Fanima! Pervasive Serious Game for Phonetic-Phonological Assessment of Children Towards Autonomous Speech Therapy

Inês Antunes[1], André Antunes[1,2], and Rui Neves Madeira[1,2(✉)]

[1] NOVA LINCS, NOVA School of Science and Technology, NOVA University of Lisbon, Caparica, Portugal
ig.antunes@campus.fct.unl.pt

[2] Sustain.RD, Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal, Setúbal, Portugal
{andre.antunes,rui.madeira}@estsetubal.ips.pt

Abstract. Many children have difficulties with speech and language, sometimes even both. Speech therapy for children is usually a tedious process where one of the drawbacks of traditional solutions is the repetition of exercises. A known strategy to increase engagement is to deliver speech therapy through mobile games. A serious game-based diagnosis tool for phonetic-phonological assessment of speech disorders focused on the European Portuguese language was designed with therapists' participation. The integration with a web platform allows real-time therapist interaction to control the game based on the classification of vocalisations made by the child in response to the gameplay, which follows the therapeutic structure for the intended diagnosis. One first user study was made with five speech therapists to validate the tool's concept and the system's usability for responding to the therapeutic requirements. The results are positive, validating the tool and suggesting its acceptance by the community of therapists, allowing us to move on to a thorough second study with therapists in a therapeutic context with children.

Keywords: Serious Games · Speech Therapy · Phonetic-Phonological Assessment · Mobile Computing · Interaction Design and Children

1 Introduction

Children at a very young age learn to use devices like smartphones, controllers, or tablets. Many children feel excited about using mobile devices [20]. The growing popularity of smart mobile devices among young children, driven by their unique characteristics and the rapid development of age-appropriate applications, has been highlighted in previous research, emphasizing these devices as the preferred

This work was supported by NOVA.ID.FCT/NOVA LINCS (UIDB/04516/2020).

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 201–220, 2024.
https://doi.org/10.1007/978-3-031-59717-6_14

11. van Greuningen, M., Borgs, B.: Feiten en cijfers over mensen met een ernstige psychiatrische aandoening (2022). https://www.vektis.nl/intelligence/publicaties/factsheet-ernstige-psychiatrische-aandoeningen
12. Hamdoun, S., Monteleone, R., Bookman, T., Michael, K.: AI-based and digital mental health apps: balancing need and risk. IEEE Technol. Soc. Mag. **42**(1), 25–36 (2023)
13. Jameel, L., Valmaggia, L., Barnes, G., Cella, M.: mHealth technology to assess, monitor and treat daily functioning difficulties in people with severe mental illness: a systematic review. J. Psychiatric Res. **145**(2021), 35–49 (2022). https://doi.org/10.1016/j.jpsychires.2021.11.033
14. James, L.J., et al.: Evaluation of personalized treatment goals on engagement of smi patients with an mhealth app. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1568–1573. IEEE (2022)
15. Litvin, S., Saunders, R., Maier, M.A., Lüttke, S.: Gamification as an approach to improve resilience and reduce attrition in mobile mental health interventions: a randomized controlled trial. PLoS ONE **15**(9), e0237220 (2020)
16. Liu, B., et al.: Adversarial attacks on large language model-based system and mitigating strategies: a case study on ChatGPT. Secur. Commun. Netw. **2023** (2023)
17. Locke, E.A., Latham, G.P.: Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. Am. Psychol. **57**(9), 705 (2002)
18. Michie, S., et al.: The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann. Behav. Med. **46**(1), 81–95 (2013)
19. Organization, W.H., et al.: World mental health report: transforming mental health for all (2022)
20. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. **55**(1), 68 (2000)
21. Shatte, A.B., Hutchinson, D.M., Teague, S.J.: Machine learning in mental health: a scoping review of methods and applications. Psychol. Med. **49**(9), 1426–1448 (2019)
22. Svensson, B., Hansson, L., Markström, U., Lexén, A.: What matters when implementing flexible assertive community treatment in a Swedish healthcare context: a two-year implementation study. Int. J. Ment. Health **46**(4), 284–298 (2017)
23. Van Gorp, P., Nuijten, R.: 8-year evaluation of GameBus: status quo in aiming for an open access platform to prototype and test digital health apps. Proc. ACM Hum.-Comput. Interact. **7**(EICS), 1–24 (2023)
24. Van Veldhuizen, J.R.: FACT: a Dutch version of ACT. Community Ment. Health J. **43**(4), 421–433 (2007). https://doi.org/10.1007/s10597-007-9089-4

treatment plans. The study suggests that it is feasible to develop a system that utilizes LLMs to allow users to generate measurable treatment plan goals. Using the prototype to create measurable goals, does make it possible to directly add meaningful goals to gamified mHealth systems like GameBus, which could potentially intrinsically motivate patients further to work on their treatment goals. Although the quality of the generated goals was not assessed by healthcare professionals or patients with SMI, the evaluation process with students indicated that incremental improvements in behavior goals were attainable. Furthermore, the study revealed that the application of SDT and Goal-Setting theory could enhance the quality of behavioral goals. However, improving a certain aspect and maintaining a consistent appreciation of all other aspects can be challenging. Now that it has been established that the proposed workflow has the potential to improve the process of case managers creating goals with their patients with SMI. We can now evaluate the system within the treatment process, and gather patient and case manager feedback. It is important to acknowledge that the results of this research should be considered preliminary, as there are currently no comparable findings in the existing literature for appropriate comparisons. These preliminary findings provide a foundation for further investigation and highlight the need for future research in this area.

References

1. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus **15**(2) (2023)
2. Arora, A., Arora, A.: The promise of large language models in health care. Lancet **401**(10377), 641 (2023)
3. Bovend'Eerdt, T.J., Botell, R.E., Wade, D.T.: Writing smart rehabilitation goals and achieving goal attainment scaling: a practical guide. Clin. Rehabil. **23**(4), 352–361 (2009)
4. Chase, J.A., Houmanfar, R., Hayes, S.C., Ward, T.A., Vilardaga, J.P., Follette, V.: Values are not just goals: online act-based values training adds to goal setting in improving undergraduate college student performance. J. Contextual Behav. Sci. **2**(3–4), 79–84 (2013)
5. Cheng, V.W.S., Davenport, T., Johnson, D., Vella, K., Hickie, I.B.: Gamification in apps and technologies for improving mental health and well-being: systematic review. JMIR Ment. Health **6**(6), e13717 (2019)
6. Dale, R.: GPT-3: what's it good for? Nat. Lang. Eng. **27**(1), 113–118 (2021)
7. Eckerstorfer, L.V., et al.: Key elements of mhealth interventions to successfully increase physical activity: meta-regression. JMIR Mhealth Uhealth **6**(11), e10076 (2018)
8. Fogg, B.J.: A behavior model for persuasive design. In: Proceedings of the 4th international Conference on Persuasive Technology, pp. 1–7 (2009)
9. Gabbard, G.O., Crisp-Han, H.: The early career psychiatrist and the psychotherapeutic identity. Acad. Psychiatry **41**, 30–34 (2017)
10. Graham, S., et al.: Artificial intelligence for mental health and mental illnesses: an overview. Curr. Psychiatry Rep. **21**, 1–18 (2019)

by the researchers. Upon analysis of the generated behavioral goals resulting from the adversarial attacks performed. We did not experience any hallucinations, however, more extensive adversarial attacks need to be conducted in order to assess the potential harm hallucinations may cause. Therefore, the evaluation of goals by the case manager is crucial before adding a goal to the treatment plan. In the exploratory adversarial attacks, we only tested the prototype with GPT-3 as its LLM, if another LLM were to be used it is unknown what the responses and how safe the generated goals would be. More elaborate and in-depth adversarial attacks need to be done in further research, to assess the safety of using LLMs to generate treatment plan goals. Lastly, this feasibility study specifically focused on setting goals for a specific target group, namely Dutch patients with SMI who are prescribed FACT treatment. Therefore, conclusions drawn from this study cannot be generalized to other (non) SMI groups without further research.

Future Work

In future work, we should recruit case managers, and patients with SMI to evaluate the goal-setting workflow and the quality of goals, as the results from healthy participants cannot be generalized to people with SMI. It is worth evaluating and tracking the progress patients with SMI are making toward the generated goal, through an mHealth application like GameBus, compared to non-LLM generated goals. The integration of SDT and Goal-setting has proven valuable in this research, but including other behavior theories, such as the COM-B, could enhance the assessment. COM-B is a valuable tool to address instances where individuals lack the necessary preconditions for certain behaviors. For example, if the generated goal says to 'ride a bicycle' it becomes an unfeasible goal if the patient does not possess a bicycle. Incorporating the COM-B model into an interactive goal-setting approach can address these situations more effectively and take the prerequisites needed to complete a goal into consideration. Mental healthcare professionals could also be given a hands-on session with the system in a usability study. An addition to the user interface that could potentially make it easier for case managers and patients to select treatment plan goals that are relevant to the treatment is to create an interface that allows users to easily swipe irrelevant goals away, and save relevant ones. Another interesting direction to explore is to investigate the willingness and privacy concerns of patients with SMI regarding using such technology in their treatment. Given the current level of distrust among patients with SMI, understanding their attitudes and perceptions toward AI-based tools and data privacy implications would provide valuable insights. Such research is essential for ensuring the ethical and effective integration of such technology in the treatment journey of patients with SMI.

6 Conclusion

This feasibility study aimed to explore the use of LLMs in enabling patients with SMI and case managers to easily create measurable treatment goals for

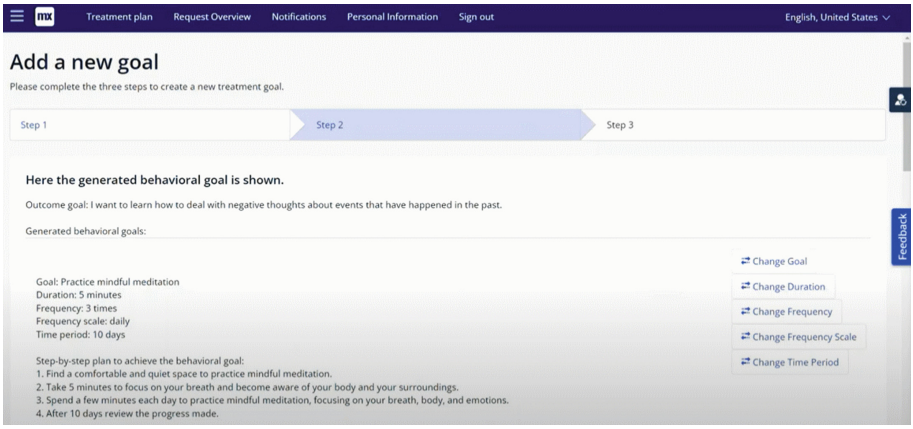


Fig. 2. Screenshot of the working prototype, showing one of the generated treatment plan goals and the option to edit the goal and its attributes.

score, autonomy, competence, relatedness, clarity, and feedback increased. Average total score: 33.063 out of 35 points. This is an increase of 4.5 compared to the previous evaluation phase.

