German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI240871

Automation Bias in AI-Decision Support: Results from an Empirical Study

Florian KÜCKING[a,1], Ursula HÜBNER[a], Mareike PRZYSUCHA[a], Niels HANNEMANN[b], Jan-Oliver KUTZA[a,b], Maurice MOELLEKEN[d], Cornelia ERFURT-BERGE[c], Joachim DISSEMOND[d], Birgit BABITSCH[b], and Dorothee BUSCH[a,c]

[a] *Health Informatics Research Group, Osnabrück University of Applied Science, Osnabrück, Germany*

[b] *Department of New Public Health, Osnabrück University, Osnabrück, Germany*

[c] *Department of Dermatology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nürnberg, Erlgange, Germany*

[d] *Department of Dermatology, Venerology and Allergology, University Hospital of Essen, Essen, Germany*

ORCiD ID: Florian Kücking https://orcid.org/0009-0005-0808-6552

Abstract. Introduction Automation bias poses a significant challenge to the effectiveness of Clinical Decision Support Systems (CDSS), potentially compromising diagnostic accuracy. Previous research highlights trust, self-confidence, and task difficulty as key determinants. With the increasing availability of AI-enabled CDSS, automation bias attains new attention. This study therefore aims to identify factors influencing automation bias in a diagnostic task. **Methods** A quantitative intervention study with participants from different backgrounds (n = 210) was conducted, employing regression analysis to analyze potential factors. Automation bias was measured as the agreement rate with wrong AI-enabled recommendations. **Results and Discussion** Diagnostic performance, certified wound care training, physician profession, and female gender significantly reduced false agreement rates. Higher perceived benefit of the system was significantly associated with promoting false agreement. Strategies like comprehensive diagnostic training are pivotal in the prevention of automation bias when implementing CDSS. **Conclusion** Considering factors influencing automation bias when introducing a CDSS is critical to fully leverage the benefits of such a system. This study highlights that non-specialists, who stand to gain the most from CDSS, are also the most susceptible to automation bias, emphasizing the need for specialized training to mitigate this risk and ensure diagnostic accuracy and patient safety.

Keywords. Automation Bias, Clinical Decision Support, Artificial Intelligence, Wound Maceration

1. Introduction

Automation bias describes the inclination of humans to trust machines more than themselves or other experts. Automation bias poses a challenge potentially

[1] Corresponding Author, Florian Kücking, Osnabrück University of Applied Science, Health Informatics Research Group, PO Box 1940, 49009 Osnabrück, Germany; E-Mail: f.kuecking@hs-osnabrueck.de

compromising the effectiveness of Clinical Decision Support Systems (CDSS). Previous research suggests that trust in the system, self-confidence, task difficulty, cognitive demand and time pressure are important determinants for the occurrence of automation bias [1-4]. Also features of the system such as high reliability and long user experience with exactly this system could be shown in a study outside healthcare to contribute to increased trust and reduced user performance [5]. Additionally, expert knowledge could play an important role to diminish the risk of falsely accepting machine recommendations. However, a study in cardiology indicated that both experts and non-experts significantly lose accuracy in their diagnostic decisions when incorrect diagnoses are recommended by a system [3]. Correct CDSS recommendations, in turn, can significantly reduce errors [6]. Automation bias can occur not only in complex environments associated with multitasking but also in simple single tasks, particularly in diagnostic tasks with high verification complexity [4]. As several studies [1-6] demonstrate, automation bias is a phenomenon that has been known and described for some time. However, with the increasing availability of AI-enabled CDSS and other AI applications, the topic should garner new attention. This study, thus, aims to address automation bias anew and to identify factors influencing automation bias in CDSS-supported diagnostics when context factors such as workload, time pressure, high cognitive demand, experience with the system and system complexity are deliberately eliminated. Therefore, our research question was: What are determinants intrinsic to decision makers leading to automation bias?

