The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems

CONFERENCE PAPER · OCTOBER 2015

READS
59

3 AUTHORS, INCLUDING:

Adrian Bussone
City University London
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Dympna O'Sullivan
City University London
42 PUBLICATIONS **119** CITATIONS
SEE PROFILE

The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems

Adrian Bussone, Simone Stumpf, Dympna O'Sullivan
Centre for Human Computer Interaction Design
School of Mathematics, Computer Science and Engineering
City University London
London, United Kingdom
Adrian.Bussone.1@city.ac.uk, Simone.Stumpf.1@city.ac.uk, Dympna.O'Sullivan.1@city.ac.uk

***Abstract*— Clinical decision support systems (CDSS) are increasingly used by healthcare professionals for evidence-based diagnosis and treatment support. However, research has suggested that users often over-rely on system suggestions – even if the suggestions are wrong. Providing explanations could potentially mitigate misplaced trust in the system and over-reliance. In this paper, we explore how explanations are related to user trust and reliance, as well as what information users would find helpful to better understand the reliability of a system's decision-making. We investigated these questions through an exploratory user study in which healthcare professionals were observed using a CDSS prototype to diagnose hypothetic cases using fictional patients suffering from a balance-related disorder. Our results show that the amount of system confidence had only a slight effect on trust and reliance. More importantly, giving a fuller explanation of the facts used in making a diagnosis had a positive effect on trust but also led to over-reliance issues, whereas less detailed explanations made participants question the system's reliability and led to self-reliance problems. To help them in their assessment of the reliability of the system's decisions, study participants wanted better explanations to help them interpret the system's confidence, to verify that the disorder fit the suggestion, to better understand the reasoning chain of the decision model, and to make differential diagnoses. Our work is a first step toward improved CDSS design that better supports clinicians in making correct diagnoses.**

Keywords*— *CDSS; Trust; Reliance; Explanations; Reliability; User Study

I. Introduction

Clinical decision support systems (CDSS) can help support choices faced by clinicians through applying stored health knowledge to observations. CDSS are already used for a variety of purposes to improve health care: screening and prevention [22], medication decision-making [7], therapeutic planning and diagnostics [14], etc. However, the use of CDSS is not without problems. They are not perfectly reliable because they operate under conditions of uncertainty, and thus the correctness of their outputs may be affected by the quality of the decision-making of the system or the data and inputs it is given [14]. Therefore it is critical that clinicians using these systems do not trust them blindly.

Previous research has shown that users are not always sensitive to the reliability of automated systems, and often trust the system more than themselves [18,21]. This can lead to misuse of the system [24] through over-reliance – also known as 'automation bias' [3] – that comes from placing too much trust in the system and results in the user's agreement with *incorrect* system suggestions. Similarly, there might be instances of disuse [24], in which users do not follow *correct* suggestions, i.e. issues of self-reliance.

It has been suggested that the intelligibility of system behavior is an important factor in ensuring that the user understands how the CDSS operates [23]. This in turn could help clinicians identify if the system has erred and also ensure that the clinician forms a more accurate picture of the system's reliability. However, current CDSS designs rarely address how to make the system's functioning intelligible to clinical users.

The role of explaining the reasoning of intelligent systems has been investigated [10,18]. Some work has been carried out on exposing the reasoning through various *explanation types* [20], such as *why* it made the suggestion and/or the system's *confidence* that the suggestion is correct. Previous work has shown that providing explanations can increase users' understanding of how the system operates [16] and that it is better to expose as much information as possible about what the system uses to make its decisions to improve intelligibility

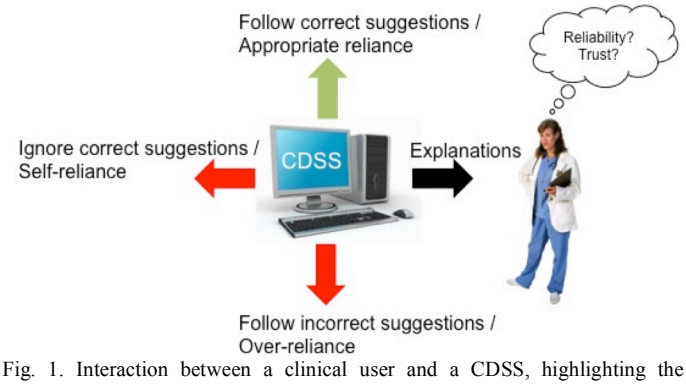

Fig. 1. Interaction between a clinical user and a CDSS, highlighting the relationship between explanations, reliance, and trust. Explanations from the system could help clinicians with assessing reliability and prevent over-reliance or self-reliance.

[17]. However, there is still a lack of knowledge about the impact of explanations on building trust and countering over-reliance, as some research suggests that explanations lead to more correct decisions [9] whilst others suggest they can also lead to worse decision-making by the clinician [4].

We explored these issues as part of the EMBalance project (http://www.embalance.eu/) in a CDSS prototype that supports primary care physicians to diagnose and treat balance disorders. While balance problems affect 40% of adults over the age of 40 in the U.K., and are the most common reason for clinical visits in individuals over 60, best practices on how to diagnose and appropriately treat balance disorders are still not widespread within primary care and usually require over four consultations and possibly referrals to specialists before a diagnosis is made [27]. Thus, a CDSS could substantially increase the correct diagnosis of conditions that cause vertigo, tinnitus, and falls (e.g. Meniere's Disease, benign paroxysmal positional vertigo, etc.) in a shorter period of time, based on information about a patient's medical history, symptoms and clinical examinations. However, it is vital in the design of such a system to consider what information would help physicians ensure that the diagnoses the system suggested are indeed correct, i.e. how explanations affect trust in the system and reliance on the system's decisions (Fig. 1).

This paper presents the results of an exploratory user study to investigate the effects of explanations on the users of a CDSS. Our research questions were:

RQ1. What effects do *Confidence* explanations have on users trusting a CDSS and relying on system suggestions?

RQ2. What effects do *Why* explanations have on users trusting a CDSS and relying on system suggestions?

RQ3. What should be explained to help clinicians better assess the appropriateness of the system's suggestion?