5 Discussion

Study Limitations

In this study, it is important to note that the evaluation of the prototype itself fell outside the scope of the study. Consequently, aspects related to user-friendliness, UI, and UX elements were also outside the scope of the study. It is also worth highlighting that no testing was conducted with case managers or patients with SMI. Our primary focus was centered on determining whether modifications to the prototype and prompts could lead to improvements in the quality of the generated goals. In the evaluation phase, only the behavioral goals of two outcome goals were evaluated by students, the quality of the generated goals may differ depending on how outcome goals are structured and the type of outcome goals provided. The goals were also not evaluated by case managers on their quality, it is possible that they could consider other elements.

A major risk of using LLMs to generate goals for patients with SMI is that LLMs may hallucinate and could potentially generate goals that may not align with standard FACT treatment which may result in further complications for the patient. Accordingly, we performed some exploratory adversarial attacks on the prototype to expose potential vulnerabilities in the system when it comes to generating goals that may potentially not be in line with FACT treatment. A list of 26 treatment outcome goals that were deemed potentially harmful by the researchers, was used as input in the latest version of the prototype. The treatment behavioral goals given as output by the prototype were then inspected

and feedback significantly increased, whereas competence and clarity significantly decreased. For goal 2, there was a significant decrease in relatedness. For the average score, clarity significantly decreased. Average total score: 28.563 out of 35 points. A slight increase compared to the previous evaluation phase.

So far, the prompt has remained general without specifying the target group for whom the goals are intended. The next direction is to explore if giving the context that the generated behavioral goals are for patients with SMI would impact the response. Additionally, it would be interesting to investigate how including specific diagnoses would influence behavioral goals. This could potentially lead to more tailored results.

Evaluation Phase 4

The prompt has undergone two modifications. Firstly, by explicitly stating in the initial sentence that the treatment goals are intended for patients with a severe mental illness and emphasizing the goal's suitability for this individual. Secondly, an additional sentence has been included to specify the patient's diagnosis with an emphasis that the goal should be attainable for someone with this diagnosis.

Tests were conducted on these two scenarios, however, the behavioral goals did not show a significant change compared to the results from the previous evaluation phase. Modifying the attributes of the goal also yielded similar responses as before. As a result, this version will not undergo further surveys since the behavioral goals did not exhibit significant changes. It remains uncertain whether this new prompt would impact other goals.

In the earlier surveys, respondents expressed uncertainty about how to carry out the goal effectively and desired additional support through guidelines or a step-by-step plan that would enable them to pursue the goal effectively. As a result, addressing this issue will be the focus of the next evaluation phase.

Evaluation Phase 5

Attributes were added to the prompt to also include a step-by-step plan on how to achieve the behavioral goals, supportive tips to aid progress, and an explanation of the goal.

The behavioral goals have the same structure as in evaluation phase 4, with these three attributes added. As occurred in evaluation phase 3 the generated goals appeared to be very similar. In the case of goal 1, three separate behavioral goals were initially generated, but two were highly familiar, and therefore one was modified by the researcher using the edit function. The main goal was to improve clarity, which increased significantly in both goals. The other survey results also varied quite from the last evaluation phase. The most important changes were for goal 1. The dimensions of autonomy, competence, relatedness, feedback, and task complexity significantly increased. For goal 2, there was a significant increase in autonomy, competence, relatedness, commitment, and feedback. For the average

not understand the goals well. It is important that goals are clearly defined. Therefore, the goals in the next evaluation phase will be presented in a more structured manner, in the format of SMART goals.

Instead of goals being generated in standard text form, each goal will now include the following attributes which are based on the SMART goal structure: goal, duration, frequency, frequency scale, and time period. An example of a structured goal can be seen in Fig. 2. This structured approach also facilitates an easier transformation to JSON format.

Evaluation Phase 2

In this version of the prototype, the patient is now able to modify the goals at the attribute level. Case managers now also have the ability to evaluate the goals of their patients, by allowing them to approve, modify, or reject goals. Once a goal is evaluated and approved by a case manager, the patient will receive a notification and be able to view the approved goals in their treatment plan.

The prompt has been modified for GPT-3 to respond in a more structured manner. To ensure conciseness, it is emphasized that only a subject, verb, and object are allowed in the generated goal.

The main goal of this evaluation phase was to improve clarity, which increased significantly for goal 1 but decreased for goal 2. Overall clarity increased by 0.25. For goal 1 the dimensions of competence and commitment significantly increased, whereas the dimensions of autonomy and challenge saw a significant decrease. For goal 2, autonomy and feedback significantly decreased. For the average score, autonomy was significantly reduced, but competence significantly increased. The average total score: is 28.5 out of 35 points. This is a slight increase of 0.187 compared to the previous evaluation phase.

As autonomy is currently one of the lower-scoring aspects, this improvement area will be investigated. Providing three behavioral goals instead of one has the potential to increase autonomy. Combined with the ability to modify attributes, this can lead to greater personalization and foster a stronger sense of ownership. This decision also may indicate that a single behavioral goal as a response per outcome goal may not be sufficient to achieve the desired outcome.

Evaluation Phase 3

In this version of the prototype, goals have the same structure as in evaluation phase 2, however, patients now receive three behavioral goals per outcome goal, instead of one. Initially, for outcome goal 2, three identical behavioral goals were generated. Although, the goal attributes varied. This occurred despite configuring GPT-3 to maximize the randomness of its responses.

The main goal of this evaluation phase was to improve autonomy. This slightly increased for goal 2, but decreased for goal 1. Overall there was a slight decrease in autonomy. The other results also varied from the last round. The biggest changes were found for goal 1. The dimensions of relatedness, challenge,

4.2 Evaluating Generated Measurable Goals

The group of 8 students that was enlisted to fill in the 5 rounds of surveys to assess the behavioral goals rated the LLM-generated behavioral goals formulated for the following two outcome goals: 1) I want to learn how to deal with negative thoughts about events that have happened in the past. 2) I want to learn to discuss my fears, but I especially want to understand my fears.

In the following section, we describe the results of each evaluation phase and the overall changes that were made to the prompts and goals based on the results of the previous phase. An overview of the survey results can be seen in Table 1.

Table 1. Displays the average scores assigned by the participants to goals 1 (G1) and 2 (G2) based on how the goals scored on the elements of the SDT and Goal Setting Theory. Scores for evaluation phases 1–3 and 5 are presented, including total and combined average scores (AS) for both goals in each phase.

Evaluation phase	1			2			3			5		
	G1	G2	AS	G1	G2	AS	G1	G2	AS	G1	G2	AS
Autonomy	4.125	4	4.063	3.625	3.125	3.375	3.375	3.25	3.313	4	4.125	4.063
Competence	3.375	3.625	3.5	4.5	3.75	4.125	3.75	3.625	3.688	4.25	4.25	4.25
Relatedness	1.875	4.5	3.188	1.5	4.5	4	3.375	3.25	3.313	4	3.875	3.938
Clarity	3.625	4.125	3.875	4.5	3.75	4.125	3.375	3.875	3.625	4.625	4.5	4.563
Challenge	3.875	3.875	3.875	3.25	4.125	3.688	4.25	4.25	4.25	4.5	4.25	4.375
Commitment	3.75	3.5	3.625	4.25	3.5	3.875	4	3.125	3.563	4	4	4
Feedback	1.75	3.875	2.813	2	3.375	2.688	2.625	3.25	2.938	3.625	3.75	3.688
Task complexity	3.375	3.375	3.375	3.625	3.625	3.625	3.75	4	3.875	4.25	4.125	4.188
Total score	25.7	28.375	28.313	27.25	29.75	28.5	28.5	28.625	28.563	33.25	32.875	33.063

Evaluation Phase 1

The first version of the prototype features a user interface where patients can input their outcome goal and receive a corresponding behavioral goal. Additionally, the case manager also has the option to enter treatment goals for their assigned patients. The prompt used in this first iteration to generate a goal from GPT-3, only includes asking for one behavioral goal based on an outcome goal.

In the evaluation, based on the survey, participants felt a high level of autonomy in pursuing both behavioral goals, with an average score of just above 4. They clearly understood the behavioral goals and found them moderately challenging. Commitment and competence to the behavioral goals were strong across both goals. However, there were differences in relatedness and feedback. Through the open text field, participants claimed to have felt stronger social support in goal 2 than in goal 1, this may be because goal 1 is carried out alone whereas goal 2 involves others. Clarity and task complexity were consistent across both goals, indicating a similar understanding of the goals and tasks' complexity. The average total score: 28.313 out of 35 points.

Despite clarity receiving a relatively high average score of 3.875, there is still room for improvement. There were several comments that the participants did

the submitted goal requests, requests for goal changes, and goal edit requests by their patients. Case managers can also accept or reject these requests, save the treatment plan in the Mendix prototype, and ultimately submit the treatment plan to the GameBus application.

GameBus. The GameBus application has been extended to receive the treatment plan goals of the Mendix prototype, through the API. The Mendix prototype sends an API request to the GameBus application, sending a patient's approved treatment plans to the GameBus application in JSON format. The GameBus application converts the treatment plan it receives into GameBus challenges and saves the treatment outcome and behavioral goals within the treatment plan, in the GameBus data model. The patients using the Mendix prototype to create the treatment plan goals, also have access to the mHealth configuration of the GameBus application. The GameBus mHealth application supports multiple modular gamification and personalization options that have been designed to increase the intrinsic motivation of GameBus users. GameBus can personalize the challenges of users by tailoring the elements within them such as adjusting the frequency of the tasks needed to be completed within the challenge, adjusting the time and date a challenge should be completed or tracked, and allowing challenges to be done in groups. The GameBus application can be configured to use gamification such as giving points to users when they complete tasks within challenges, displaying a leaderboard displaying user scores, and awarding users loot boxes. The system administrator of the GameBus application can easily configure these personalization and gamification elements to be visible to any of the GameBus users. Each of the patient accounts in the Mendix prototype has a GameBus account assigned and can log into the GameBus application. Due to the challenges in the GameBus application now deriving directly from the treatment plan of the patients the challenges present in the application are even more relevant to the patient's treatment.

We extended the GameBus application to include the gamification element of levels. The relevant Mendix treatment plan goals of the patients are mapped into a level structure where each outcome goal is separated into different levels with increasing difficulty. The levels can be unlocked by completing the relevant behavior goals associated with the outcome goal. This level extension showcases the potential for patients to be more engaged with their treatment through this mHealth application, as the gamification and personalization elements are designed to maximize intrinsic motivation to the challenges in the application. The GameBus application also makes it possible to monitor which challenges each user is taking part in and track the progress users made on each of their challenges. Case managers can use this data to track which challenges their patients are making progress on, when their patients work on tasks within their challenges, and when patients have completed their challenges. The data could then potentially be used as a proxy for user engagement with their treatment.

from the Eindhoven University of Technology to assess the generated goals, over 5 iterations. The students received a survey that included an explanation of the project and a video of the latest version of the prototype. They were asked to rate the generated behavioral goals for two outcome goals and provide tips for improving the goals. The two outcome goals chosen for evaluation were based on data on treatment plan goals from patients with SMI, provided by a Mental Health and Addiction Care Institute in the Netherlands. Participants rated the goals based on the following elements of the SDT and Goal Setting theory: autonomy, competence, relatedness, clarity, challenge, commitment, feedback, and task complexity to assess the quality of the goals. This rating used a five-point Likert scale which consists of 1 = Strongly Disagree (SD); 2 = Disagree (D); 3 = Neutral (N); 4 = Agree (S); and 5 = Strongly Agree (SA). Each aspect was analyzed, and the average score was used to evaluate the goals.

