Assessing the impact of inaccurate decision support systems on experts' behavior and decisions

Clemens Schöll
Quality and Usability Lab,
TU Berlin, Germany
clemens@neopostmodern.com

Sebastian Möller
Quality and Usability Lab,
TU Berlin, Germany
sebastian.moeller@tu-berlin.de

August Longino
University of Washington,
Seattle, USA
august.longino@gmail.com

***Abstract*— Decision support systems increasingly support experts' work, but errors caused by them may have severe consequences, especially in the medical domain. To better understand how these inaccuracies affect experts' behavior and decisions, a quality of context model was constructed. Based on the model a controlled online experiment was conducted in which physicians had to treat hypothetical cases while seeing inaccurate medication suggestions. Initial results indicate that raw information quality and cover rate significantly improve patient outcomes and reduce required time, whereas high misleadingness leads to poorer patient outcomes and more time being required.**

***Keywords*— decision support systems; recommender systems; quality of context; risk modelling**

I. Introduction

Imagine you are the doctor in a clinic with limited resources, where only two anti-epileptic medications are available: lamotrigine and carbamazepine. A pregnant woman with seizures comes in and your software system suggests to treat her with carbamazepine. You accept, forgetting that only lamotrigine is suitable for pregnant patients. An *adverse drug event* (ADE) occurs and might harm the patient.

In such situations a crucial task is the identification of relevant information within an overwhelming quantity of possibilities. Decision support systems (DSS) offer a mechanism to help cope with this [1], [2]. Given the ease of building, training and deploying them nowadays, on the one hand, and great success stories, on the other hand, many rush to adopt them in their systems [3]. They are starting to enter fields where decision support is not only speeding up workflows but decisions bear a high responsibility [4], and where wrong decisions can have negative or fatal consequences. Thus, the aim of the present paper is to analyze how the quality of the decision support influences the behavior and the decisions taken by experts.

II. Related work

Traditionally, the decision making process in interaction with such systems was regarded as a "black box" [5]. For *casual tasks* numerous influences of recommender systems on the decision have been observed, such as context effects, primacy/recency effects, and framing effects [6]; or participants following a communicated rating prediction in their own rating [7]. However, it is unclear how these can be transferred to routine tasks carried out by experts.

QoMEX2017 – Erfurt, Germany; 978-1-5386-4024-1/17/$31.00 ©2017 IEEE

Decision-making applied by experts can be expressed in a cognition model of three performative modes [8]: *skill-based*, the largely unconscious application of schemata; *rule-based*, the conscious solving of problems that are rather familiar, based on simple, stored logical rules; and *knowledge-based*, conscious, analytical thinking and the processing of stored knowledge when facing novel problems. Errors may occur for each of these modes, called *slips* when a skill-based action is broken while not being consciously monitored, or *mistakes* for rule-based cognition (usually a misapplied rule) and for knowledge-based cognition (misinterpretations of the problem or a mere lack of knowledge to be applied).

Clinical decision support systems (CDSS) were shown to have an impact on practitioner performance. Most studies find evidence for an improvement in practitioner performance, while the impact on patient outcomes is less definitive but tends to be positive. Surprisingly, for the latter, several reviews reported no statistical significance [9], [10]. Recent research addressed the problem that CDSSs themselves can fail, and that this may decrease the trust in CDSSs [11].

III. Quality of decision context model

In scenarios where relevant information is being displayed to aid a decision and data-entry process, Möller et al. define two information sets: The required useful context (R) and the actually provided context (C) [12]. Based on this they argue that only two parameters are necessary to accurately model QoC: *raw information quality* (RIQ), the ratio of provided required context versus the entire provided context and *cover rate* (CR), the ratio of provided required context versus the entire required context respectively. Quantitative definitions can be expressed based on the information sets with $Q_{\text{RIQ}} = {}^{|R \cap C|}/_{|C|}$ and $Q_{\text{CR}} = {}^{|R \cap C|}/_{|R|}$ respectively.

For mistakes facilitated by a DSS, such as the one described in the introduction, these two parameters fail to describe the complexity of the situation that leads to an erroneous decision, such as an ADE. We therefore introduce the *misleadingness*, i.e. the likelihood of information within a context to lead to the misapplication of a rule, by seeming adequate and resulting in an erroneous or negative influence on the user's decision. This definition is targeted at rule-based errors. Slips are not explicitly covered, but are assumed to be less relevant in expert settings where action and decision are held to be conscious. The misleading effect increases with the 'strength' of the rule, i.e. how frequently it is used. A user is misled if the context *appears* to be correct and trust-worthy and they is therefore seduced to apply said rule or more generally speaking a thought process resulting in a not desired action.

The following criteria were defined to lead to a high misleadingness score, rated by a human expert, in order of importance: *verisimilitude*, i.e. how adequate the harmful items seem and the strength of rules supporting them; *absence of alternatives*, the lack of alternatives which also fulfill the goal but without negative consequences; and *order effects*, when items with negative consequences are listed higher than their better counterparts.

IV. Initial experiment

To evaluate the practical implications a (so far under-powered) experiment was conducted, taking on the example of medication prescription in the medical field. Physicians, both in the choices they face and the increasingly digitalized environment, give a good example of experts with decisions bearing a high a responsibility. In an online environment physicians were presented eight text-based case descriptions along with a minimal *electronic medical record* (EMR) system indicating vital signs and, where appropriate, lab results. On the pretext that a new EMR user interface was being evaluated, they were asked to diagnose and medicate hypothetical patients from a reduced database of available diagnoses and drugs.