Our results provide a first insight into understanding the role of CDSS explanations on users' trust and reliance, and propose how explanations can be used as part of CDSS design to improve and facilitate further assessment of system reliability.

We will first provide an overview of related work in this area. We then describe the study set-up in more detail and present the results of our investigations. We conclude with a discussion of the implications of our work.

II. Related Work

A. System Reliability and Use

Previous work has shown that users often misuse or disuse an automated system such as a CDSS [9, 24]. Suggestions made by a CDSS can be wrong and thus the user cannot rely on them completely. However, it has also been found that users may not always be sensitive to the system's reliability, and perceive it as highly accurate or reliable, even more so than other humans [21,24,29]. Thus, users may believe a system to be more reliable than it actually is and misuse it by agreeing with incorrect suggestions, a behavior known as over-reliance [3]. Previous studies have shown that over-reliance is widespread in CDSS use [2,4,14,25,26] and that it can be more pronounced in users with low confidence in their abilities or judgment [18]. This is a major concern when targeting a CDSS at clinicians who do not have specialized knowledge. On the other hand, there is also evidence that an unreliable system may cause disuse, which might cause clinicians to over-ride suggestions [8, 9], even if those suggestions are correct, causing self-reliance. How to *counteract* over-reliance and self-reliance has not yet been extensively researched in the CDSS community.

B. Assessing System Reliability and Trust

A CDSS often applies highly complex and multi dimensional reasoning that is difficult for a user to understand [13,23] and explanations are often provided by the system to help the user understand what it is doing [11,12,26]. Previous work has stressed the importance of explaining various aspects of the decision-making process to users [10,14,20], and these different kinds of explanation types – for example, *Confidence* explanations showing the probability of the diagnosis being correct and *Why* explanations providing facts used in reasoning about the diagnosis – have been used previously in CDSS [12]. Recent work has suggested that explanations need to be carefully designed to be *sound* and *complete* but without overwhelming users with unnecessary information [17].

Systems that are understandable seem to help users determine if the output is appropriate [12,15,19] and could therefore form an effective solution to address the problem of assessing a system's reliability. However, explanations have also been shown to increase user trust in the system's reliability [16,28], which could in turn lead to over-reliance. Therefore, trust and reliance need to be judiciously balanced so that users do not trust the system too much nor understand too little. Our research addresses explanations and their relationship to trust and reliance, as well as what types of explanations would help practitioners better assess the reliability of CDSS suggestions.

III. Study Set-up

To investigate the relationship between explanations, trust and reliance, we conducted an exploratory between-group user study employing two different versions of a CDSS prototype in which we manipulated the explanations shown to participants. We conducted a qualitative analysis on the participants' decision-making, their trust, and their 'think-aloud' and interview responses.

A. Participants

Seven primary care practitioners and one nurse practitioner (5 male, 2 female), with an average of 6.5 years experience took part in the study. All had completed clinical training in general medicine and had been involved in examining and diagnosing patients, but none had specialized knowledge on balance disorders. Since less experienced healthcare practitioners with no specialist knowledge are target users for a CDSS, these participants made ideal candidates for our study.

Participants were recruited through advertisements sent to medical network groups, forums, medical schools, and local primary care offices. They were given a small incentive for participating in the study.

B. Vignettes

In order to simulate the experience of diagnosing patients with balance-related complaints and to maintain consistency of information provided, we created eight clinical vignettes. Each clinical vignette described a fictitious patient's age and gender, their medical history (including occupation and drinking/smoking habits), symptoms, and the results of four clinical examinations, based on clinical expertise in our research project. Each key fact in a vignette was printed on a separate piece of paper that was given to participants as the study progressed, mimicking the iterative disclosure of information during a consultation. Table 1 shows an overview of the information provided for all eight vignettes.

C. Prototypes

As part of our study, we developed a CDSS interface with which participants could interact (Fig. 2). Participants entered medical history, symptoms, and examination results from the clinical vignettes and then received a suggested diagnosis from the prototype. Instead of a fully automated system, we used a "Wizard of Oz" approach in which the behavior of the software is controlled by the researcher unbeknown to the participant; in our prototype we simply mocked up the diagnoses and associated explanations.

The diagnoses were explained in two ways within the prototype. First, alongside each diagnosis the prototype showed a *Confidence* explanation in the form a percentage of certainty. In order to investigate the effect of this type of explanation on reliance, i.e. whether a high percentage caused over-reliance, we manipulated this to be either high or low: half of the vignettes' associated diagnoses were randomly given a percentage below 30% while the other half were given a high percentage above 75%.

Second, the prototype showed a *Why* explanation through a list of facts based on the vignettes that were associated with the formation of the diagnosis. We created two versions of the prototype interface in order to manipulate the level of information provided as part of these explanations. The *Comprehensive* version showed all items from medical history, symptoms, and examination results (Fig. 3, left) whereas the *Selective* prototype listed only examination results (Fig. 3, right).

A danger of using a CDSS is that the user agrees with an incorrect suggestion, known as over-reliance. To investigate this aspect in our study, four out of the eight vignettes concluded with suggested diagnoses that were incorrect. We ensured that the incorrect diagnoses were not trivial to identify by participants: the diagnosis still shared some symptoms or examination results with the correct diagnosis. To be able to isolate the impact of *Confidence* explanations and *Why* explanations on trust and reliance, we balanced incorrect diagnoses across these conditions (see Table 1, last column).

D. Procedure

Participants either used the Comprehensive or Selective version of the prototype in a between-group study design. We considered a within-subject design but decided that this would be too confusing to participants, making the prototype appear to be unpredictable. To counter any confounding effects due to participant's level of experience, we balanced group assignment based on their stated level of knowledge of balance disorders. Overall, four participants used the Comprehensive version while three used the Selective version. Each participant was asked to consider eight vignettes; one participant was only able to complete four vignettes. This resulted in a total of 52 vignettes completed altogether: 28 by the Comprehensive group and 24 by the Selective group.