We define a change in the average score of an element, compared to the previous average score of that element, as significant when there was a difference of at least 0.5. With eight aspects, the maximum score for a goal was 35. This approach allowed participants to provide specific ratings for each aspect, ensuring an assessment of their perceptions. By implementing the dimensions within the goal-setting theory and SDT frameworks into the goals, we increase the potential intrinsic motivation a person has to complete the goals. The group of students was also given a plain text field to provide feedback.

The information gathered, along with the provided feedback, guides the areas for improvement. The prompt in the prototype was revised according to the chosen improvements. Once these changes were implemented, the group was requested to complete the updated survey again. This iterative process aims to use prompt engineering to refine the goals based on participant feedback. By incorporating this, the prompt sent to GPT-3 can be improved and the quality of the generated behavioral goals can potentially be increased.

4 Results

4.1 Prototype

Mendix. The created Mendix prototype allows users to log in as patients or case managers. Patients can create, edit, and delete treatment outcome goals. For each treatment goal, the system generates 3 behavioral goals using GPT-3's API. Patient information and goal format instructions are included in the prompt sent to GPT-3. GPT-3's response is mapped and stored in the data model created in Mendix, and the goals are assigned to the patient. Patients can view and modify their behavioral goals. They can also remove or regenerate specific behavioral goals. After finalizing their treatment outcome goals, patients submit them to their case manager for approval before adding them to the treatment plan in the Mendix prototype. Patients have an overview page displaying all their approved and non-approved treatment outcome goals along with associated behavioral goals, each with a step-by-step guide. Case managers can also generate, edit, and remove goals for their assigned patients. Case managers are able to see all

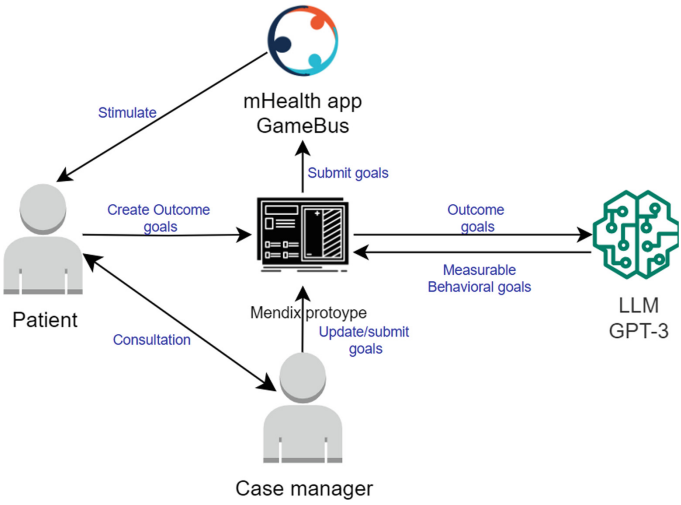


Fig. 1. The architecture of the prototype.

appropriate to first evaluate the goals with students before entering the ethical procedures. Once the basic capabilities have been established we will again involve patients and case managers in the process.

3.1 Prototype

Mendix. Mendix is a low-code development platform that was chosen to develop the prototype of the new goal-setting workflow, because of its ease of use built-in capabilities with REST API architectures, and ability to facilitate rapid prototype development. GPT-3 was accessed using the available API. The responses were formatted and directly implemented in both Mendix and the GameBus system. As the new workflow is being constructed, specific milestones were identified for evaluating the goals. For this purpose, behavior change theories were utilized. While three main theories were outlined in the theoretical background, in the scope of this study we focus on SDT and goal-setting theory.

GameBus. GameBus is a digital platform that promotes a healthy lifestyle by hosting and facilitating healthy challenges and competitions. Users are motivated to continue working on the challenges available to them through the use of gamification elements such as leaderboards and points [23]. For this study, we configured the GameBus platform as an mHealth application. The GameBus application can use measurable goals as input for challenges.

3.2 Measurable Goals

To assess if it is possible to improve the quality of the measurable goals that GPT-3 generates, by prompt engineering. We recruited a group of 8 students

healthcare stems from the ability to process and learn from massive amounts of free-text data. In theory, LLMs could generate measurable treatment goals based on data available on the web. Currently, much of the responsibility of creating treatment plan goals falls under the workload of the case manager [24]. In previous work, we created measurable goals for patients with SMI together with case managers [14]. Even when provided with a protocol for creating measurable goals, case managers found creating measurable goals for their patients a time-consuming and difficult task. Generating measurable goals using LLMs, could potentially alleviate some of the workload from case managers to patients, as it could empower patients to easily formulate measurable goals for their treatment. To the best of our knowledge, no literature is available that discusses the use of LLMs to generate treatment goals for patients with SMI.

The recent LLMs that have been released for public use usually strictly abide by the relevant laws and regulations of the countries in which they are released [16]. The LLMs usually do not mention anything offensive, violent or criminal in their conversations, and do not give any unethical medical advice. However, it is known that LLMs are known to hallucinate [1]. Hallucinations in the context of LLMs are when the system generates information not found in its training set [1]. This could potentially cause the generated goals to contain harmful elements. It is important to note that case managers are responsible for ensuring the safety of the goals within the patient's treatment plan. If patients are allowed to generate treatment plan goals unchecked using LLM technology there could be potential risks associated with that. Therefore it is recommended for both patients and case managers to assess the generated plan goals before adding them to the treatment plan.

Currently, electronic treatment records are utilized by mental health institutes to keep records of patients, including their treatment goals. However, these systems lack the inclusion of measurable goals. In our prototype, we explore features that can possibly add measurable goals to these systems.

3 Methods

To explore the feasibility of using LLMs to generate treatment plan goals for patients with SMI, a prototype of such a system was built. The functioning prototype was built using the GPT-3 LLM, the prototyping tool Mendix, and the gamification engine Gamebus. The architecture of the system can be seen in Fig. 1. The goals generated by the prototype were then evaluated using behavior change theory guidelines by a group of students. To evaluate a prototype with patients with SMI, a prerequisite entails securing the approval of the study by at least three ethical committees. In addition to obtaining ethical approval from the university, the mental health institute mandates an external committee to ascertain whether the study qualifies as medical-scientific research before conducting its own ethical review. Considering that the study is in the exploratory phase, evaluating the basic capabilities of LLMs and prompt engineering strategies to evaluate if generated goals could possibly be improved upon, it was deemed

According to the Self-Determination Theory (SDT), in the context of an mHealth tool, a tool that satisfies the need for autonomy (i.e., the desire to have control over tasks), competence (i.e., the need to acquire new skills), and relatedness (i.e., the desire to feel connected to others) can enhance intrinsic motivation [15,20]. The proposed workflow of using LLMs to generate treatment plan goals can potentially enhance the patient's intrinsic motivation, which in turn can lead to better engagement with the treatment goals. Finally, the Goal Setting Theory states that specific, challenging goals and appropriate feedback contribute to higher and better results. For engaging a person in a target behavior, goals must follow five principles: clarity, challenge, commitment, feedback, and task complexity [17]. Clear goals are Specific, Measurable, Attainable, Realistic, and Time-oriented (SMART) [4]. We will be using the SDT and Goal Setting Theory as criteria for evaluating the quality of the goals that are generated by LLMs. The generated goals will also be structured in the format of SMART goals, to the extent that we can accomplish this.

Treatment outcome goals focus on a result (e.g., losing 2 kg of weight), while treatment behavioral goals focus on an individual's action (e.g., going for a walk) [18]. For this study, we consider outcome goals as overall treatment plan goals, and behavior goals as smaller goals patients should achieve to reach their desired outcome goals. In order to track the progress toward the treatment goals, the behavior goals need to be measurable (e.g., I will go for a walk twice a week throughout the month of September). Integrating behavioral goals in mHealth interventions could potentially lead to more successful interventions [7]. An additional useful tool to motivate users in mHealth interventions is Gamification, which is the use of game design elements in non-game contexts [5].

2.2 AI in Mental Healthcare

While there has been a growing trend toward integrating AI technology in physical health applications, adopting such technology within the mental health domain has been comparatively slow [10]. Mental health practitioners emphasize patient-centered care and a hands-on approach in their clinical practice in contrast to non-psychiatric practitioners. This approach involves softer skills, including establishing strong relationships with patients and closely observing their behaviors and emotions [9]. Mental health clinical data is frequently in qualitative and subjective patient statements and written notes. Because of this, there is still much to be gained in the field of mental health practice through the incorporation of AI technology [10]. The application of AI techniques could present the opportunity to develop more accurate pre-diagnosis screening tools and risk models to determine an individual's susceptibility or likelihood of developing a mental illness [21]. AI-based LLMs have already showcased their efficacy in diverse areas, such as explainable AI, conversational agents, education, information retrieval, and text summarizing [6]. With their remarkable capabilities, LLMs can potentially transform various industries [10]. Research on LLM technology for mental health has yielded mixed results, and the enduring effects of using LLM on mental health remain unexplored [12]. The potential of LLMs in

This results in the majority of the goals currently found in the treatment plans not being suitable to be used in mHealth applications [14].

To assist case managers with making measurable treatment plan goals that can be used in mHealth applications, we introduced a protocol for a structured approach to creating measurable treatment plan goals for patients with SMI. Despite case managers being positive about the protocol, due to its time-consuming nature, it will have a low adoption rate. Given the high workload of case managers, they are not always open to new time-consuming tasks. Following all the steps in the protocol is a more time-consuming process for the case manager, compared to their usual goal-setting and tracking workflow. To reduce the time case managers would spend following the protocol when creating and tracking measurable goals, we will explore ways to leverage AI to potentially automate a part of the workflow.

The recent breakthroughs of Large Language Models (LLMs), such as BERT, GPT-3, and GPT-4 have been disruptive. These LLMs have been trained on large data sets and the models available have produced impressive results when prompted by humans. When prompted correctly, these models have the ability to respond to specific queries given to them when provided a specific context [2]. Due to this technology's ability to generate relevant text responses when given a context, we will explore the potential use of LLMs to introduce a new workflow for case managers to create treatment plan goals with their patients.

This study will aim to create a prototype that explores using LLMs to create AI-generated treatment plan goals, potentially improving the workflow of creating treatment plan goals with patients. Evaluating the prototype is outside the scope of this project. Nevertheless, we will assess the goals generated by the prototype to evaluate whether the quality of generated goals can be improved through modifications to the prototype and the prompts sent to the LLM.

