Tell Me or Show Me? Evaluating Diagnostic Support for Electrocardiogram Interpretation

Tobias Grundgeiger
Julius-Maximilians-University
Würzburg, Germany
tobias.grundgeiger@uni-wuerzburg.de

Nicole Dering
Julius-Maximilians-University
Würzburg, Germany
nicole.dering@stud-mail.uni-wuerzburg.de

Julia Wienkop
Julius-Maximilians-University
Würzburg, Germany
julia.wienkop@stud-mail.uni-wuerzburg.de

Lena Kutter
Julius-Maximilians-University
Würzburg, Germany
lena.kutter@stud-mail.uni-wuerzburg.de

Sabrina Baum
Julius-Maximilians-University
Würzburg, Germany
sabrina.baum@stud-mail.uni-wuerzburg.de

Leo Schwarzkopf
Julius-Maximilians-University
Würzburg, Germany
leo.schwarzkopf@stud-mail.uni-wuerzburg.de

Carlos Hölzing
University Hospital Würzburg
Würzburg, Germany
hoelzing_c@ukw.de

Oliver Happel
University Hospital Würzburg
Würzburg, Germany
happel_o@ukw.de

ABSTRACT

The interpretation of electrocardiograms (ECGs) is a difficult and error-prone task that can be supported by computer algorithms. In this exploratory study, we investigated diagnostic support in the form of a text-based full diagnosis (ECG diagnosis condition), ECGs with marked segments that are relevant for the diagnosis (ECG marking condition), or only the ECGs as a baseline (ECG only condition). Support improved diagnosis accuracy compared to the ECG only condition, and the ECG diagnosis condition resulted in the highest accuracy. Support negatively affected the feeling of autonomy compared to ECG only condition, with the largest effect in the ECG marking condition. Finally, most participants preferred the ECG diagnosis condition. We discuss possible explanations for the in-part contradicting result patterns of accuracy, psychological need satisfaction, and preference and suggest avenues for future research.

CCS CONCEPTS

• Information systems~Information systems applications~Decision support systems • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI • Applied computing~Life and medical sciences~Healthcare information systems

KEYWORDS

Eye tracking, diagnosis, decision support, artificial intelligence, electrocardiography

ACM Reference format:

Tobias Grundgeiger, Nicole Dering, Julia Wienkop, Lena Kutter, Sabrina Baum, Leo Schwarzkopf, Carlos Hölzing, & Oliver Happel. 2025. Tell Me or Show Me: Evaluating Diagnostic Support for Electrocardiogram Interpretation. CHI 2025 Workshop Envisioning the Future of Interactive Health. Poster contribution, *4 pages.*

1 Introduction

The interpretation of electrocardiograms (ECGs) is a difficult and error-prone task [1, 11]. Research demonstrated that computer algorithms can provide accurate ECG diagnosis interpretation; however, these advances are limited to specific abnormalities in ECG recordings [10] and our clinical collaborator reported that the recommendations in existing products do not provide reliable support. Finally, the clinician may accept or dispute the support of computer algorithms, and as a result, even a perfect computer algorithm may not result in a correct ECG diagnosis [2, 3, 9].

In this exploratory study, we investigated the effect of providing diagnostic support for ECG interpretation. For diagnostic support, we provided ECGs with a text-based full diagnosis (ECG diagnosis condition), ECGs with marked segments that are relevant for the diagnosis (ECG marking condition), or only the ECGs as a baseline (ECG only condition). To evaluate the effect, we measured (1) ECG diagnosis accuracy in percentage, (2) dwell time percentages on segments of the ECG that were relevant for the diagnosis, (3) psychological need satisfaction of autonomy and competence, and (4) subjective preference for the three different ECG presentations.

2 Method

2.1 Participants

A total of 23 physicians (13 females and 10 males; average age 31 years, SD=6 years; average work experience 3.9 years, SD=4.2 years) from the University Hospital Würzburg, Germany, participated in this study. The eye-tracking data of five participants were excluded because the tracking ratio was below 60%. Nevertheless, all responses regarding ECG diagnosis, autonomy, competence, and preference were included in the final analysis. All participants had either normal vision or vision corrected by contact lenses. The study was approved by the institutional ethics committee, and written consent was obtained from each participant.