TABLE 1. THE CLINICAL VIGNETTES USED IN THE STUDY

Medical history	Symptoms	Examinations	Diagnosis, Confidence, In/Correct
Male, 30, High Stress	Left sided tinnitus, Vertigo upon quickly standing	Gaze test: Normal, Gait Test: Small steps, Romberg: Normal, Head hang: Positive	Vestibular Neuritis, 24% (Incorrect)
Female, 42, menopausal, job affected by vertigo	Recent falls, Right-sided tinnitus, Vertigo triggered by diet	Gaze test: Motion intolerance, Gait Test: Normal, Head hang: Normal, Smooth Pursuit: Motion intolerance	Vestibular Migraine, 26% (Correct)
Male, 82, Retiree	Recent falls, hearing loss, Vertigo	Gait test: Small steps, Head Thrust: Normal , Smooth Pursuit: Normal , Romberg: Normal	Age-related Imbalance, 28% (Incorrect)
Male, 51, Drinks 10 units/week, non-smoker, recent change in sleep habits	Recent falls, vertigo triggered by sitting up	Semont: Normal, Smooth pursuit: Normal, Head thrust: Normal, Dix-Hallpike: Positive	Anterior Canal BPPV, 78% (Correct)
Female, 64, Retiree, smokes 1 pack/week	Hearing loss evolution fluctuating, bi-lateral tinnitus, spontaneous vertigo	Romberg: Normal, Smooth pursuit: Normal, Semont: Normal, Dix-Hallpike: Normal	Meniere's Disease, 19% (Correct)
Female, 43, Non-smoker, Drinks 5 units/week	Recent falls, Bi-lateral tinnitus, Vertigo triggered by rolling in bed	Gait test: Normal, Romberg: Normal, Dix-Hallpike: Normal, Horizontal Roll Test: Downbeat Nystagmus	H-BPPV, 84% (Correct)
Female, 22, Student, Drinks 65 units/week	Recent falls, Tinnitus, Vomiting	Gait test: small steps taken, Dix-Hallpike: Normal, Horizontal Roll Test: Normal, Semont: Normal	Vestibular Schwannoma, 81% (Incorrect)
Male, 58, Unemployed Smokes 2 packs/week Drinks 5 units/week	Vertigo triggered by rolling in bed	Romberg: Normal, Horizontal Roll test: Positional nystagmus, Head thrust: Positional Nystagmus, Semont: Positional Nystagmus	Vertebro-basiler Ischaemia, 77% (Incorrect)

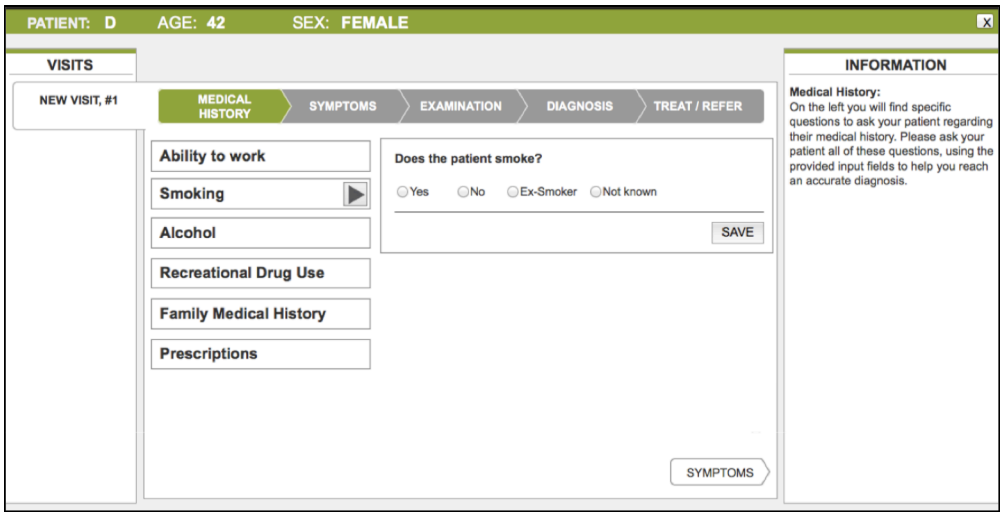

Fig. 2. Example interface of the CDSS prototype used in the study. Participants were able to enter medical history, symptoms and the results of clinical examinations to be shown a diagnosis.

Each study session lasted approximately 1.5 hours. Before they were asked to consider the vignettes, participants were familiarized with the CDSS prototype and were told that the suggestions made by the prototype might not always be correct. Regarding how the system determined a diagnosis, the participants were told only that it has a database; no further detail was provided. They then rated their trust of the system, before using the system for the first time. The main part of the study consisted of a participant considering each of the eight vignettes in turn, entering the information provided into the prototype and then considering the suggested diagnoses, either accepting the diagnosis as correct or rejecting it as incorrect. Hence, the task performed by participants in our study is akin to CDSS use in a real-world application during a typical consultation workflow (Fig. 4). As they worked through the vignettes using the prototype, we asked them to "think aloud" to verbalize their thoughts and reasoning. At the end of the study, participants were asked to rate their trust toward the prototype post-use and they were interviewed about the explanations' impact on their experience.

E. Data Collection and Analysis

We used the difference between the trust ratings participants provided pre- and post-use, rated on a 7-point Likert scale ranging from 'distrust completely' (a rating of 1) to 'trust completely' (a rating of 7), to measure the impact of the explanations on their assessment of the reliability of the tool.

We also investigated the effects of the explanations on system reliance through their verbal responses by noting how often participants agreed with a diagnosis made by the system, how often they made the 'right' decision (i.e. they agreed with the correct diagnosis or rejected the wrong diagnosis) and how often participants made the 'wrong' decision (i.e. they agreed with the wrong diagnosis or disagreed with a right diagnosis).