2 Theoretical Background

To establish how to create measurable goals, in this section, we will review relevant work in the areas of behavior change, the use of AI in mental healthcare, and the possible risks of generating goals for patients with SMI using LLMs.

2.1 Behavior Change Theories

According to behavior change theories such as the COM-B system and the Fogg behavior model, behavior is a product of three fundamental factors: capability, opportunity, and motivation [8,18]. To successfully perform a targeted behavior at a particular time, it is essential to have the capability and opportunity, including an enabling environment. The strength of motivation to engage in the behavior must be higher than any other competing behaviors. These three factors interact to produce the desired behavior [20]. Motivation can be split into intrinsic and extrinsic motivation. Where extrinsic motivation is being motivated by external factors, intrinsic motivation is motivated by an inherent interest which leads to more persistence [20].

PTSD, schizophrenia, and severe depression pose unique challenges for both patients and mental healthcare providers [19]. One such challenge is mental health professionals' need to dynamically scale the care for each patient diagnosed with Severe Mental Illnesses (SMI) depending on the current state of the patient. Among people with ill mental health, those diagnosed with one or more severe mental illnesses for a period of over two years, and who struggle socially from their mental illnesses, are considered patients with SMI [24]. In the last few decades, there has been a growth in the number of patients with SMI treated by mental health care institutes, increasing the workload and work pressure of healthcare professionals [11]. Many patients with SMI live independently at home or in assisted living facilities without direct help from friends or family. Due to the associated vulnerabilities experienced by patients with SMI, there is a need for long-term treatment, with the ability to scale the care to the needs of the patient. Flexible Assertive Community Treatment (FACT) is a type of treatment for patients with SMI, that treats patients within their own home environment and provides care that matches their current needs [24]. Several countries have opted to prescribe FACT to patients with SMI, in an attempt to treat patients within their home environment [22]. A case manager is a healthcare professional who is directly responsible for monitoring the effects of the treatment. During FACT treatment, patients work together with their case manager to create a treatment plan, which includes the treatment plan goals that they will work on over a one-year period [24]. The goals of the treatment plan not only target symptom recovery but also seek to empower patients, enhance their self-reliance, and provide support in addressing social issues such as employment and housing. The goals serve as guiding principles of the treatment, ensuring that the care provided to patients is in accordance with their own goals, focusing on a person-oriented and holistic approach [24].

Currently, case managers assess the functioning of their patients through direct contact with the patients or their surroundings (e.g., family, general practitioner, etc.). Outside of direct contact, case managers do not have any tools to monitor the state of their patients, which results in them not being able to assess when to scale care for patients efficiently. Previous research shows that mobile Health (mHealth) applications are promising when it comes to positively influencing and tracking the behaviors of patients with SMI [13]. Such an mHealth application could potentially also improve FACT by assisting case managers in assessing their patients' functioning and monitoring the progress that a patient is making on their treatment [13]. This could potentially be a support tool to help case managers efficiently assess when to scale care for patients.

In order to monitor patients with SMI who are prescribed FACT, an mHealth application should track the progress a patient is making on their treatment plan goals. However, in previous research in collaboration with FACT teams at a Dutch Mental Health and Addiction Care Institute, we have discovered that on average only 25% of the available treatment plans have a form of measurable goals available within them [14]. A well-defined Measurable goal is a goal that can be tracked to monitor progress [3]. The treatment plan goals were revealed to have a significant lack of structure, consistency, and difference in level of detail.

Towards Augmenting Mental Health Personnel with LLM Technology to Provide More Personalized and Measurable Treatment Goals for Patients with Severe Mental Illnesses

Lorenzo J. James[1,2(✉)], Maureen Maessen[1], Laura Genga[1], Barbara Montagne[2], Muriel A. Hagenaars[3], and Pieter M. E. Van Gorp[1]

[1] Industrial Engineering and Information Systems, Eindhoven University of Technology, Eindhoven, The Netherlands
l.j.james@tue.com
[2] Treatment Center for Personality Disorders, GGZ Centraal, Center for Mental Health Care, Disorders, Amersfoort, The Netherlands
[3] Department of Clinical Psychology, Universiteit Utrecht, Utrecht, The Netherlands

Abstract. Mobile health (mHealth) tools are increasingly being used in various mental health domains to monitor patients with Severe Mental Illnesses (SMI), with the aim of potentially increasing patient engagement with their treatment. Patients with SMI who are prescribed Flexible Assertive Community Treatment (FACT) create a treatment plan together with their case manager, which serves as the leading document describing the goals that will be worked on during treatment. In order to incorporate the treatment plan goals of a patient in an mHealth application, the treatment plan goals need to be measurable. However, in previous work, we discovered that on average, only 25% of the available treatment plans include measurable goals. We have developed a protocol for making measurable goals with patients with SMI to address this issue. However, we anticipate low adoption of the protocol due to the potentially time-consuming nature of the steps involved. To mitigate this, we are exploring the use of AI to generate measurable treatment plan goals for patients with SMI and introduce a new workflow. In our exploratory study, we created a prototype of a system that may enable case managers and patients with SMI to generate measurable treatment plan goals using Large Language Models.

Keywords: SMI · mHealth · LLM · Gamification · Goals

1 Introduction

The impact of mental health is a growing concern for individuals and communities in our present-day society. Severe mental illnesses such as bipolar disorder,

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 186–200, 2024.
https://doi.org/10.1007/978-3-031-59717-6_13

17. Kaye, W., et al.: The problem of poor retention of cardiopulmonary resuscitation skills may lie with the instructor, not the learner or the curriculum. Resuscitation **21**(1), 67–87 (1991)
18. Kirkbright, S., Finn, J., Tohira, H., Bremner, A., Jacobs, I., Celenza, A.: Audiovisual feedback device use by health care professionals during CPR: a systematic review and meta-analysis of randomised and non-randomised trials. Resuscitation **85**(4), 460–471 (2014)
19. Kuyt, K., Park, S.H., Chang, T.P., Jung, T., MacKinnon, R.: The use of virtual reality and augmented reality to enhance cardio-pulmonary resuscitation: a scoping review. Adv. Simul. **6**(1), 1–8 (2021)
20. Laerdal: Little Anne CPR dummy. Official page (2018). https://www.laerdal.com/de/
21. Leary, M., McGovern, S.K., Balian, S., Abella, B.S., Blewer, A.L.: A pilot study of CPR quality comparing an augmented reality application vs. a standard audio-visual feedback manikin. Front. Digital Health **2**, 1 (2020)
22. LeBlanc, V.R.: The effects of acute stress on performance: implications for health professions education. Acad. Med. **84**(10), S25–S33 (2009)
23. Leong, B.: Bystander CPR and survival. Singapore Med. J. **52**(8), 573 (2011)
24. Medical, L.: CPRmeter 2 (2023). https://laerdal.com/de/products/medical-devices/cpr-feedback-devices/cprmeter-2/
25. Microsoft: Hololens. Official page (2018). https://www.microsoft.com/en-us/hololens
26. Olasveengen, T.M., et al.: European resuscitation council guidelines 2021: basic life support. Resuscitation **161**, 98–114 (2021)
27. Paglino, M., et al.: A video-based training to effectively teach CPR with long-term retention: the scuolasalvavita. it ("schoolsaveslives. it") project. Internal Emerg. Med. **14**(2), 275–279 (2019)
28. Perkins, G.D., et al.: European resuscitation council guidelines 2021: executive summary. Resuscitation **161**, 1–60 (2021)
29. Ricci, S., Calandrino, A., Borgonovo, G., Chirico, M., Casadio, M.: Virtual and augmented reality in basic and advanced life support training. JMIR Serious Games **10**(1), e28595 (2022)
30. de Sena, D.P., Fabrício, D.D., da Silva, V.D., Bodanese, L.C., Franco, A.R.: Comparative evaluation of video-based on-line course versus serious game for training medical students in cardiopulmonary resuscitation: a randomised trial. PLoS ONE **14**(4), e0214722 (2019)
31. Technologies, U.: Unity 3D game engine (version 2018.4.36f1) (2018). http://www.unity.com
32. Tsou, J.Y., Kao, C.L., Hong, M.Y., Chang, C.J., Su, F.C., Chi, C.H.: How does the side of approach impact the force delivered during external chest compression? Am. J. Emerg. Med. **48**, 67–72 (2021)
33. Wanner, G.K., Osborne, A., Greene, C.H.: Brief compression-only cardiopulmonary resuscitation training video and simulation with homemade mannequin improves CPR skills. BMC Emerg. Med. **16**(1), 1–6 (2016)
34. Yeung, J., Meeks, R., Edelson, D., Gao, F., Soar, J., Perkins, G.D.: The use of CPR feedback/prompt devices during training and CPR performance: a systematic review. Resuscitation **80**(7), 743–751 (2009)
35. Yigitbas, E., Krois, S., Renzelmann, T., Engels, G.: Comparative evaluation of AR-based, VR-based, and traditional basic life support training. In: 2022 IEEE 10th International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–8. IEEE (2022)

challenges. By doing so we aim to contribute to expanding the field of research on medical training and inform the development of more effective CPR training programs that enhance learner performance, confidence, and ultimately save more lives.

Acknowledgement. This research was supported by HumanE-AI-Net.

References

1. An, M., Kim, Y., Cho, W.K.: Effect of smart devices on the quality of CPR training: a systematic review. Resuscitation **144**, 145–156 (2019)
2. Arduino: Arduino pro mini (2018). https://www.arduino.cc/
3. Balian, S., McGovern, S.K., Abella, B.S., Blewer, A.L., Leary, M.: Feasibility of an augmented reality cardiopulmonary resuscitation training system for health care providers. Heliyon **5**(8), e02205 (2019)
4. Berg, R.A., et al.: Assisted ventilation does not improve outcome in a porcine model of single-rescuer bystander cardiopulmonary resuscitation. Circulation **95**(6), 1635–1641 (1997)
5. Braslow, A., Brennan, R.T., Newman, M.M., Bircher, N.G., Batcheller, A.M., Kaye, W.: CPR training without an instructor: development and evaluation of a video self-instructional system for effective performance of cardiopulmonary resuscitation. Resuscitation **34**(3), 207–220 (1997)
6. Brown, T.B., et al.: Relationship between knowledge of cardiopulmonary resuscitation guidelines and performance. Resuscitation **69**(2), 253–261 (2006)
7. Corp., I.: IBM SPSS statistics for windows (version 22.0) (2017). https://www.ibm.com/products/spss-statistics
8. Everson, T., Joordens, M., Forbes, H., Horan, B.: Virtual reality and haptic cardiopulmonary resuscitation training approaches: a review. IEEE Syst. J. (2021)
9. Greif, R., et al.: European resuscitation council guidelines 2021: education for resuscitation. Resuscitation **161**, 388–407 (2021)
10. Gruber, J., Stumpf, D., Zapletal, B., Neuhold, S., Fischer, H.: Real-time feedback systems in CPR. Trends Anaesthesia Crit. Care **2**(6), 287–294 (2012)
11. Higashi, E., Fukagawa, K., Kasimura, R., Kanamori, Y., Minazuki, A., Hayashi, H.: Development and evaluation of a corrective feedback system using augmented reality for the high-quality cardiopulmonary resuscitation training. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 716–721. IEEE (2017)
12. Hong, J.Y., Oh, J.H., Kim, C.W., Lee, D.H.: Hand injuries caused by feedback device usage during cardiopulmonary resuscitation training. Resuscitation **107**, e3–e4 (2016)
13. Ingrassia, P.L., et al.: Augmented reality learning environment for basic life support and defibrillation training: usability study. J. Med. Internet Res. **22**(5), e14910 (2020)
14. Issleib, M., Kromer, A., Pinnschmidt, H.O., Süss-Havemann, C., Kubitz, J.C.: Virtual reality as a teaching method for resuscitation training in undergraduate first year medical students: a randomized controlled trial. Scandinavian J. Trauma, Resuscitation Emerg. Med. **29**(1), 1–9 (2021)
15. Johnson, J.G., Rodrigues, D.G., Gubbala, M., Weibel, N.: HoloCPR: designing and evaluating a mixed reality interface for time-critical emergencies. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 67–76 (2018)
16. Jung, C., et al.: Virtual and augmented reality in cardiovascular care: state-of-the-art and future perspectives. Cardiovascular Imaging **15**(3), 519–532 (2022)

we standardized teaching routines without diminishing learners' interest, promoting verbal communication between teacher and student.