2. Methods

2.1. Study Design, Sample and Study Execution

In the present study, automation bias was operationalized through the rate of agreement with false AI-enabled diagnostic recommendations in the diagnosis of chronic wounds. For this purpose, a quantitative intervention study was conducted, statistically controlled by diagnostic performance, profession, certified wound care training, perceived benefit, gender, and age. The intervention involved providing AI-based suggestions for diagnosing a specific complication of wound healing, namely wound edge maceration. A corresponding algorithm had been trained, tested, and validated prior to the study [7]. No explicit control group was implemented as all participants had obtained six images without AI support to assess their diagnostic performance (expressed as the rate of correct answers from 0 to 1) prior to the intervention. Subsequently, all participants undertook the same task with the AI recommendations. The participants were informed in advance about the recommendation, but not that the recommendations could also be incorrect. Six wound images were presented together with recommendations (Figure 1).

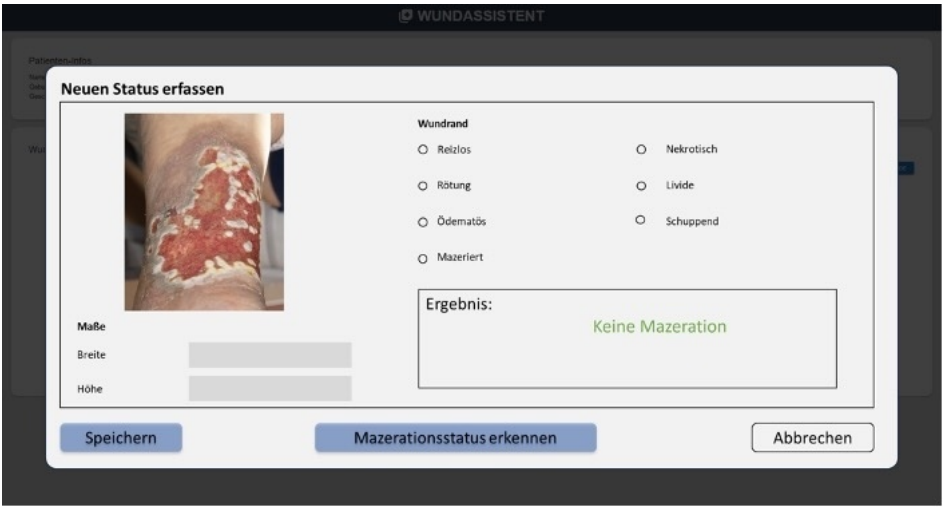

Figure 1. User Interface of CDSS used in this study showing the wound image together with the recommendation (Ergebnis). Only diagnosing maceration was enabled.

However, in 50% of the cases, incorrect recommendations were given. The presentation of the images occurred randomly to minimize potential order effects.
To establish a valid basis for correct and incorrect answers, all images had been independently assessed beforehand by experts from two German university hospitals (Essen and Erlangen). At the end, the items regarding perceived effectiveness, perceived efficiency and perceived usefulness, age, gender, profession and certified wound care training were captured via an online questionnaire (LimeSurvey).

The study took place from November 2023 to January 2024. Participants were recruited via email invitation from 1,893 hospitals in Germany and registered nurses from study programs at the two universities in Osnabrück. In total 333 persons participated of whom 210 provided complete answers and were included in the analysis. The sample showed a gender distribution, with 42.9% male (n = 90) and 57.1% female (n = 120) respondents. Furthermore, there was a professional composition of 63.3% (n = 133) nurses and 36.7% (n = 77) physicians. Regarding age, there was a distribution of 33.8% (n = 71) of participants aged 39 years or younger and 66.2% (n = 139) aged 40 years or older. In terms of the healthcare setting, 90% (n = 189) of respondents worked in the inpatient setting, while 10% (n = 21) worked in the outpatient setting.