2.3 Material and Procedure

The ECGs were selected in advance by an experienced anesthesiologist (author OH) in order to have a similarly difficulty level. All ECGs were selected from the teaching website ECG Made Simple (https://ecgmadesimple.ca). To emulate the diagnostic proposal of an AI for the condition ECG with diagnosis, we adapted the ECG diagnostic proposal of the ECG device at the University Hospital Würzburg, Germany. For the condition ECG with marking, the color selection of the markers was based on the color scheme suggested by the anesthesiologist (author OH). We prepared a total of nine ECGs. Each ECG was equally often shown in each of the three conditions. We could not fully counterbalance the ECGs and the order of the conditions. We decided to use each ECG equally often in the three conditions but did not fully counterbalance the order of the conditions.

In the study, we explained the study to participants, and participants provided consent. We asked participants to diagnose ECGs based on various ECG graphs. In some conditions, the process would be supported by either marked ECG segments or a written diagnosis (Figure 1). Participants were told that the support was generated by a very good but not perfect algorithm in both conditions. The participants were seated in front of the eye tracker (SMI Eye Tracker mRED with 120 Hz) in an office room at the university hospital. In order to maximize eye tracking quality, the lights were left on for all participants, and the eye tracker was always placed in the same position. After the participants were seated, they were advised to move as little as possible, keep their eyes open and straight forward, and avoid head movement. Before each condition, we calibrated the eye tracker. The calibration was carried out by the experimenter and repeated until the accuracy of every subject was below 1°.

Once the calibration was done, the first condition started. Participants viewed and diagnosed the three ECGs of the first condition. There was no time limit while performing the task. After the three ECGs, participants rated their experienced autonomy (two questions) and competence (two questions) on a five-point Likert scale (1="strongly disagree" and 5="strongly agree") [5]. This procedure was repeated for each condition (within-subject design). At the end, we collected demographic data and data on the preferred condition.

2.4 Analysis

For ECG diagnosis accuracy, an experienced anesthesiologist (author OH) analyzed the answers and marked 0, 0.5., or 1 point for wrong, partly correct but not fully specified answers, and correct answers, respectively. The anesthesiologist was blinded to the conditions while analyzing the answers. Using the SMI Experiment Center, areas of interest (AOIs) for the ECGs (i.e., the larger area around the segments