In line with the findings of Balian *et al.* [3], our study further substantiates the potential of AR in CPR training. By incorporating a control group and measuring baseline performance, we provide additional evidence to support the effectiveness of AR for CPR training. This strengthens the understanding of the benefits and possibilities that AR technology offers in enhancing CPR education through various designs and implementations.

Although Leary [21] did not find a statistically significant difference between CPRReality training and a standard audio-visual feedback manikin, our study demonstrates significant improvements in CPR performance within the experimental group, suggesting the advantages of our AR-based approach in enhancing CPR quality compared to traditional methods.

7.1 Limitations

While our system design and study results provide valuable insights, it is essential to acknowledge its limitations. Certain technical issues, such as loss of environment mapping, occasional teacher avatar misplacement, and limited field of view, were identified during the experiment indicating the need for further optimization. Additionally, the study's controlled environment and participants' awareness of evaluation may have influenced their performance, necessitating further validation of the app's real-world applicability in diverse settings with participants unaware of being evaluated. Furthermore, further research is needed to assess the long-term durability and retention of the improved CPR performance observed in our study. Lastly, our study primarily focused on compression depth and frequency as key performance indicators, future studies could consider incorporating a more comprehensive assessment, including other critical elements such as hand placement, posture, and breath technique.

8 Conclusion

In conclusion, our study provides compelling evidence of the effectiveness of RescuAR, an AR-based CPR teaching and training app, in improving compression depth, frequency, and overall CPR performance. The insights gained from our study inform the development of guidelines and best practices for incorporating AR technology in CPR training programs. Policymakers, educators, and healthcare professionals can use this information to establish standards and recommendations for the implementation and utilization of AR-based training tools. This includes considerations such as the design of instructional content and feedback mechanisms, and the integration of AR training into existing curriculum frameworks. While this study primarily focused on the fundamental teaching of CPR, our future work will delve into exploring the integration of more realistic and immersive simulation scenarios. By incorporating advanced training systems, we aim to investigate ways of providing learners with a comprehensive and realistic training experience that prepares them for a wider range of CPR situations and

7 Discussion

The results of our study provide compelling evidence of the effectiveness of RescuAR in self-directed CPR teaching and training. The significant improvements observed in both compression depth and frequency among participants in the experimental group demonstrate the positive impact of RescuAR on CPR skills. Moreover, the experimental group exhibited significantly higher improvements in effective CPR performance compared to the control group, underscoring the comprehensive benefits of RescuAR's AR-based approach. These findings suggest that traditional CPR teaching methods may have limitations in adequately addressing crucial aspects of skill acquisition and retention.

Acquiring physical skills like CPR relies on developing accurate muscle memory, and timely intervention is crucial to prevent the formation of incorrect habits. However, accurately assessing learner performance is challenging for teachers without measurement devices. Even with measurement systems, the timing and manner of intervention can impact learner performance. Previous research has explored the use of measurement systems to provide insights for teachers, but their effectiveness is influenced by various factors, including intervention timing and the method used [10,34].

Our system showed that using real-time feedback modalities integrated into the teaching routine could help to overcome this challenge. These findings align with previous research highlighting the benefits of audio-visual feedback in CPR education [10,18]. Furthermore, our approach for real-time detection and visualization of CC depth and frequency helped to seamlessly integrated the real-time feedback into the training routine overcoming the limitations caused by using commercial CPR feedback devices such as hand injuries due to the placement of the device, or instability of using smart devices for measurement during CC [1,12].

Moreover, successful teaching requires sufficient dedicated time to achieve specific skills or abilities. However, in a traditional setting, time constraints and variations in individual capabilities make it challenging to provide adequate attention to each student. In a crowded classroom setup, many students may hesitate to ask questions or request repetitions during training sessions. This can hinder their learning experience and limit their understanding of the subject matter. **A self-directed approach — as implemented in RescuAR — promotes active participation, encourages question asking, and allows students to engage with the material at their own pace**.

Additionally, instructors themselves differ in their approaches and abilities to teach and evaluate effectively [17]. Consequently, evaluations or presentations dependent on human teachers may lack objectivity and standardization. To address these challenges, previous studies have attempted to unify teaching routines and enhance training sessions using various technologies. Some studies focused on teaching CPR principles through videos [5,27,33]. While unifying teaching is essential, learners' active engagement also plays a significant role in improving outcomes [30]. For instance, de Sena *et al.* found that although video-based teaching improved CPR skills, participants preferred more interactive and engaging self-training over passive video-based instruction [30]. Our study demonstrates that merging approaches can provide an efficient alternative for CPR education. **By utilizing a virtual teacher with programmed curricula,**

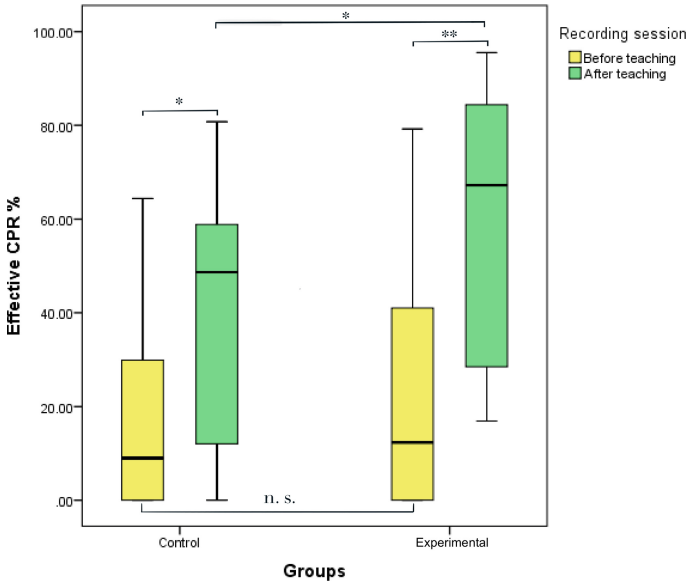


Fig. 5. The figure demonstrated the statistical analysis of control and experimental groups - n.s. = no significant difference between indicated groups ($p < 0.05$) * = Significant difference between indicated groups ($p < 0.05$), ** = Significant difference between indicated groups ($p < 0.01$)

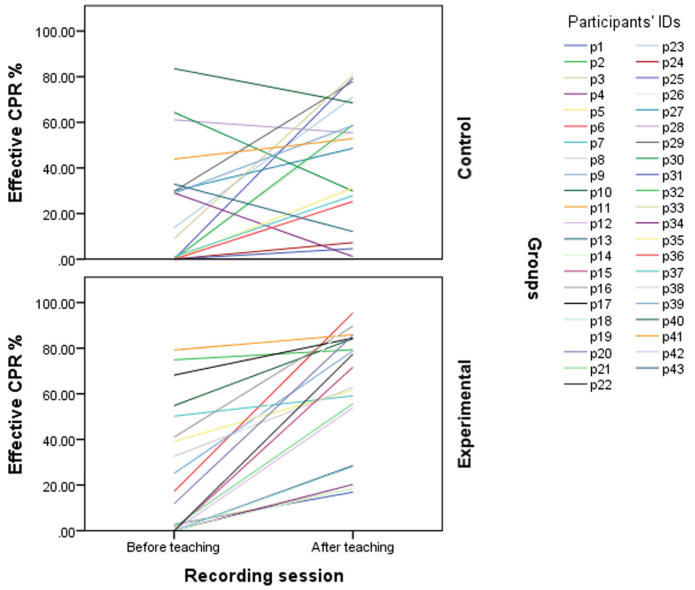


Fig. 6. Participant's performance lines before and after teaching sessions

In this study, the percentage of effective CPR performance of a participant was calculated based on the percentage of the time that the participants complied with all the above-mentioned guidelines at the same time while doing CPR.

Statistical analysis was performed using IBM SPSS Statistics for Windows, Version 22.0 [7]. Numerical data were presented as means ± standard deviations. To compare the improvement in two groups before and after teaching sessions, Paired t-test was used. To compare the outcomes between two study groups, Chi-square test (for categorical data) and Student's t-test (for numerical data) were used. A two-sided p-value less than 0.05 was considered significant in all analyses.

6.5 Results

RescuAR Evaluation. The participants in the experimental group showed significant improvements in performing correct depth and frequency. While the calculated p-value for CC depth improvement was 0.003, the p-value for frequency improvement was below 0.0001. A comparison between both criteria before the experiment session showed that more participants had issues finding the correct frequency (39.8% ± 36.9% correct frequency) than the correct depth (58.7% ± 39.8% correct depth). Even though for most of the participants, both depth and frequency were improved after the tutorial and training session, the frequency improvement rate was higher (73.8% ± 28.2%) than the improvement rate in CC depth (84.3% ± 23.2%).

Moreover, the results showed that the overall effective CPR performance of the experimental group significantly increased ($p < 0.0001$) after teaching session with RescuAR. While the mean of participants' performance was 23.3% ± 27.2% before the teaching session, it was improved to 61.3% ± 27.4% after training with the proposed system.

Experimental Group vs. Control Group. The study analysis showed no significant difference regarding characteristics distribution between the two groups concerning sex ($p = 0.332$) and knowledge backgrounds ($p = 0.543$) of the participants. Based on the findings, RescuAR helped to achieve higher effective CPR rates compared to traditional teaching. Analyzing the final performances of both groups' participants, it was observed that, even though both groups had no significant difference in doing effective CPR before the teaching session (control = 20.3% ± 25.4%, experimental = 23.3% ± 27.2%, $p = 0.594$), the experimental group performed significantly better than the control group after the teaching session (control = 40.4% ± 28.5, experimental = 61.3% ± 27.4%, $p = 0.019$) (Fig. 5).

Moreover, performance degradation only occurred in some of the control group's participants, and all participants in the experimental group showed improvement after the teaching session (Fig. 6).