2.2. Model Development and Parameter Estimation

To answer the research question, a regression on the dependent variable "agreement rate with incorrect recommendations" provided by the AI-based CDSS was calculated according to Eq. 1. The dependent variable was based on the counts of false recommendations normalized by the number of images. In case the participants said that they were not sure, which was neither false nor right, it was counted as 0.5 (An alternative dependent variable just counting the false and right answers had been tested and resulted in the same findings). The variable thus ranged from 0 to 1 with a value of 0 meaning

that no false recommendations were accepted while a value of 1 meant that all of them were assumed as true.

Overall, six predictor variables were included into the model: *Diagnostic Performance, Certified Wound Care Training*, *Profession*, *Perceived Benefit, Gender* and *Age.* The *Diagnostic Performance* was computed using the rate of correctly identified wound macerations in the first part of the study prior to the intervention. The predictors *Certified Wound Care Training*, *Profession*, and *Gender* were dichotomized, and for *Age*, class midpoints were formed. A *Perceived Benefit Score* was calculated summarizing the TAM 2 [8] items "Effectiveness", "Efficiency", and "Usefulness", each measured on a 7-point Likert scale, as their internal consistency (Cronbach's α = 0.964) was very high.

$$\begin{aligned} false\ agreement\ rate \\ &= \beta_0 + \beta_1(diagnostic\ performance_{score}) \\ &+ \beta_2(perceived\ benefit_{score}) \\ &+ \beta_3(certified\ wound\ care\ training) + \beta_4(profession) \\ &+ \beta_5(gender) + \beta_6(age) + \varepsilon \end{aligned} \quad (1)$$

The regression analysis was conducted using the multiple linear regression function in the statistical software IBM SPSS (version 29.0). The normal distribution of the residuals was visually confirmed. Homoscedasticity (Breusch-Pagan test p = .892) was present, and no hint of multicollinearity were found (all variance inflation factors were <1.5). The significance level was set to α =.05.

3. Results

The descriptive results of the false agreement rate are shown in Table 1.

Table 1. Descriptive statistics for false agreement rate for different groups (range 0 to 1).

	Age[a]		Profession		Training[b]		Gender	
	Young	**Old**	**Nurse**	**Physician**	**No**	**Yes**	**Male**	**Female**
Mean	0.376	0.399	0.394	0.375	0.458	0.291	0.444	0.343
SD	0.248	0.236	0.249	0.231	0.227	0.230	0.232	0.242
N	71	139	133	77	120	90	90	120

Annotations: [a] young (≤ 39 years), old (≥ 40 years), [b] certified wound care training

The regression analysis revealed a variance explanation of 23.8% (n = 210, R = 0.238, Adj. R^2 = 0.215) with a significant F-statistic for the model (F (6;203) = 10.552, p < .001). All regression results are shown in Table 2.

The results also revealed several significant associations between the examined factors and the agreement rate with false AI recommendations. A higher *Diagnostic Performance* of wound macerations was significantly associated with a lower false agreement rate (p = .002). Similarly, individuals with *Certified Wound Care Training* demonstrated a significantly lower false agreement rate (p < .001). Additionally, the *Profession* physician compared to nurses agreed less often with false recommendations

(p = .015). Female *Gender*, compared to male, exhibited a significantly lower rate of agreeing with false AI results (p = .025). Furthermore, the *Perceived Benefit* was associated with a higher false agreement rate (p = .007). No significant influence was found for *Age* (p = .991).

Table 2. Results of the multiple linear regression model. Dependent variable: agreement rate with false AI recommendations. Legend: SEM standard error of means, CI confidence interval.