All sessions were video recorded, screen captured and

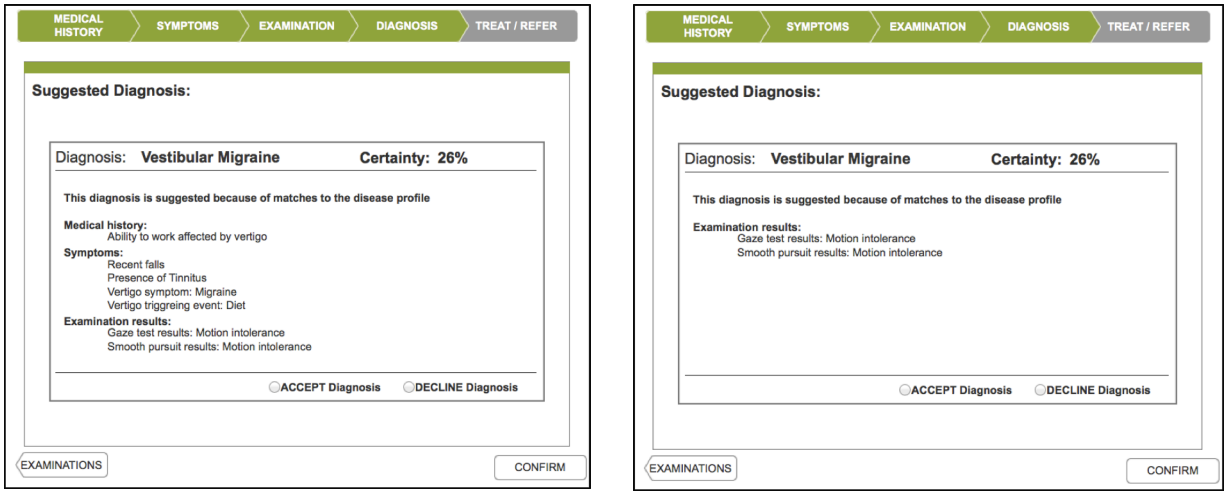

Fig. 3. The two versions of the prototype. The Comprehensive version (left) showed all information associated with a diagnosis while the Selective version (right) showed a less detailed explanation.

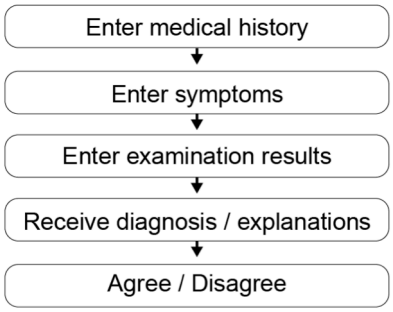

Fig. 4. Flow of main study task

transcribed. We used a thematic analysis approach [6] to better understand the participants' thoughts about the explanations' impact on trust, and what further information would have helped them understand the reasoning of the CDSS.

We used a qualitative approach to analyze all of our data, giving raw counts to illustrate our findings. Due to the low sample size, we did not conduct any statistical tests but instead include visualizations which provide an intuitive description of the data distribution.

IV. Results

We address our research questions in turn. Table 2 shows, for each participant using the two versions, the frequency of the diagnoses made by the system ('System Suggestions' columns) and the frequency of decisions made by participants ('Decisions by Participants' columns).

A. Confidence Explanations (RQ1)

To understand the effect that *Confidence* explanations had on reliance, we analyzed participants' decisions to agree with the system based on the percentage – either high or low – that was shown with the diagnostic suggestion (Table 2, 'By Confidence' columns). Overall there were 52 suggestions that participants saw, equally split between high and low percentages. Participants agreed with both roughly equally (Fig. 5): 21 when they were being shown with high confidence percentages, compared with 18 with low confidence percentages. In addition, only one participant mentioned the system's confidence as an important factor in trusting the system:

"There is this degree of certainty that makes me trust more in this system." [C04]

These results seem to indicate that high system confidence had only a slight effect that led participants to over-rely. The small impact is surprising considering existing design guidelines [12]; possible reasons for this result will be further described in section IV.C.

B. Why Explanations (RQ2)

We next turned our investigation to the impact of *Why* explanations on participants' tendency to rely on the system. To do this, we first compared how many times participants who had been shown either Comprehensive or Selective explanations made 'right' decisions (i.e. agreeing with a correct suggestion or disagreeing with an incorrect one) vs. 'wrong' decisions (i.e. disagreeing with a correct suggestion or agreeing with an incorrect one). Again, we found that both groups did roughly equally well (Table 2, 'By Right/Wrong Decisions'): out of 52 suggestions, participants in the Comprehensive group made 16 'right' decisions, whereas the Selective group made 14 'right' decisions (Fig. 6, black ticks and crosses). It would therefore appear that the amount of information provided in an explanation had no impact on the correctness of the decisions made by participants.

However, there is an important difference between agreeing with a system suggestion that is incorrect versus disagreeing with a correct one. While both are wrong, the former indicates an over-reliance on the system, whereas the latter shows that the user does not simply follow what the system presents. In both groups, participants made wrong decisions in over half of the instances. We therefore looked into the pattern of making 'wrong' decisions in more detail, first when participants agreed

TABLE 2. Frequencies of Suggestions and Decisions, by Participant

Prototype Version	Participant ID	System Suggestions			Decisions by Participants							
					By Confidence		By Right/Wrong Decisions					
		Diagnoses Shown	Of Which Correct	Of Which High Confidence	Agree w/ High Confidence	Agree w/ Low Confidence	'Right' Decisions	'Wrong' Decisions	Agree w/ Correct	Disagree w/ Incorrect	Agree w/ Incorrect	Disagree w/ Correct
Comprehensive	C01	8	4	4	3	1	6	2	3	3	1	1
	C02	8	4	4	3	4	5	3	4	1	3	0
	C03	4	1	2	2	2	1	3	1	0	3	0
	C04	8	4	4	4	4	4	4	4	0	4	0
Selective	S01	8	4	4	4	4	4	4	4	0	4	0
	S02	8	4	4	2	2	6	2	3	3	1	1
	S03	8	4	4	3	1	4	4	2	2	2	2

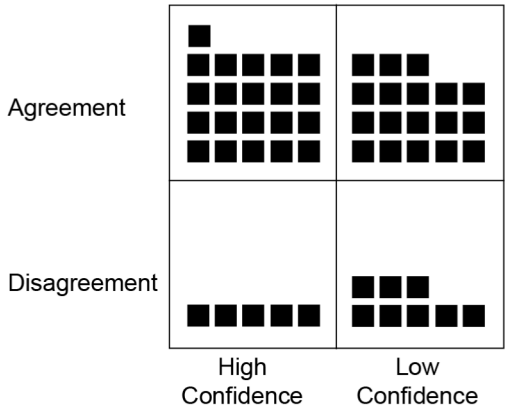

Fig. 5. Number of instances across all participants of either agreeing or disagreeing with a suggestion when associated with a high or low *Confidence* explanation. Participants agreed slightly more often with suggestions associated with high percentages than with low ones.

with incorrect suggestions and then when they disagreed with a correct one.