6.3 Participants

Total of 44 persons volunteered to participate in this study. Among those, 43 persons' data were included and one person's data were excluded from the study due to data corruption. After exclusion, the experimental group consisted of 22 volunteer participants and the control group contained 21 participants. The participants were nurse students and laypeople who were randomly assigned to study groups. The nurse students were recruited by Southampton University and laypeople were recruited by DFKI. All participants were required to be aged above 18. Characteristics of participants are presented in Table 2.

Table 2. The participants' characteristic distributions

Characteristics	Control	Experimental
Age (years), Median (IQR)	22, (20–24)	21, (19–24)
Female	5 (23%)	9 (41%)
Male	16 (77%)	13 (59%)
Non-binary	0 (0%)	0 (0%)
Nurse Student	9 (43%)	9 (41%)
Laypeople	12 (57%)	13 (59%)
With prior knowledge	8 (38%)	11 (50%)
Tried or Familiar with AR (yes)	2 (9%)	1(4%)

IQR, interquartile range; AR, augmented reality; With prior knowledge, who is CPR certified and/or completed a first aid course

6.4 Data Collection and Analysis

Over two 1-minute cycles, two CC measurements were recorded before and after the teaching session using the same sensor-equipped CPR manikin described in Sect. 5.2 with a sampling rate of 100 Hz. The effective CPR performances of the participants were analyzed according to the latest evidence-based guidelines for resuscitation officially published by European Resuscitation Council [28]. These guidelines suggest an effective CPR as follows:

- CC in a frequency of at least 100/min but not exceeding 120/min.
- CC with a depth of at least 5 cm but not exceeding 6 cm

Control Group. The control group underwent a traditional CPR teaching and training session, which involved classroom-based instruction and practice using the CPR manikin. The same manikin as the experimental group was used in the control group to avoid any biases. A certified teacher was recruited to provide essential information on performing CPR based on European resuscitation council guidelines [26]. The session began with a theoretical introduction to the airway, breathing, and circulation techniques, followed by a demonstration from the teacher on the correct CPR procedure using the CPR manikin. Participants were then given the opportunity to practice CPR on the same manikin under the observation of the teacher. The teacher interrupted and gave feedback whenever they felt essential. Throughout the session, participants were encouraged to ask questions, repeat the training, and perform additional CC cycles if they desired. No time constraints were applied during teaching and training session. To minimize bias between the groups, the training sessions were conducted on an individual basis.

Both groups received instruction on airway, breathing, and circulation techniques. However, during the data recording session, participants were only asked to perform chest compressions, as the main focus of our study was to evaluate the quality of chest compressions in terms of depth and frequency while neglecting the potential effects of the breath technique. No device or extra feedback method was used during data recording. The performances of participants were assessed before and after the study to measure their baseline CPR skills and the improvements achieved through the assigned teaching training method. By comparing the performance improvements between the experimental and control groups, the study aimed to evaluate the impact of RescuAR system on enhancing CPR skills.

6.2 Study Protocol

This protocol was reviewed and approved by the ethical committee of Southampton University. All of the participants were informed about being free to participate in the research and nondisclosure of personal information. They all agreed and signed written informed consent. Upon written consent, the experiment protocol was performed in five ordered stages: (1) Demographic survey, (2) baseline CPR recording (3) randomized group assignment (4) teaching and training session, and (5) post-training CPR recording. The survey collected demographic and characteristic information including age, gender, experience background in CPR, and familiarity with wearable AR devices. After completion of the survey participants were asked to perform two cycles of 1-minute hands-only CPR. Participants rested at least two minutes between each cycle. Later all participants were randomly divided into two experimental groups to receive teaching and training session. After the teaching session participants rested at least for 10 min to avoid the effect of tiredness on their performance. Later they performed another two cycles of 1-minute hands-only CPR (without usage of any feedback device or extra help) with at least two minutes rest between each cycle. As various studies demonstrated the importance of hands-only CPR in bystanders [4,23] this study only focused on the evaluation of hands-only CPR performance, and any effect regarding performing rescue breath was neglected.

$$x = \frac{F}{k}$$

Where F is the force applied to the compression spring (reading from the pressure sensor converted to Newtons, N), k is the spring constant ($\frac{N}{m}$), x is the displacement of the spring (m).

The accuracy of the employed method for calculating the depth of CC was verified against a commercial CPRmeter device [24] to ensure the validity of the measurements.

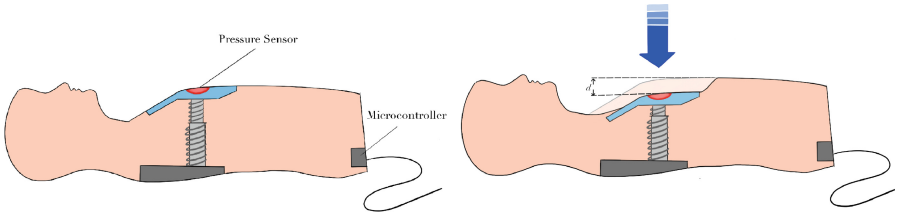


Fig. 4. Placement of Pressure sensors and controller inside the manikin. F = Applied force to the spring (N), k = spring constant ($\frac{N}{m}$), x = displacement of the spring (m)

6 User Study and System Evaluation

To evaluate the effectiveness of our designed RescuAR system, we conducted a user experiment. The study took place in two different experiment centers, Southampton University Southampton, UK, and German Research Center for Artificial Intelligence (DFKI) Kaiserslautern, Germany. The primary objective of this experiment was to assess and compare the CPR performance of participants who used the RescuAR system with those who underwent traditional CPR training methods (RQ2).

6.1 Study Design

The study utilized a randomized controlled trial design to investigate the effectiveness of different teaching methods for CPR training. Participants were randomly assigned to either the experimental or a control group.

Experimental Group. The experimental group received a self-directed CPR teaching and training session using the RescuAR system (Application and the CPR manikin). The teaching session began by calibrating the HoloLens for each individual and starting the application. Each participant completed the teaching and training session without any time limit.

To teach the correct frequency of CC, two distinct audio cues were employed. Initially, the iconic "Stayin' Alive" song, known for its rhythm matching the recommended CPR procedure (104 bpm), was utilized to encourage participants to perform CC in sync with the song's beats. Subsequently, a metronome sound with the same frequency (104 bpm) rate was introduced to enhance learner focus and synchronization.

Regarding CC depth, a trial-and-error approach was adopted to instill the appropriate technique. To align with the preferences of professionals who recognized the benefits of using a design similar to standard commercial devices, real-time feedback on compression depth was provided through a visual depth panel resembling CPRmeter devices [24]. This allowed learners to assess their performance and make adjustments accordingly (Fig. 3). The application allowed for the repositioning of the virtual depth feedback panel using the manipulation gesture of HoloLens. To address the shaking effect occurred during CPR performance, we implemented a functionality to fixate the depth feedback panel to the corner of the field of view. Participants could enable or disable this feature using a voice command. To eliminate the need and urge to look at the teacher avatar during chest compressions, which could be affected by the shaking effect, the visual rendering of the avatar was disabled during the practice session and reappeared after the completion of the stage. This effect was also explained by the teacher to the user to avoid any confusion. Additionally, a click sound reminiscent of the standard CPR dummy's internal clicker (which had been previously removed to avoid confusion) indicated the moment when the correct depth was achieved.

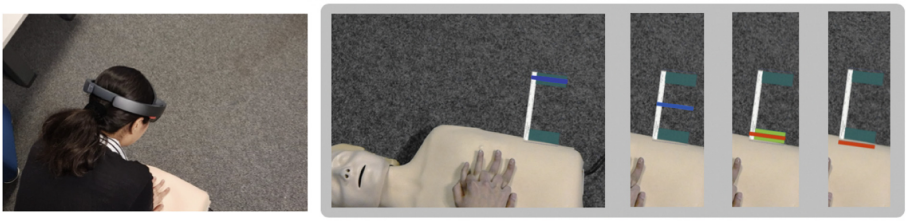
Upon completion of training for both frequency and depth criteria, a final session provided an opportunity for learners to combine both elements, aiming to perform CC with the correct frequency and depth in a synchronized manner. This comprehensive training approach equipped participants with the necessary skills to deliver effective CC during CPR.

5.2 CPR Manikin

To acquire information such as depth, frequency, and pressurized position of the CC on the manikin, a single FSR (Force-Sensing Resistor) was used. The sensor was attached beneath the skin layer of the CPR manikin's chest plate, precisely positioned in the center of the chest where the hands should be placed during correct CPR (Fig. 4). To facilitate seamless data acquisition and control, the pressure sensors were directly connected and managed by an Arduino Pro Mini board [2]. The analog signals generated by the embedded sensors were transformed into digital signals through the Arduino platform. Subsequently, these digital signals originating from the sensors were transmitted via a serial port to a local computer for further utilization and analysis.

To calculate the depth of chest compressions, we adopted the method proposed by Tsou *et al.* in their study [32]. This method utilizes Hooke's law to calculate the depth based on the pressure applied to the chest spring. We assumed the spring inside the dummy to be a linear spring due to the same diameter along its entire length. We converted the FSR readings to Newton to calculate the spring constant. Before the experiment, we measured the spring constant using the FSR force value and the spring displacement using the following formula:

(a)



(b)

Fig. 3. Third-person and first-person (in-app) views of RescuAR application. (a) Theoretical phase. (b) Practical phase

feel like you need further practice, you can let me know at any time. Just use the word 'repeat!"'

To ensure a clear understanding of proper technique, animated 3D avatars were employed to visually demonstrate the correct posture and hand positioning during CPR (Fig. 3). By utilizing 3D virtual models and realistic placement within the room, participants were able to walk around the models and observe the CPR technique from different angles, providing a more immersive and interactive 3D training experience that cannot be achieved using 2D displays. By combining audio instruction, written scripts, and animated avatars, the theoretical phase of the training aimed to maximize participant comprehension and knowledge acquisition.

The application commenced with a calibration phase, which involved scanning the environment and room using the capabilities of the HoloLens and Mixed Reality Toolkit (MRTK) to obtain a spatial mesh of the environment. We utilized plane-finding methods to detect suitable surfaces for the placement of virtual teacher avatars and other demonstrative avatars. Colliders were added to the room mesh to create a more realistic movement area for the teacher avatar.

Practice Phase. The second phase of the application focused on practical training to cultivate the essential muscle memory required for CPR proficiency. Based on the feedback from survey participants, separate hands-on practice addressing each essential criterion of CPR was emphasized as important. As a result, two key criteria, frequency and depth, were targeted individually and then combined to ensure comprehensive learning.

Given that the primary objective of this study was to evaluate participants' performance specifically regarding CC, the instructions pertaining to rescue breaths were exclusively included in the theoretical section and not incorporated into the practical training. This deliberate decision allowed for a more targeted assessment of participants' proficiency in CC, aligning with the study's primary focus.