Coefficient	b	SEM	β	T	p	95% CI	
						lower	upper
Constant	0.600	0.105		5.711	<0.001	0.393	0.807
Diagnostic Performance	-0.247	0.080	-0.201	-3.070	0.002	-0.406	-0.088
Perceived Benefit	0.025	0.009	0.176	2.720	0.007	0.007	0.043
Certified Wound Care Training	-0.125	0.036	-0.257	-3.497	<0.001	-0.196	-0.055
Profession	-0.091	0.037	-0.181	-2.442	0.015	-0.164	-0.017
Gender	-0.079	0.035	-0.162	-2.255	0.025	-0.149	-0.010
Age	<0.001	0.001	0.001	0.011	0.991	-0.003	0.003

4. Discussion

The results indicate that various factors influence automation bias, expressed as the rate of agreement with false AI-supported recommendations in the diagnosis of macerations in chronic wounds. It could be demonstrated that higher diagnostic performance, i.e. the ability to correctly identify macerations, was significantly associated with lower agreement with false AI recommendations. This implies that enhanced diagnostic performance could mitigate susceptibility to automation bias, complementing prior research identifying experience as factor in safeguarding against the acceptance of incorrect recommendations [1]. This result can be seen as a confirmation of the importance of thorough training and regular review of the diagnostic skills. Furthermore, it was found that individuals with specific certified wound care training agreed less often with false AI recommendations. This suggests that expertise and experience in wound care may help reduce the potential impact of automation bias. This finding strongly argues against the risk of de-professionalization when using CDSS properly. On the contrary, it implies that CDSS usage requires professional expertise. Similar findings have been reported in other studies, indicating that experts are less susceptible to automation bias compared to non-experts [3]. These two findings summarize that domain skills and competencies safeguard clinicians from trusting false AI-enabled recommendations.

Physicians outperformed nurses by exhibiting lower agreement with false AI recommendations. This could suggest that physicians may be better equipped to critically question the recommendations or to be less prone to doubt their own diagnostic abilities, ultimately leading to reduced susceptibility to automation bias. Gender differences also showed an influence on agreement with false AI recommendations, with women demonstrating lower susceptibility to automation bias compared to men. Gender differences were not expected and could indicate different approaches to diagnostic decisions or a more critical attitude towards AI-enabled recommendations among

women. Regarding age, no significant influence on agreement with false AI recommendations was found. This is consistent with some previous studies that showed mixed results regarding the relationship between age and automation bias [4]. Perceived benefit of the AI-supported system showed a significance for a higher agreement rate with false recommendations. Supporting this finding, a previous study had also demonstrated higher perceived benefit to positively influence trust in AI-based systems [9]. This can be seen as a hint that trust in the effectiveness, efficiency and usefulness of the system may lead to increased susceptibility to automation bias. These findings are an important aspect to consider when implementing AI-supported systems in clinical settings to avoid excessive reliance and blindness to potential errors.

This study comes along with some limitations that have to be considered when interpreting the results. The investigation took place in a simulated environment and the assessment of wound maceration was based on images only. Furthermore, the use case reflected only a single task which, however, had the advantage of assuming the same level of task difficulty despite variations in image quality. Nevertheless, it cannot be ruled out that factors not accounted for in this study could also play a role in the occurrence of automation bias. This assumption is indicated by the moderate percentage of explained variance. Beyond these concerns, future studies should clarify the impact of explainable AI (XAI) on automation bias and the role of related user training how to interpret the XAI results.

In addition to the literature, which had identified trust in the system, self-confidence, and task difficulty to lead to automation bias [1,2] our findings complement the picture by adding diagnostic skills as an inhibitor to automation bias. This contrasts the findings that both experts and non-experts can be affected [3]. We did not study trust directly, however, perceived benefit of the system could be interpreted as a proxy for trust potentially leading to automation bias as indicated in this study. Our findings also confirm that there was no age effect. All in all, the findings from the present study add a new perspective with practical implications.

