie, C.R. M.: "Hypothesis testing and evaluation", in D.J. Koehler, N.Harvey (eds); Blackwell handbook of judgment and decision making, Blackwell, Oxford, (2004).

[Miller et al., 2003] Miller R.M., Elixhauser A., Zhan C.: ”Patient safety events during pediatric hospitalizations“; Pediatrics, 111, (2003) 1358–1366.

[Pennebaker, King, 1999] Pennebaker, J.W., King, L.A.: „Linguistic styles: Language use as an individual difference“; Journal of personality and social psychology, (1999), 77(6), 1296-1312.

[Peterson, Beach, 1967] Peterson, C.R., Beach, L.R.: „Man as an intuitive statistician“; Psychological bulletin, 68, 1, (1967), 29-46.

[Phillips, Edwards, 1966] Phillips, L.D., Edwards, W.: „Conservatism in a simple probability inference task“; Journal of experimental psychology, 72, 3, (1966), 346-354.

[Ramnarayan, 2005] Ramnarayan, P.: „Decision support systems: A how-to guide to their evaluation“, Sage Publications electronic resources, http://c.sagepub.com (2005)

[Ramnarayan, Britto, 2002] Ramnarayan, P., Britto, J.: “Paediatric clinical decision support systems”; Archives for diseases in children, (2002), 87, 361-362.

[Ramnarayan, Cronje, 2005] Ramnarayan, P., Cronje, N.: „Report on Isabel adult medical validation study“; Isabel Healthcare report, London, UK (2005) [Schiff et al., 2006] Schiff, G.D., Kim, S., Abrams, R., Cosby, K., Lambert, B., Elstein, A.S., Hasler, S., Krosnjar, N., Odwazny, R., Wisniewski, M.A., & McNutt, R.A. Advances in patient safety, 2, (2006), 255-278.

[Seidel, Breslin, Christley, Gettinby, Reid, Revie, 2003] Seidel, M., Breslin, C., Christley, R.M., Gettinby, G., Reid, S.W.J., Revie, C.W.: Agricultural systems, (2003), 76, 527-538.

[Tung, Quek, 2005] Tung, W.L., Quek, C.: “GenSo-FDSS: A neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data”; Artificial intelligence in medicine, (2005), 33, 61-88.

[Wason, 1968] Wason, P.C.: „Reasoning about a rule“; Quarterly journal of experimental psychology, 12, (1968), 129-140.

[Wason, 1960] Wason, P.C.: „On the failure to eliminate hypotheses in a conceptual task“; Quarterly journal of experimental psychology, 20, (1960) 273-281.

[Wells, Lindsay, 1980] Wells, G.L., Lindsay, R.C. L.: „On estimating the diagnosticity of eyewitness nonidentification“; Psychological bulletin, 88, 3, (1980), 776-784.

[Yeoh, Ross, Shutleff, Williams, Patel, Malifonz, Behm, Raiwady, Relling, Relling, Patel, 2002] Yeoh, E.J., Ross, M.E., Shutleff, S.A., Williams, W., Patel, D., Malifonz, R., Behm, F., Raiwady, S., Relling, M., Patel, A.: “Classification subtype discovery and prediction of outcome in pediatric acute lymphoblastic leukaemia by gene expression profiling”; Cancer cell, (2002), 1 (2), 133-144.