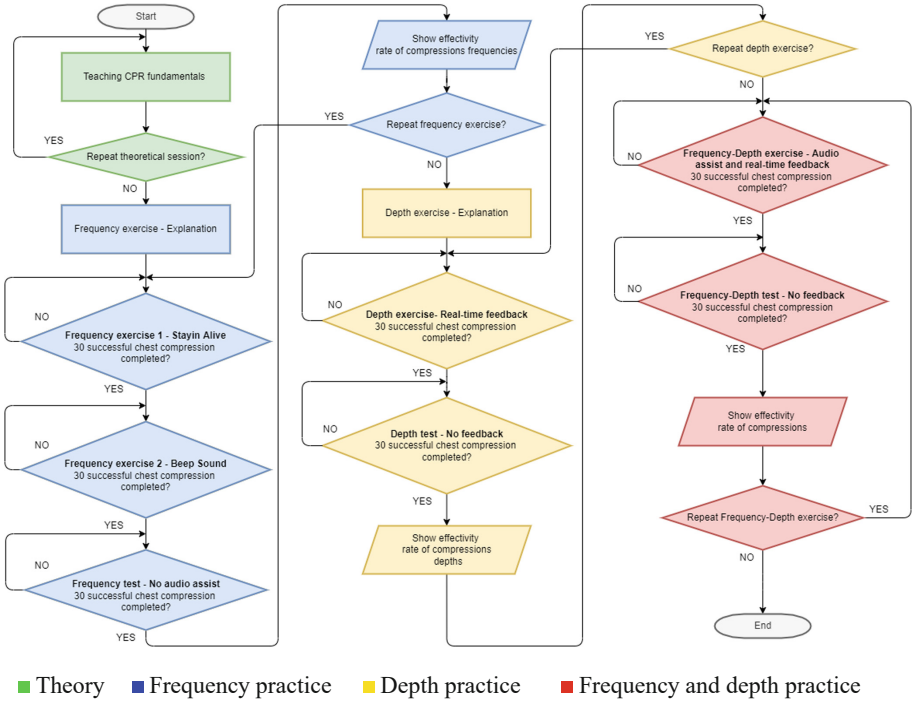


Fig. 2. The flowchart explaining the routine of RescuAR CPR tutorial and training application

Theory Phase. During the theoretical phase, users were introduced to the fundamentals of CPR (airway, breathing, and circulation techniques) based on European resuscitation council guidelines [26] through an engaging virtual teaching experience. A virtual teacher avatar was designed in following the survey results and served as the guide, delivering the essential CPR basics both audibly and visually. To enhance comprehension, the information was presented not only through the avatar's spoken instructions but also as easily understandable written scripts. This method was deemed to be the most suited approach for this application by survey participants. The communication between the virtual teacher and the participant was facilitated using voice commands. The virtual teacher provided the voice commands at the end of the conversation to reduce the need for memorization. For example, the virtual teacher would say, "If you

In terms of frequency acquisition, healthcare professionals recommended the inclusion of audio elements to facilitate proper timing during CPR. They expressed the belief that incorporating the iconic "Stayin' Alive" sound, which aligns with the required rhythm for CPR, would enhance the training experience. However, to promote stronger muscle memory and improve precision, professionals also suggested the inclusion of a metronome sound.

For correct depth acquisition, the respondents emphasized the importance of incorporating a simple graphical visualization, similar to the devices used in traditional teaching setups to provide visual cues on the depth of CC. Participants pointed out the importance of optimizing the orientation and location of the visualization within the user's field of view to ensure clarity and accuracy during CPR training. They recommended that the depth display be positioned near the manikin's chest within the user's field of view, in a fixed position to maintain proper posture during performance. However, considering that CPR is a highly physical activity and individuals may have different preferences, participants suggested that it would be beneficial if the position of the depth display could also be adjustable by the user based on their personal preference. This customization feature would allow users to optimize their viewing experience and ensure optimal training engagement.

5 RescuAR System Design and Implementation

Based on survey findings, an AR-based system prototype was developed with the aim of creating an immersive multi-sensory CPR teaching and training tool, incorporating audio, visual, and tactile elements, to enhance the learning experience and real-time feedback to foster a comprehensive understanding of the CPR technique.

The application was designed and developed using the Unity 3D game engine [31]. Microsoft HoloLens [25] was used as a wearable AR device to run the application.

The system consisted of two main parts: RescuAR application, and a standard CPR manikin [20] covered with custom pressure sensors. The RescuAR application utilized real-time data streams from the CPR manikin to provide feedback on compression rate, depth, and recoil. A local wireless communication scheme facilitated seamless interaction between the manikin and the application, enabling accurate and immediate feedback on CPR performance.

5.1 RescuAR Application

The formulation of the CPR training app's prototype design was based on the results derived from both quantitative and qualitative data collected through the survey. According to the survey participants, a two-phase delivery of teaching material was found to be appropriate. They recognized that understanding the underlying principles and concepts of CPR is crucial in order to perform the techniques accurately and confidently. Hence the application routine was divided into two distinct phases: Theory and practice. During the initial theory phase, the application focused on imparting the fundamental principles of CPR. The subsequent practice phase emphasized hands-on practical training. The flow chart of the application routine is presented in Fig. 2.

qualitative data that complemented the quantitative findings, allowing for a richer and more nuanced interpretation of the survey results. These results highlighted the essential design elements and interaction methods that were considered crucial for an effective CPR training application based on the professionals' expertise and daily experiences in the healthcare field.

Instructional Content, Routine, and Design Elements. The survey respondents provided valuable insights and reached a consensus on several key aspects of the CPR training application design. They emphasized the importance of incorporating a combination of audio, scripted text, and a virtual human teacher within the app to effectively deliver theoretical materials and provide personalized guidance and support. Additionally, participants strongly favored the use of animated human avatars as a demonstration tool for CC and breath techniques compared to other methods such as 2D videos of real persons performing CPR or audio instructions and verbal cues. The dynamic and interactive nature of animated avatars was perceived as more engaging and effective in conveying the correct techniques for CPR training. In terms of the teaching routine, participants suggested that it should begin with a theoretical representation of the concepts, followed by hands-on practice. They further recommended that the teaching of compression depth and frequency should be initially conducted separately before combining them in the training sessions. Lastly, participants highlighted the significance of enabling repetitive theoretical and practical sessions. They emphasized the importance of incorporating functionality to navigate through the different stages of the training. This feature would provide users with the ability to easily access previous and next steps, facilitating repetitive practice sessions. The opportunity for repetitive practice allows learners to review and reinforce their CPR skills, ultimately improving their performance and confidence.

Input and Interaction. Regarding input and interaction, the survey respondents expressed a strong preference for the integration of interaction methods through physical buttons or verbal communication via voice commands. Both methods were considered intuitive and practical for engaging with the app during training sessions. The respondents emphasized the advantages of voice commands, as they enable users to perform actions without the need for manual input, allowing them to focus on performing CPR on the manikin. This hands-free interaction was seen as a convenient and efficient way to engage with the app. On the other hand, physical buttons were identified as a suitable option for environments with high noise levels or crowded settings. Furthermore, participants expressed the belief that using virtual buttons or hand gestures for interaction would not be suitable for this task, as they would require users to have prior knowledge of how to interact with such input methods.

Methods for Frequency and Depth Acquisition. The survey respondents highlighted the crucial role of immediate and real-time feedback during hands-on training sessions for acquiring the correct frequency and depth in CPR. Among various types of feedback, the combination of audio and visual feedback emerged as the most preferred method for both frequency and depth acquisition.

material. Additionally, participants were asked about their preferred delivery modes, such as text, images, videos, or interactive elements. Furthermore, more detailed questions focused on finding the most effective teaching methods for acquiring correct CPR skills such as correct CC depth and frequency, along with suggestions for optimization.

Under the interaction and feedback section, participants were asked to express their preferences for input methods, such as voice commands, button interactions, and gestures. Additionally, participants were asked to indicate their preferred methods of real-time feedback from app, whether visual, audio, or haptic.

Lastly, the miscellaneous section encompassed more general questions, allowing participants to share their opinions on gamification aspects of the AR-based CPR training app. Additionally, participants were given the opportunity to provide any overall suggestions or feedback they deemed relevant to the development of the app.

By capturing diverse perspectives from healthcare professionals, the survey played a crucial role in informing the design and development of an effective AR-based CPR training app.

4.2 Participants

A survey was conducted among 11 healthcare professionals. The participants' demographic characteristics, including their professional backgrounds and relevant experience, are summarized in Table 1.

Table 1. The participants' characteristic distributions

Characteristics	
Age (years), Median (IQR)	39, (32–48)
Female	2
Male	9
Non-Binary	0
Experience in healthcare (years), Median (IRQ)	20 (14–23)
Tried or Familiar with AR (yes)	6 (54%)

IQR, interquartile range; AR, augmented reality

4.3 Survey Findings

The open-ended questions in the survey were analyzed using qualitative data analysis methods to evaluate the responses and gain deeper insights into participants' perspectives. The analysis process involved systematically reviewing the open-ended responses, coding them for key themes and patterns, and organizing the data into meaningful categories. Through this approach, we were able to identify common themes, explore variations in participants' experiences and opinions, and gain a comprehensive understanding of the topics under investigation. The qualitative analysis provided valuable

underwent traditional teaching methods. The CPR performances of participants were evaluated before and after the training sessions, specifically focusing on parameters such as compression depth and frequency. As various studies demonstrated the importance of hands-only CPR in bystanders [4,23], this study's primary objective was to evaluate participants' CPR performances specifically regarding hands-only CC and any effect related to breath technique was neglected.

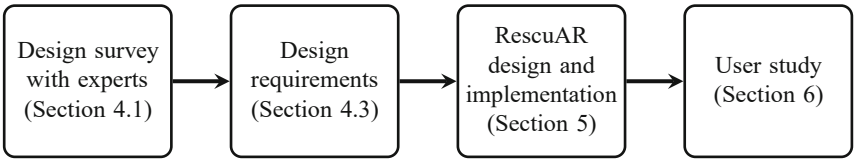


Fig. 1. The study flow process starting with the specification of design requirements based on expert survey findings and later evaluation of the designed system in a user study

4 Design Requirements for Self-directed AR-Based CPR Teaching and Training Tool

To inform the design of our interactive CPR teaching and training tool, we conducted a survey among experienced healthcare professionals. All professionals were familiarized with the AR device capabilities that will be used for the study prior to conducting the survey. The purpose of the survey was to identify the key requirements (RQ1) for developing an AR-based self-directed learning system that could effectively support CPR education. By gathering insights from these professionals, we aimed to create a tool that addresses the specific needs and challenges of learners in the context of CPR training.

4.1 Design Survey

The survey employed a combination of open-ended and multiple-choice questions. To provide a more structured approach the questions were categorized into five categories: User needs and requirements, user interface and design, instructional content, interaction and feedback, and miscellaneous.

While the questions under the user needs and requirements category focused on the participants' opinions on the features and functionalities that are considered essential for an AR-based CPR training app, the user interface and design questions focused more on the appearance and aesthetic aspects of the app.