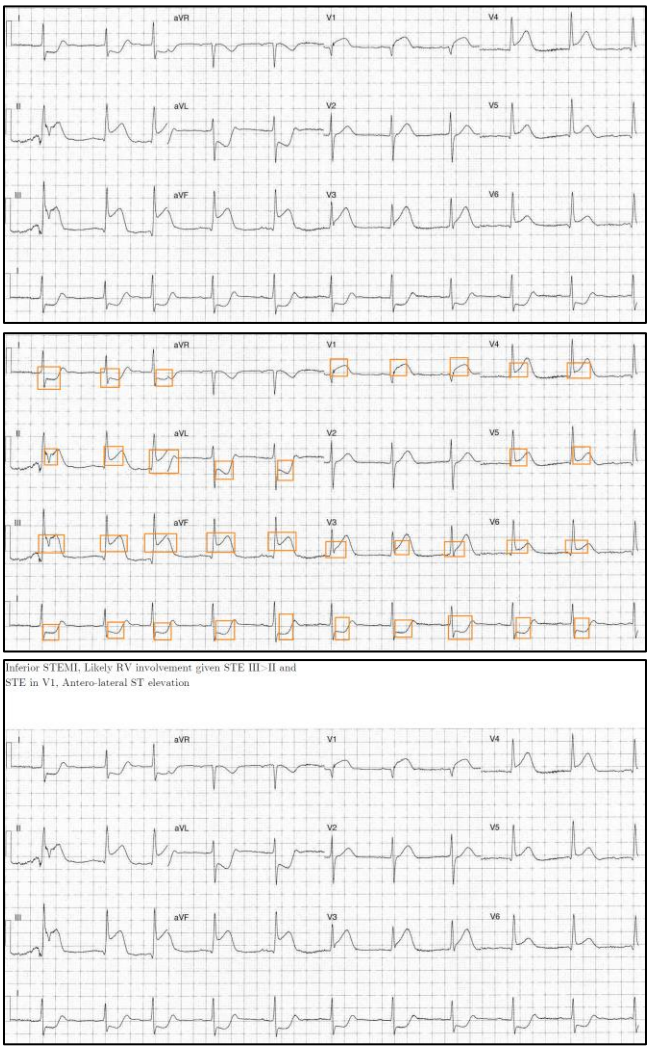

Figure 1: Demonstration of the electrocardiogram (ECG) layout for three conditions (top) ECG only, (middle) ECG marking, and (bottom) ECG diagnosis. Note that these illustrations have been created of illustration purposes, only, and were not used in the study. The ECG was taken from https://jhcedecg.blogspot.com/ (CC BY-NC-SA 4.0).

that were relevant for the diagnosis, which were also highlighted in the ECG marking condition) were selected. Subsequently, the viewing duration for both AOIs and whitespace (i.e., the rest of the ECG image that was not part of the AOIs) was analyzed, and we calculated the percentage dwell time in the AOI for each participant in each condition. Autonomy and competence were analyzed by averaging the five-point Likert scale answers for each psychological need.

3 Results

3.1 ECG Diagnosis Accuracy

A repeated measure ANOVA showed a significant main effect, $F(2,40)=5.129$, $p=.010$, $\eta p^2=0.204$. Bonferroni-corrected post hoc test indicated no significant differences between the ECG only - ECG marking ($M_{diff}=-9.5\%$, $p=.678$, $d=-0.383$), the ECG marking - ECG diagnosis condition ($M_{diff}=15.1\%$, $p=.176$, $d=-0.606$) but a significant difference between the ECG only - ECG diagnosis condition ($M_{diff}=-24.6$, $p=.009$, $d=0.989$). Note that N was only 21 because data of two participants were missing due to technical errors (see Tabel 1 for means and standard deviations).

Table 1: Dependent variables separated by condition. Values indicate mean (standard deviation) or preference in percentages.

	ECG only	ECG marking	ECG diagnosis
ECG Diagnosis Accuracy	56% (26)	66% (26)	80% (23)
Percentage Dwell Times	31% (15)	50% (15)	32% (12)
Autonomy	3.48 (1.07)	2.89 (1.23)	3.20 (0.90)
Competence	3.13 (0.57)	2.84 (0.66)	2.94 (0.59)
Preference	17.4% (N=4)	21.7% (N=5)	60.9% (N=14)

3.2 Dwell Times

A repeated measure ANOVA showed a significant main effect of ECG presentation on dwell times of diagnosis-relevant AOIs, $F(2,34)=11.35$, $p<.001$, $\eta p^2=0.400$. Bonferroni-corrected post hoc test indicated significant differences between the ECG only - ECG marking condition ($M_{diff}=-19\%$, $p<.001$, $d=-1.381$) and the ECG marking - ECG diagnosis condition ($M_{diff}=18\%$, $p=.001$, $d=1.279$) but not between the ECG only - ECG diagnosis condition ($M_{diff}=-1\%$, $p=1$, $d=-0.101$).

3.3 Autonomy and Competence

For autonomy, a repeated measure ANOVA showed a significant main effect, $F(2,44)=1.982$, $p=.010$, $\eta p^2=0.189$. Bonferroni-corrected post hoc test indicated a significant difference between the ECG only - ECG marking condition ($M_{diff}=0.587$, $p=.008$, $d=0.566$) but no significant differences between the ECG only - ECG diagnosis condition ($M_{diff}=0283$, $p=.392$, $d=0.272$), and the ECG marking - ECG diagnosis condition ($M_{diff}=0.304$, $p=.313$, $d=0.293$).