5. Conclusion

Considering factors influencing automation bias when introducing a CDSS is important to fully leverage the benefits of such a system. The findings of this study are of practical interest because they reveal that the group of professionals, i.e. non-specialists, who could most profit from CDSS embraces those who were most prone to rely on the CDSS even if its recommendations were false. Furthermore, diagnostic skills and competencies combined with special training could mitigate the susceptibility to automation bias. These results underscore the notion that CDSS cannot replace prior intensive training of diagnostic skills leading to critical thinking and potentially also to self-confidence. Physicians exhibited a lower susceptibility to automation bias compared to nurses, suggesting a more critical attitude towards AI recommendations. This propensity may also resonate with women, who demonstrated a decreased inclination to accept false suggestions. Furthermore, it is plausible that physicians view AI as subordinate to their expertise. Although the aim of developing AI-based CDSS should be to minimize prediction errors they cannot be preempted completely. Hence, considering automation bias is crucial for the CDSS implementation of strategies aimed at ensuring diagnostic accuracy and patient safety. These findings therefore call for training to develop

appropriate skills including plausibility checks to prevent automation bias when utilizing CDSS in patient care.

Declarations

Conflict of interest: The authors declare that there is no conflict of interest.

Ethics: Ethical approval of this study was obtained from the Ethics Committee of Osnabrück University of Applied Sciences (no. HSOS/2023/1/6).

Authors contributions: FK, MP, NH, and JK developed the concept and design of the study. Data collection was predominantly conducted by FK. The images for the study were provided by DB, MM. Statistical analysis was performed by UH and FK. The manuscript was primarily written by FK and UH. The review and revision of the manuscript were conducted by BB, JD, UH, and CE. Supervision was provided by DB and UH.

Acknowledgements: This study was funded by the German Federal Ministry of Education and Research (BMBF) (grant: 16SV8616).

References

[1] Goddard K, Roudsari A, and Wyatt JC, Automation bias: a systematic review of frequency, effect mediators, and mitigators, J. Am. Med. Inform. Assoc. 19 (2012) 121–127. doi:10.1136/amiajnl-2011-000089.

[2] Goddard K, Roudsari A, and Wyatt JC, Automation bias: Empirical results assessing influencing factors, Int. J. Med. Inf. 83 (2014) 368–375. doi:10.1016/j.ijmedinf.2014.01.001.

[3] Bond RR, Novotny T, Andrsova I, Koc L, Sisakova M, Finlay D, Guldenring D, McLaughlinc J, Peace A, McGilligan V, Leslie SJ, Wang H, and Malik M, Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms, J. Electrocardiol. 51 (2018) S6–S11. doi:10.1016/j.jelectrocard.2018.08.007.

[4] Lyell D, and Coiera E, Automation bias and verification complexity: a systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 423–431. doi:10.1093/jamia/ocw105.

[5] Bailey NR, Scerbo MW, Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust, Theor. Issues Ergonom. Sci. 8 (2007) 321–348. doi:10.1080/14639220500535301.

[6] Lyell D, Magrabi F, Raban MZ, Pont LG, Baysari MT, Day RO, and Coiera E, Automation bias in electronic prescribing, BMC Med. Inform. Decis. Mak. 17 (2017) 28. doi:10.1186/s12911-017-0425-5.

[7] Hüsers J, Hafer G, Heggemann J, Wiemeyer S, Przysucha M, Dissemond J, Moelleken M, Erfurt-Berge C, and Hübner U, Automatic classification of diabetic foot ulcer Images – A transfer-learning approach to detect wound maceration, in: Mantas J, Hasman A, Househ MS, Gallos P, Zoulias E, and Liaskos J (Eds.), Stud. Health Technol. Inform., IOS Press, 2022. doi:10.3233/SHTI210919.

[8] Venkatesh V, and Davis FD, A Theoretical extension of the technology acceptance model: Four longitudinal field studies, Manag. Sci. 46 (2000) 186–204. doi:10.1287/mnsc.46.2.186.11926.

[9] Babitsch B, Hannemann N, Kutza JO, and Hübner U, Trust in digitalization and AI: Findings from a qualitative study on healthcare professionals in Germany. Stud. Health Technol. Inform., IOS Press, 2023. doi:10.3233/SHTI230810.