Our results show that the amount of information presented in the *Why* explanation does seem to matter in *agreeing* with *incorrect* suggestions. Participants in the Comprehensive group agreed with more suggestions and also with more incorrect ones, whereas participants in the Selective group agreed with only 7 incorrect ones (Fig. 6, top row). Hence, these findings suggest that the participants in the Comprehensive group tended to agree with incorrect suggestions made by the system and suffered from over-reliance.

A possible reason for this over-reliance was that the participants receiving Comprehensive explanations were exposed to additional justifications, convincing them to go along with the system even though they knew that the system sometimes erred. The verbalizations of participants show the persuasive nature of the Comprehensive explanation, disregarding their own diagnostic hypothesis and agreeing with an incorrect suggestion:

"I guess this thing knows more than me. The system knows more than me. I'll accept [the diagnosis]." [C02]

"I would never have thought it would be this, but the software is telling me it is. It's made me reconsider." [C04]

Further evidence for this persuasive effect of *Why* explanations comes from the participants' trust ratings. Three of the four participants trusted the system more after being given Comprehensive explanations, while only one of the Selective group showed an increase in trust (Fig. 7).

We wondered about what aspects of the Comprehensive *Why* explanations could have persuaded participants to trust the system. Our analysis showed that this seemed to occur in three main ways. First, the explanations convinced participants that the system used up-to-date and detailed medical knowledge to determine the suggestion. For example:

"More, [the explanations] make me trust it more. It's proof that the system is matching the information with digital medical knowledge." [C04]

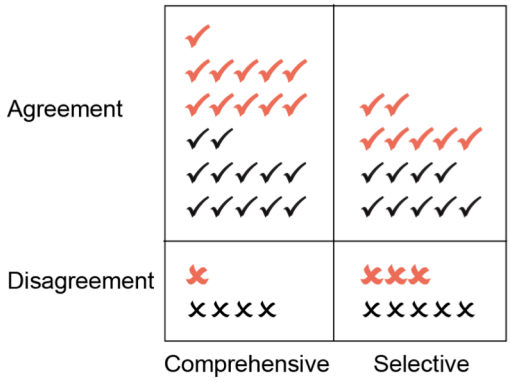

Fig. 6. Number of instances in which participants using either the Comprehensive or Selective version agreed (✓) or disagreed (✗) with the system's suggestion, and whether those decisions were 'right' (black) or 'wrong' (red). The Comprehensive group agreed more frequently, while the Selective group often disagreed, even if the suggestion was correct.

Second, participants thought that the system had a way to determine salient features that mattered in a diagnosis:

"The things it gives me there - on that last screen - for the most part, the salient features, they're the most important thing when making a diagnosis of x or y." [C02]

These prominent features were thought to directly lead to a diagnosis:

"There's a link behind that – an algorithm that links with my patient's information. So there is an algorithm that knows the latest research in medical knowledge – this test is positive, so literature says that this test is related to specificity or sensitivity with this disease." [C04]

Third, comprehensive explanations also seemed to inspire greater trust because it led participants to believe that the system used a method of reasoning similar to their own. Work in automated and context-aware systems has suggested that a user's trust is impacted by their perception of the system's abilities and perception that the system follows their own reasoning [16,27]. The results from this study suggest the same to hold true for users of clinical decision support systems:

"They do impact my trust. It seems to me like it follows the system of, you know, the same system of decision-making that we use." [C03]

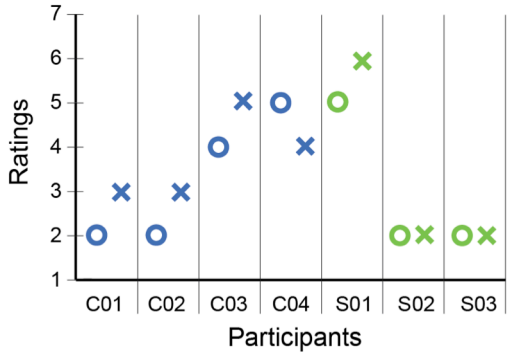

Fig. 7. Trust ratings per participant before using the system (circles) and after system use (crosses). Participants who saw Comprehensive *Why* explanations (blue) usually increased their trust increased trust compared to the Selective group (green).

We then turned our investigation back to the pattern of wrong decisions made when participants *disagreed* with a *correct* system suggestion, i.e. the flip-side of over-reliance. Nearly one third of the decisions made by the Selective group were disagreements, including three with correctly suggested diagnoses, whereas the ratio of disagreements in the Comprehensive group was much lower, and only one of them was with a correct suggestion. This suggests that showing less information in the explanations caused unwarranted self-reliance, that is, if not given enough information a user may choose to rely on their own limited knowledge rather than that of a CDSS.

We then looked at possible reasons for self-reliance and how it related to trust. Because the selective explanations showed only the matching examination results, these participants assumed the CDSS did not consider the symptoms or medical history of the patient:

"It's focusing solely on the investigation, the examinations, whereas the way that I work, I focus more on the history and the story that [the patients] give me. It only says exams, so it does lower my confidence." [S03]

Thus, the Selective explanations caused the participants to perceive that the system applied a reasoning process that is inadequate compared to their own:

"I know it doesn't take into account the things that I would have been looking for, for example the clinical details that I would have thought were relevant. So it doesn't take that into account so I know it's not thinking along the same lines that I am." [S02]

The results presented here suggest that more information presented as part of system explanations might lead users of a CDSS to over-rely on the system and accept more incorrect diagnoses. It is hence tempting to avoid explanations altogether so as not to persuade users to trust suggestions blindly and cause over-reliance. However, this approach might also be counter-productive as less information in explanations led to self-reliance and made participants choose wrong diagnoses; this effect might be worsened if no explanations are provided at all.