For competence, a repeated measure ANOVA showed no significant main effect, $F(2,44)=2.63$, $p=.116$, $\eta p^2=0.093$. Bonferroni-corrected post hoc test indicated no significant differences between the ECG only - ECG marking condition ($M_{diff}=0.196$, $p=.473$, $d=0.321$), the ECG only - ECG diagnosis condition ($M_{diff}=0.283$, $p=.131$, $d=0.464$), and the ECG marking - ECG diagnosis condition ($M_{diff}=0.087$, $p=1$, $d=0.143$).

3.4 Preference

Out of 23 participants, 60.9% (N=14) preferred the ECG diagnosis condition, 21.7% (N=5) preferred the ECG only condition, and 17.4% (N=4) preferred the ECG marking condition. In the qualitative feedback, many participants indicated that the marked segments were distracting and hard to ignore. As a result, they felt biased towards these segments or had problems following their regular patterns of interpreting an ECG. The participants who preferred the ECG marking mentioned that the marking enabled a fast and independent assessment (compared to the ECG diagnosis condition). In relation to support, many participants appreciated the support in form of having a suggested diagnosis that they could confirmed, and some participants indicated that they conducted their own interpretation and only then considered the suggestion.

4 Discussion

In this exploratory study, we investigated the effect of providing diagnostic support for ECG interpretation. Diagnostic support significantly increased ECG diagnostic accuracy, in particular in the ECG diagnosis condition. However, despite providing 100% correct answers in the ECG diagnosis condition, the average accuracy was only 80%. Simply providing the “correct diagnosis” did not always result in choosing this diagnosis. One explanation might be that the anesthesiologists were not always able to find all relevant segments in the ECG and therefore did not consider the suggested (correct) diagnosis as valid. If this was the case, combining the marking and the written diagnosis might improve accuracy further.

In relation to the dwell times, we observed that the ECG marking condition resulted in the longest dwells on the segments that are relevant for interpretation. One may think, that longer dwells on relevant AOIs may foster better interpretation. This was not the case in our study. Others

observed no correlation between the total time spend on interpretation an accuracy [1, 11]. We also observed no correlation between dwell time on the relevant AOI and diagnosis accuracy in any condition.

The decreased feeling of autonomy with diagnosis support might be explained by the idea that automation may take away the feeling of decision-making autonomy from users [4, 7, 8, 12]. The highlighting of specific ECG segments might have resulted in an even bigger feeling of being other-directed compared to the stated diagnosis. Similarly, the descriptive differences in relation to competence indicate that the feeling of competence was highest with no support (ECG only) and similarly lower for both support conditions.

Contrary to the findings in relation to psychological need satisfaction, the anesthesiologists preferred support (>80%), and most anesthesiologists preferred the ECG with diagnosis. This contradiction may be explained by an increased feeling of safety in case of the support conditions that may have outweighed the reduced feeling of autonomy and competence in the final preference rating. Future research should consider safety as a further psychological need.

When taking the accuracy, need satisfaction, preference ratings and the qualitative feedback together, one may also provide support on request, only. Support on request would allow physicians to use their regular interpreting strategies in an unbiased manner and confirm or challenge their interpretation. Support on request might be a possibility to consider need satisfaction but still provide support.

The study has limitations. First, as part of the exploratory nature of the study, we did not run a fully counterbalanced design in relation to ECG distribution in the different conditions and the order of the conditions. Second, the ECG graph paper corresponded to the American standard, but we had German anesthesiologists as participants. Third, we provided no additional information but only the ECGs. For example, patient history and patient presentation may provide further important information for an ECG diagnosis. Fourth, it has been noted that the scales for autonomy and competence address life in general and need to be adapted to the specificities of the healthcare context [6]. Fifth, diagnosis support was always correct, but no algorithm will be 100% correct. Future research should consider the effect of reduced reliability on ECG diagnosis.

In conclusion, providing ECG diagnosis support increased ECG interpretation accuracy. Further research is needed to understand the best presentation of the diagnosis support to further increase diagnosis accuracy while maintaining a positive user experience (i.e., psychological needs satisfaction) rather than reducing user experience.