Beyond the control group – who did not receive any decision support – two treatments were assigned in a blocking randomized way, altering the EMR behavior: when opening the medication search field they would see a list of hand-selected medication suggestions. Treatment II would see a list of varying length indicating all required medications according to current best-practices (thus $Q_{\text{RIQ}} = Q_{\text{CR}} = 1$, $Q_{\text{M}} = 0$) while treatment I saw a fixed-length list of five medications designed to have varying levels of RIQ, CR, and misleadingness. Neither group were explained the origin of the suggestions.

23 participants were selected by their allowance to perform medical prescriptions in their home country: 21 were medical doctors and two physician assistants, with most having studied in Germany and the USA. Despite the limited participant number, the data already offers a rather clear trend. The treatments selected by the participants were rated by clinical experts on a scale from 0 ("severe harm") to 5 ("ideal"). Both RIQ and CR can be associated with a statistically significant increase of treatment rating ($p < .01$ and $p < .001$ respectively) while misleadingness statistically decreased treatment ratings ($p < .001$), see Tab. I. However, given the sample size, no significant model including all three parameters could be evaluated.

Participants that saw ideal suggestions ($Q_{\text{RIQ}} = Q_{\text{CR}} = 1$, $Q_{\text{M}} = 0$) completed the tasks faster and with better patient outcomes. Participants with inaccurate suggestions had the same outcome and times as participants with no suggestions whatsoever. The parameters proposed to model quality of context showed significant links to both patient outcomes and speed, where RIQ and CR can be linked to better outcomes and faster times, and misleadingness produces worse outcomes and slower times.

One possible limitation of the experiment is the uncertain fidelity of the participants. In an informal follow-up interview one participant, referring to the case described in the introduction, stated that "[they] would have referred such a patient to a neurologist anyway."

	T_X rating (demeaned)	T_X rating (demeaned)	T_X rating (demeaned)
RIQ	0.868^{**}		
	(3.29)		
CR		1.244^{***}	
		(4.15)	
Misleadingness			-1.676^{***}
			(-4.57)
Constant	-0.397^{*}	-0.836^{**}	0.428^{***}
	(-1.99)	(-3.20)	(3.74)
Observations	103	103	103
R-Square	0.0968	0.145	0.172
Adj. R-Square	0.0879	0.137	0.163

t statistics in parentheses; $^{+}p < 0.1, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

TABLE I. Regressions for therapy (T_X) ratings, by quality of context

V. Conclusion

Decision support systems enjoy widespread adoption in critical areas, such as the medical field, yet failures and misuse are common. One such failure are mistakes induced by erroneous suggestions. Quality of context needs to address these risks and introduce parameters assessing them. We proposed a new QoC model including *misleadingness* and experimentally validated its relevance.

The proposed model needs to be validated in a real-world setting. The risks of misleading decision support systems call for more research into improving and developing coping mechanisms for these specific, but inevitable failures.

Acknowledgment

The author would like to thank Niko Schöll (Universitat Pompeu Fabra, Barcelona) and Henry Hoffmann (ada/medx, Berlin) for their support.

REFERENCES

[1] P. Lops, M. De Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender systems handbook*. Springer, 2011, pp. 73–105.

[2] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[3] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.

[4] E. S. Berner and T. J. La Lande, "Overview of clinical decision support systems," in *Clinical decision support systems*. Springer, 2016, pp. 1–17.

[5] L. Chen, M. de Gemmis, A. Felfernig, P. Lops, F. Ricci, and G. Semeraro, "Human decision making and recommender systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 3, no. 3, p. 17, 2013.

[6] M. Mandl, A. Felfernig, E. Teppan, and M. Schubert, "Consumer decision making in knowledge-based recommendation," *Journal of Intelligent Information Systems*, vol. 37, no. 1, pp. 1–22, 2011.

[7] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, "Is seeing believing?: how recommender system interfaces affect users' opinions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 585–592.

[8] J. Rasmussen and A. Jensen, "Mental procedures in real-life tasks: a case study of electronic trouble shooting," *Ergonomics*, vol. 17, no. 3, pp. 293–307, 1974.

[9] D. L. Hunt, R. B. Haynes, S. E. Hanna, and K. Smith, "Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review," *Jama*, vol. 280, no. 15, pp. 1339–1346, 1998.

[10] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. Devereaux, J. Beyene, J. Sam, and R. B. Haynes, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review," *Jama*, vol. 293, no. 10, pp. 1223–1238, 2005.

[11] A. Wright, T.-T. T. Hickman, D. McEvoy, S. Aaron, A. Ai, J. M. Andersen, S. Hussain, R. Ramoni, J. Fiskio, D. F. Sittig *et al.*, "Analysis of clinical decision support system malfunctions: a case series and survey," *Journal of the American Medical Informatics Association*, pp. 1068–1076, 2016.

[12] S. Möller, H. Liu, and A. Spriestersbach, "Contextual sidebar: Effect of context information quality on the usability of a context-aware application," in *PQS 2016 5th ISCA/DEGA Workshop on Perceptual Quality of Systems*, 2016, pp. 34–38. [Online]. Available: http://dx.doi.org/10.21437/PQS.2016-8