C. Desired Explanations (RQ3)

It has been suggested that explanations can help a user to identify a system's mistakes [15] but maybe different information is needed than that provided in our study. We next investigated the information that could be included in explanations to support clinical users in assessing the reliability of system suggestions. In this section, we consider the statements made by all participants together, rather than by group. From our thematic coding of the data four desirable types of explanatory information emerged, each of which will be described in turn.

1) How sure are you, and how do you know that?

Simply showing a *Confidence* explanation as a certainty percentage proved problematic. Four participants made comments that showed they did not understand what the percentage meant. Indeed, there are many ways that the term 'certainty' can be interpreted, and our participants' understanding of the percentage reflected this ambiguity. For example:

"So when it says the certainty is 19%, um… What does that mean? Because it's got a diagnosis of Meniere's here and so is that sort of like saying that the certainty of the diagnosis is a 19% chance of that being correct?" [S03]

In some instances, participants not only questioned the meaning of the percentage but also showed that it was important to them to understand how the system derived the Confidence explanation:

"Where does this figure come from? Where does the software calculate the low degree of this figure?" [C04]

The need for additional explanation appeared to be even higher when the percentage was low or counter to what was expected. Perhaps these questions are triggered by the *surprise* experienced by the participants; it has been shown that explanations become important 'when the user perceives an anomaly' [11]. The instances we observed show this kind of critical thinking followed after a surprise, for example:

"The ones I would agree with are the ones with the lowest degree of certainty and that kind of puzzles me because the only time I had a plausible diagnosis it was something completely unrelated, or something I hadn't even remotely thought about and it had a high degree of certainty. You see what I mean? That was the only time I had something in mind and something completely different comes up with a pretty high degree of certainty." [C04]

"It sounds like that's what it is, but I don't know why the system is not certain." [C02]

In addition, it appeared that these participants were also looking for *Confidence* explanations that would cover the suggested diagnosis in relation to all other possibilities:

"Is that sort of like, that's basically it and all other things are rejected?" [S03]

"Why some of it is 81% and some of it is 17%? Does it mean it's not too sure but it is pointing to this [diagnosis]?" [C03]

"If this is the only diagnosis that it has come up with, then why is the certainty so low?" [C02]

Providing a *Confidence* explanation for a suggestion in the form of a percentage is common practice in CDSS design and has been included in design guidelines for increasing system intelligibility [20]. However, our results suggest that this is not enough. First, the system needs to clearly communicate what the figure means; this could be done through careful selection of terminology or a general definition available on request. Second, system confidence should be shown for *all* possible suggestions so that users can understand the suggestion being made in context of the wider decision-making. Last, it is important to show how confidence was derived from the evidence presented, perhaps upon the user's request when the percentage is felt to be low or surprising.

2) Does the disorder fit the suggestion?

Four participants also requested a disease description in addition to the suggested diagnosis. The additional details requested could be broken down into two types: risks and symptoms, and a summary of a typical case.

Some of the statements made by three participants showed that they wanted the system to show a list of common risk factors, causes, or symptoms associated with the suggested diagnosis, even when the system provided the facts it used to make a diagnosis. For example:

"I'd want to see other things that would cause it. Because it's some sort of stroke. So any other risk factors – patient weight, history of heart disease, history of angina?" [C02]

"Okay, like... Having a little summary. Like, if it's Meniere's, to say 'Features such as hearing loss and tinnitus suggest Meniere's disease or Acoustic Neuroma.' Something like that." [C03]

"So, I don't know what side it's on and I can't remember, I think with Vestibular Schwannoma you tend to get it on one side… That would've helped me make my diagnosis." [S01]

Another common request was for a description of a *typical* case, for example:

"For example, it'd be a little blurb. They would be 90 years old, they have this unexplained imbalance, all other tests are normal, all other things have been ruled out, it's for days rather than minutes that she gets these intermittent dizziness things, and then it would be easy to be like 'she fits this criteria, this criteria, and this criteria, so I believe it's this diagnosis.'" [S02]

This suggests that users could benefit from being able to verify that the suggestion does indeed correspond with the suggested diagnosis. This seems especially important if the user lacks experience in the domain yet still carries the responsibility of making the final diagnostic decision. It appears that providing risk factors and symptoms could help clinicians identify whether the suggestion fits the described disorder, thereby double-checking the system's reasoning. The typical case information, on the other hand, could help them to assess *how much* the suggestion fits – whether this is a typical or atypical presentation of the disorder.

3) Why, in detail?

The *Why* explanations provided a list of the patient information that led to the suggested diagnosis, but four participants mentioned that three additional types of information would be appreciated: the signs pointing *against* the diagnosis, the pathological link between the findings and the diagnosis, and the weights that the decision model assigns to different parts of the patient's information.

In decision-making, it is important to weigh up the pros and cons, and to seek both confirming and counter-evidence [5]. Statements made by participants in our study reiterated this, pointing out that the system should also explain the arguments against the current suggestion:

"So, in medicine it's not just the positives, it's the negatives that you look out for as well. If I have a very clear positive, if that was the only thing I had about the patient then that's fine, but I'd have to have the strong negatives. Only then would I accept a diagnosis." [S02]

Other instances showed that the list of facts was too simplistic and needed to be supplemented by some deeper chain of reasoning. Statements made by two participants suggest that presenting the pathological link between the patient information and the suggested diagnosis would be desirable. For example:

"So, like, if it said 'Dix Hallpike shows rotational nystagmus which is suggestive of autillis in the ear canal causing patient's vertigo symptoms in combination with positional element' then I would understand that." [C01]

Without the system describing the pathological link, the list of matching information was seen as merely a regurgitation of the information that was previously input:

"All it says is that they've got migraines, which they've obviously come in and told me. It's triggered by diet, which it could be. They've got some tinnitus and vertigo. So, it just tells me what I know already. It doesn't explain why it's come to that conclusion. See what I mean?" [C01]

Previous research on medical reasoning show that less experienced physicians rely heavily on their 'knowledge of underlying pathophysiology and anatomy,' or biomedical knowledge, when trying to reach a diagnosis [1]. In many of the decision-making instances the participants, who did not have expertise in balance disorders, drew biomedical inferences from the provided patient data. These statements suggest that if the system were to provide these connections for them, the users would feel more confident in understanding if the diagnosis is correct.