ACKNOWLEDGMENTS

We thank the medica staff of the University Hospital Würzburg for participating in the study.

REFERENCES

[1] R. R. Bond, T. Zhu, D. D. Finlay, B. Drew, P. D. Kligfield, D. Guldenring, . . . G. D. Clifford. 2014. Assessing computerized eye tracking technology for gaining insight into expert interpretation of the 12-lead electrocardiogram: an objective quantitative approach. *Journal of Electrocardiology, 47*(6), 895-906. https://doi.org/https://doi.org/10.1016/j.jelectrocard.2014.07.011

[2] Andrew W. Cairns, Raymond R. Bond, Dewar D. Finlay, Cathal Breen, Daniel Guldenring, Robert Gaffney, . . . Pat Henn. 2016. A computer-human interaction model to improve the diagnostic accuracy and clinical decision-making during 12-lead electrocardiogram interpretation. *Journal of Biomedical Informatics, 64*, 93-107. https://doi.org/https://doi.org/10.1016/j.jbi.2016.09.016

[3] Andrew W. Cairns, Raymond R. Bond, Dewar D. Finlay, Daniel Guldenring, Fabio Badilini, Guido Libretti, . . . Stephen J. Leslie. 2017. A decision support system and rule-based algorithm to augment the human interpretation of the 12-lead electrocardiogram. *Journal of Electrocardiology, 50*(6), 781-786. https://doi.org/https://doi.org/10.1016/j.jelectrocard.2017.08.007

[4] Tobias Grundgeiger, Jörn Hurtienne, and Oliver Happel. 2021. Why and how to approach user experience in safety-critical domains: The example of healthcare. *Human Factors, 63*(5), 821–832. https://doi.org/10.1177/0018720819887575

[5] Marc Hassenzahl, Sarah Diefenbach, and Anja Göritz. 2010. Needs, affect, and interactive products–Facets of user experience. *Interacting with Computers, 22*(5), 353-362. https://doi.org/10.1016/j.intcom.2010.04.002

[6] Anna Hohm, Oliver Happel, Jörn Hurtienne, and Tobias Grundgeiger. 2022. User experience in safety–critical domains: a survey on motivational orientations and psychological need satisfaction in acute care. *Cognition, Technology & Work*(24), 247–260. https://doi.org/10.1007/s10111-022-00697-0

[7] Anna Hohm, Oliver Happel, Jörn Hurtienne, and Tobias Grundgeiger. 2023. *"When the beeping stops, you completely freak out" - How acute care teams experience and use technology.* In the Proceedings of the ACM on Human-Computer Interaction 6.CSCW2 (2022): 1-32, Minneapolis, MN USA. https://doi.org/10.1145/3579590

[8] Sara Klüber, Franzisca Maas, David Schraudt, Gina Hermann, Oliver Happel, and Tobias Grundgeiger. 2020. *Experience matters: Design and evaluation of an anesthesia support tool guided by user experience theory.* In the Proceedings of the ACM Conference on Designing Interactive Systems (DIS), Eindhoven, Netherlands. https://doi.org/10.1145/3357236.3395552

[9] Tomas Novotny, Raymond Bond, Irena Andrsova, Lumir Koc, Martina Sisakova, Dewar Finlay, . . . Marek Malik. 2017. The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. *International Journal of Medical Informatics, 101*, 85-92. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2017.02.007

[10] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, . . . Antonio Luiz P. Ribeiro. 2020. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications, 11*(1), 1760. https://doi.org/10.1038/s41467-020-15432-4

[11] Mohammed Tahri Sqalli, Dena Al-Thani, Mohamed B Elshazly, and Mohammed Al-Hijji. 2021. Interpretation of a 12-Lead Electrocardiogram by Medical Students: Quantitative Eye-Tracking Approach. *JMIR Med Educ, 7*(4), e26675. https://doi.org/10.2196/26675

[12] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. *Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes.* In the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Glasgow, Scotland Uk. https://doi.org/10.1145/3290605.3300468