Finally, three participants requested an indication of how much the features contributed to the diagnosis. These instances showed a desire for insight on the decision model, specifically the weights that were assigned to the various pieces of patient information:

"I think if it explained the significance of positive findings on an examination then I would be more willing to accept it." [C01]

These results seem to suggest that clinicians would appreciate explanations that more closely reflect how they reason and that also explain the underlying decision model used to make these suggestions. This could lead to better understanding why a particular suggestion was made; however, this also has be carefully balanced against leading users to over-rely on the diagnosis.

4) Differential diagnosis

A differential diagnosis is, simply put, an alternative diagnosis that a physician tries to 'rule out' or disprove. In effect, there is a process of elimination in this reasoning procedure, instead of using inference from symptoms. This is associated with, but distinct from, wanting both positive and negative evidence for a diagnosis. In one quarter of the suggestions made by the system, participants desired the system to follow or show this procedure:

"It would be helpful to sort of say: 'The second differential is xyz and these are the things which it fulfills. But the Hallpike result is normal and therefore it seems against it.'" [S02]

The instances in which it was mentioned often coincided with uncertainty on the clinician's behalf:

"As I said before, when I'm not sure of the diagnosis I would want it to suggest other things it could be." [C02]

"So it's sometimes quite useful to look at differential

diagnoses because it kind of prompts you to think 'Actually, yeah, that might be a possibility.'" [S03]

Moreover, a differential was seen as a way to determine if the system was correct:

"Lots of things match lots of disease profiles. That's why we work with a differential diagnosis." [C01]

These results suggest that showing competing hypotheses – a differential diagnosis – is a necessary explanation to include in any CDSS designed to aid a user in assessing a suggestion. Providing one or more differentials would enable the user to make a more informed decision by weighing the positives and negatives of each disorder. A system capable of computing and presenting differential diagnoses in addition to a primary diagnosis would allow the clinical user to explore and examine the broader picture.

V. Discussion

We acknowledge some limitations of our study. First, our results are limited in their generalizability due to the low sample of participants, despite our best efforts. Other work endeavoring to target primary care physicians would do well to consider the difficulty in recruiting these end users, especially as their time constraints places considerable burdens in involving them in any research projects. Second, our study did not follow a full factorial experiment which also would have required a large number of participants to validate any hypotheses. At this stage, to inform a model of how trust, reliance and explanations are possibly related, we focused on a small set of aspects aimed at gathering a snapshot of the decision-making process of clinicians interacting with a CDSS. Further studies are required to include other contextual factors and considerations to flesh out this model, and larger studies are needed to validate this slowly evolving model in clinical settings.

Although there are limitations, our results provide preliminary indications on some important points. Our study has both practical implications regarding CDSS design and the future direction of research in this area, arising from a greater understanding of the relationship between trust and reliance, more knowledge about the role of explanations in trust and reliance, and how to create better explanations to help users assess system reliability.

First, our findings indicate a strong relationship between a user's trust in a CDSS and their reliance on the system. CDSS users who trust the system highly are also likely to over-rely on the system's suggestions, while users who distrust the system are likely to rely on their own knowledge, even if it is poor. This is an issue that is difficult to address. In the early stages of system use, promoting user trust is often necessary in order for the system to be adopted. However, as shown here, this trust could then also lead to over-reliance and potentially dangerous clinical outcomes. Thus, a system needs to inspire an *appropriate* amount of trust but more work is needed to find the right balance between trust and reliance.

This issue also needs to be considered in the wider context of how trust in a system is established. Trust can also be influenced by other factors in system use, such as previous experience with similar systems, personal characteristics of the user, and the reliability of the system's reasoning. Previous work has suggested that presenting metrics about the accuracy of the decision model – like the sensitivity and specificity – could provide the user with a more rounded understanding of the system's reliability [9]. Given the confusion regarding the *Confidence* explanation, it is possible that these metrics would provide additional help to clinician to calibrate their trust and reliance. However, we currently do not know enough about these aspects and their relationship to trust and, in turn, reliance.

Second, our results indicate that explanations have effects on trust and reliance. Whilst a more detailed explanation may promote over-reliance, we argue that providing no explanation at all is not a viable option as they are desirable and necessary. Our results showed that without providing explanations there is a danger that users will rely too much on themselves because they do not understand how the system works. The four new explanation types identified in this study suggest that explanations are a necessity for users in order for them to accurately assess the veracity of a CDSS's suggestion. However, more work is needed to establish the impact of explanations on perceived intelligibility of the prototype and the link to trust and reliance.

Finally, our work has suggested the need for more and better CDSS explanations. This raises questions about how these explanations should be communicated in the interface as well as their effects of trust and reliance. Explanations are designed and hence choices have to be made as to *how much, when,* and *how* this information is presented, not only to avoid over-reliance, but also to avoid information overload. Only careful and extensive evaluation with users will be able to establish appropriate design guidelines. Furthermore, these user studies also will be able to tell us more about the impact of these new explanation types on the trust and reliance of users.

VI. Conclusion

In this paper we have presented the results of an exploratory user study investigating the effects of explanations on trust and reliance of CDSS suggestions, through a prototype developed as part of the EMBalance project.

We have shown that:

- *Confidence* explanations did not seem to sway participants to increase their trust and reliance on the system. However, explaining more about how sure the system is could still be helpful to clinicians to assess reliability;
- Comprehensive *Why* explanations promoted over-reliance while Selective *Why* explanations promoted self-reliance, and both can lead to incorrect diagnoses. This seemed to be because explanations had an effect on trust in the system. CDSS design will need to strike a careful balance to result in *appropriate* trust.
- Clinicians may be better situated to assess the appropriateness of a system's suggestion if provided with explanations that allow them to verify the disorder fits the suggested diagnosis, to follow the reasoning to obtain the

suggested diagnosis, and be shown differential diagnoses. However, more work is needed to understand the impact of these explanation types on intelligibility, trust and reliance.

Ultimately, the findings of this research take a first step toward understanding how explanations can support healthcare professionals in better decision-making with a CDSS so that the true benefits of this collaborative work can be realized by clinicians and patients alike:

"Working in general practice is a hard job. I sit here on my own. I have to use my own knowledge. So this is like having another person. I think that's very good." [C03]

ACKNOWLEDGMENT

We thank the participants of our study. This research has been supported by EU FP7, EMBalance project grant number 610454.

REFERENCES

[1] E. Alberdi, J. C. Becher, K. Gilhooly, J. Hunter, R. Logie, A. Lyon, N. McIntosh, and J. Reiss, "Expertise and the interpretation of computerized physiological data: implications for the design of computerized monitoring in neonatal intensive care," International Journal of Human-Computer Studies, vol. 55(3), pp.191-216, 2001.

[2] E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, "Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography", Academic Radiology, vol. 11(8), pp.909-18, 2004.

[3] E. Alberdi, P. Ayton, A. Povyakalo, and L. Strigini, "Automation bias and system design: a case study in medical application," People and Systems - Who Are We Designing For, The IEE and MOD HFI DTC Symposium, London, pp. 53-60, 2005.

[4] E. Alberdi, L. Stringini, A. Povyakalo, and P. Ayton, "Why are people's decisions sometimes worse with computer support?" Computer Safety, Reliability, and Security, pp. 18-31, 2009.

[5] J. Baron, Thinking and Deciding, 3rd ed., Cambridge: Cambridge University Press, 2000.

[6] V. Braun, and V. Clarke, "Using thematic analysis in psychology," Qualitative research in psychology, vol. 3(2), pp. 77-101, 2006.

[7] J. Calvillo-Arbizu, and L. M. Roa-Romero, "Design of a clinical decision support system for assisting in empiric antibiotic treatments," in The International Conference on Health Informatics (ICHI), 2014, pp. 304-307.

[8] D. Dowding, N. Mitchell, R. Randell, R. Foster, V. Lattimer, and C. Thompson, "Nurses' use of computerised clinical decision support systems: a case site analysis," Journal of Clinical Nursing, vol. 18, pp.1159-1167. 2009.

[9] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," International Journal of Human-Computer Studies, vol. 58(1), pp.697-718, 2003.

[10] M. S. Gonul, D. Onkal, M. Lawrence, "The effects of structural characteristics of explanations on use of a DSS," Decision Support Systems, vol. 42(1), pp.1481-93, 2006.

[11] S. Gregor, and I. Benbasat, "Explanations from intelligent systems: Theoretical foundations and implications for practice," MIS Quarterly, vol. 23(4), pp.497-530, 1999.

[12] D.W. Hasling, W.J. Clancey, and G. Rennels, "Strategic explanations for a diagnostic consultation system," International Journal of Man-Machine Studies, vol. 20(1), pp.3-19, 1984.

[13] R. Islam, C. Weir, and G. del Fiol, G, "Heuristics in Managing Complex Clinical Decision Tasks in Experts' Decision Making," International Conference on Healthcare Informatics (ICHI) (pp. 186-193). IEEE.

[14] G. Kong, D.-L. Xu, and J.-B. Yang, "Clinical Decision Support Systems: A review on knowledge representation and inference under uncertainties," International Journal of Computational Intelligence Systems, vol. 1(2), pp.159-67, 2008.

[15] T. Kulesza, M. Burnett, S. Stumpf, W.-K. Wong, S. Das, A. Groce, A. Shinsel, F. Bice, and K. McIntosh, "Where are my intelligent assistant's mistakes? A systematic testing approach," End-User Development, pp. 171-186, 2011.

[16] T. Kulesza, S. Stumpf, M. Burnett, I. Kwan, "Tell me more?: The effects of mental model soundness on personalizing an intelligent agent" Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1-10, 2012.

[17] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.K. Wong, "Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models," IEEE Symposium on Visual Languages and Human-Centric Computing, pp 3-10, 2013.

[18] J.D. Lee, and K.A. See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors: The Journal of Human Factors and Ergonomics Society, vol. 46(1), pp.50-80, 2004.

[19] B.Y. Lim, A.K. Dey, and D. Avrahami, "Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems," The SIGCHI Conference on Human Factors in Computing Systems - Studying Intelligent Systems, pp. 2119-2128, 2009.

[20] B.Y Lim, and A.K Dey, "Toolkit to Support Intelligibility in Context-Aware Applications," The 12th ACM International Conference on Ubiquitous Computing, pp. 13-22, 2010.

[21] P. Madhavan, D.A. Wiegmann, "Similarities and differences between human-human and human-automation trust: an integrative review," Theoretical Issues in Ergonomics Science, vol. 8(4), pp.277-301, 2007.

[22] S. Medlock, S. Eslami, M. Askari, D.L. Arts, S.E. de Rooij, A. Abu-Hanna, and A.M. Lagaay, "Development of Computerized Clinical Decision Support to Assist in Detecting and Preventing Delirium in the Hospital Setting," The International Conference on Healthcare Informatics, pp. 95-100, 2014.

[23] D. O'Sullivan, P. Fraccaro, E. Carson, E. and P. Weller, "Decision time for clinical decision support systems," Clinical Medicine, vol. 14(4), pp.338-41, 2014.

[24] R. Parasuraman, and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 39(2), pp.230-53, 2007.

[25] A.A. Povyakalo, E. Alberdi, L. Strigini, and P. Ayton, "How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography," Medical Decision Making, vol. 33(1), pp.98-107, 2013.

[26] P. Pu, and L. Chen, "Trust-inspiring explanation interfaces for recommender systems," Knowledge-Based Systems, vol. 20, pp.542-56, 2007.

[27] Royal College of Physicians, "Hearing And Balance Disorders: Achieving Excellence In Diagnosis And Management," Report Of A Working Party, London: RCP, 2007.

[28] Y. Seong, and A.M. Bisantz, "The impact of cognitive feedback on judgment performance and trust with decision aids," International Journal of Industrial Ergonomics, vol. 38(7), pp.608-625, 2008.

[29] D.A. Wiegmann, A. Rich, and H. Zhang, "Automated diagnostic aids: the effects of aid reliability on users' trust and reliance," Theoretical Issues in Ergonomics Science, vol.2(4), pp.352-67, 2001.