In the instructional content section, several questions were asked to gather insights regarding the content, flow, and delivery mode of the teaching materials. Participants were asked to provide feedback on the clarity and comprehensiveness of the instructional content, as well as their preferences for the organization and sequence of the

end, we provide insights into the design, implementation, and evaluation stages of an AR-based CPR training application suited for wearable AR devices. In addition, we introduce a unique and cost-effective solution for real-time measurement of compression depth, frequency, and recoil during CPR training. Moreover, our study strengthens the evidence supporting the effectiveness of AR-based CPR training by incorporating a control group and baseline performance measurements. By combining these efforts, we hope to enhance CPR training effectiveness and ultimately improve the outcomes of life-saving interventions.

3 Methodology

Upon reviewing the existing literature, we identified a gap in the research regarding the detailed exploration of the application design process for AR-based CPR training tools, as well as the limited availability of quantitative data for a comprehensive comparison between these tools and traditional CPR training methods. Motivated by these gaps in the existing literature, our study aims to fill this research void by answering the following research questions:

RQ1: What are the required design features and functionalities for an AR-based self-directed CPR teaching and training tool?

We answer this research question by conducting an extensive design survey among healthcare professionals and investigating their professional opinion on the required features and specifications for such training tool.

RQ2: Are the performance results of the designed system comparable with traditional teaching methods?

Our objective was to address this question by conducting an experimental trial where we compared the CPR performances of participants, who used the designed system to learn and practice CPR routine, with participants who underwent traditional teaching. We measured their performances before and after teaching sessions in terms of correct depth, correct frequency, and overall effective CPR (correct depth and correct frequency) to develop a better understanding of the efficiency of the designed system.

To this end, we employed a mixed-method design consisting of three consecutive steps: data collection for design requirements, system design and implementation, and system evaluation (Fig. 1).

Initially, a survey was conducted among a group of health professionals to gather valuable insights and identify the design requirements and constraints for an AR-based self-directed CPR teaching and training tool. The survey aimed to understand the preferences, needs, and expectations of professionals regarding CPR training, as well as their perceptions of the potential benefits and challenges of using such an interactive system.

Based on the insights and design requirements identified through the survey, RescuAR, an AR CPR self-training tool, was conceptualized and designed. The primary goal of RescuAR was to enable users to learn and improve their CPR performing skills by providing them with interactive features and real-time data visualization.

To assess the effectiveness of RescuAR, a user study was conducted involving nurse students and laypeople. The participants were randomly assigned to either the experimental group, which utilized RescuAR for CPR training, or the control group, which

education through survey findings from healthcare professionals. Furthermore, we conducted a user evaluation to assess the impact of RescuAR on learners' ability to acquire and retain essential CPR skills. Our study demonstrated the effectiveness of RescuAR in improving CC depth, frequency, and overall CPR performance, providing valuable insights into the potential of AR as a transformative tool for CPR education and skill development. The findings of this study will contribute to the growing body of research on AR-based medical training and inform the development of more effective CPR training programs that enhance learner performance, confidence, and ultimately save more lives.

2 Related Work

In recent years, several approaches have been investigated to improve the effectiveness of CPR training, ranging from traditional classroom-based instruction to advanced technological interventions, such as virtual reality simulations and AR applications [8, 14, 16, 19, 29, 35].

In a recent study done by Balian *et al.* [3], the feasibility of an AR CPR training system (CPReality) for healthcare providers was tested. Their study results showed that the integration of AR into CPR training has the potential to be a valuable educational strategy that goes beyond simply translating knowledge and skills. However, their study has limitations, including inherent selection bias, a potential learning effect, a lack of baseline CPR performance assessment, and the absence of a control group for comparison. These limitations should be considered when interpreting the study's findings.

Leary *et al.* [21] focused on the limitations of Balian *et al.* work [3] and compared the use of CPRReality training with a standard audio-visual feedback manikin in terms of improvement in CPR quality. The findings of their study indicated that there was no statistically significant difference observed between the two groups. This implies that further, more extensive studies should be conducted to explore whether AR CPR training has the potential to enhance overall CPR quality for both new and re-certifying healthcare providers.

In another study done by Ingrassia *et al.* [13], an AR-based basic life support training system (Holo-BLSD) was proposed and evaluated in terms of feasibility and acceptability. While their study evaluates users' experiences and perceptions through a survey, it does not provide comprehensive evidence or direct comparisons of performance results in comparison to traditional CPR training methods. Therefore, the study may lack the empirical data needed to evaluate the effectiveness and efficacy of the Holo-BLSD system in terms of skill acquisition, retention, and performance outcomes.

In a related study, Johnson *et al.* [15] proposed the use of mixed reality (MR) to support time-critical emergencies. Their work introduced HoloCPR, an MR application that provided real-time instructions for resuscitation through a combination of visual and spatial cues. While their study demonstrated the potential of MR in decreasing reaction time and improving procedural accuracy, it primarily focused on the evaluation of these specific outcomes.

In this study, we aim to contribute to the existing body of research by addressing the limitations and building upon the findings of previous studies in the field. To this

have shown limitations in terms of skill acquisition, retention, and providing immediate feedback on performance.

One prominent challenge in traditional CPR training lies in the gap between knowledge acquisition and practical application. While learners may grasp the theoretical concepts, transferring that knowledge into effective hands-on performance can be challenging [6]. Research has shown that individuals often struggle to translate their theoretical knowledge into practical skills when faced with high-stress situations, such as cardiac arrest scenarios [22].

Furthermore, the ability to receive real-time feedback on performance during CPR training is critical for learners to correct errors, refine their technique, and build confidence. Immediate feedback allows learners to adjust their actions, ensuring the application of correct chest compression (CC) depth, rate, and recoil. However, traditional methods of CPR training often lack the means to provide instantaneous and accurate feedback, leaving learners uncertain about their proficiency and limiting their ability to improve their skills effectively.

To address these challenges and bridge the gap in CPR education, the combination of augmented reality (AR) technology with sensing modalities emerges as a promising solution. AR integrates virtual elements into the real-world environment, offering learners an immersive and interactive training experience, while sensing modalities provide an opportunity for the integration of real-time feedback. By leveraging computer vision and sensing technologies, these systems can analyze the learner's movements and provide immediate feedback on the accuracy and effectiveness of their CPR technique. This real-time feedback enables the learner to make adjustments on the spot, improving the quality and consistency of their compressions. This approach enhances skill acquisition, retention, and performance by overlaying instructional guidance and visual cues onto the learner's view of a CPR scenario [1, 10, 34].

While existing studies on AR-based CPR training have shown promise in enhancing training experiences and outcomes [3, 11, 13, 16], there is still a need for further research to fully explore the effectiveness of AR in addressing the current limitations of traditional CPR training methods. Although AR has demonstrated potential in providing immersive and interactive training environments, its specific impact on skill acquisition, retention, and performance during CPR training is an area that requires more investigation. Additionally, the design and implementation of AR-based CPR training systems can vary, and it is essential to evaluate the effectiveness of different approaches to optimize their educational value. Furthermore, understanding the potential benefits and challenges associated with incorporating AR technology into CPR training can inform the development of evidence-based guidelines and best practices for its integration.

In this paper, we introduce RescuAR, a self-directed AR-based CPR teaching and training system that addresses the gap between theoretical knowledge and practical applications in CPR education. RescuAR leverages AR technology to provide real-time feedback on performance during CPR training, enhancing the acquisition and retention of crucial CPR skills. We provide a comprehensive overview of the system, covering its pre-design stage, implementation process, and post-development evaluation. The design and implementation of RescuAR are carefully tailored to meet the specific needs of CPR

RescuAR: A Self-Directed Augmented Reality System for Cardiopulmonary Resuscitation Training

Hamraz Javaheri[1(✉)], Agnes Gruenerbl[1], Eloise Monger[2], Mary Gobbi[2], Jakob Karolus[1,3], and Paul Lukowicz[1,3]

[1] DFKI GmbH, Kaiserslautern, Germany
Hamraz.Javaheri@dfki.de
[2] University of Southampton, Southampton, UK
[3] University of Kaiserslautern-Landau, Kaiserslautern, Germany

Abstract. In recent years, the adoption of augmented reality (AR) technology for healthcare education has gained significant attention. Especially in life-critical situations, such as cardiopulmonary resuscitation (CPR) where sufficient medical training is essential and traditional methods are often limited due to availability constraints. We present RescuAR, a self-directed AR-based CPR training system enhancing CPR skill acquisition and retention by leveraging immersive AR experiences and real-time feedback using sensing modalities.

RescuAR was designed and implemented as a self-directed AR application based on survey findings with 11 healthcare professionals, incorporating both theory and practice phases. To evaluate the effectiveness of RescuAR, a randomized controlled user study was conducted involving $n = 43$ participants, including nurse students and laypeople. The experimental group used RescuAR for CPR training, while the control group underwent traditional teaching and training sessions. The results of the user study revealed that RescuAR significantly improved the overall effective CPR performance, surpassing the outcomes achieved through traditional teaching methods. In conclusion, RescuAR's self-directed and autonomous approach to CPR training shows promising results in improving CPR performance and has the potential to transform CPR education.

Keywords: Augmented-Reality · Cardiopulmonary Resuscitation · Self-Education

1 Introduction

Cardiopulmonary resuscitation (CPR) plays a crucial role in saving lives during cardiac arrest, a medical emergency with high mortality rates. The timely and effective administration of CPR significantly increases the chances of survival [9]. Therefore, it is vital to ensure that individuals are well-trained in this life-saving technique. Traditional CPR training methods, such as classroom-based instruction and mannequin practice, have been the cornerstone of CPR education for decades. However, these methods

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 166–185, 2024.
https://doi.org/10.1007/978-3-031-59717-6_12

31. Webster, L., Spiro, R.F.: Health information technology: a new world for pharmacy. J. Am. Pharm. Assoc. **50**(2), e20–e34 (2010). https://doi.org/10.1331/JAPhA.2010.09170
32. WHO, W.H.O.: Non communicable diseases (2021). https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases Accessed 14 Feb 2022
33. Zhang, H., et al.: A mobile health solution for chronic disease management at retail pharmacy. In: 2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom), pp. 1–5 (2016). https://doi.org/10.1109/HealthCom.2016.7749455

14. Lapão, L.V., et al.: EHealth services for enhanced pharmaceutical care provision: from counseling to patient education. In: 2013 IEEE 2nd International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–7 (2013). https://doi.org/10.1109/SeGAH.2013.6665308
15. Lee, K.P., Hartridge, C., Corbett, K., Vittinghoff, E., Auerbach, A.D.: Whose job is it, really? physicians, nurses and pharmacists perspectives on completing inpatient medication reconciliation. J. Hosp. Med. **10**(3), 184–6 (2015). https://doi.org/10.1002/jhm.2289
16. Lofland, J., Snow, D.A., Anderson, L., Lofland, L.H.: Analyzing Social Settings: A Guide To Qualitative Observation And Analysis, 4th edn. Wadsworth Publishing, Belmont, CA, USA (2005)
17. Martin, A., et al.: The evolving frontier of digital health: opportunities for pharmacists on the horizon. Hosp. Pharm. **53**(1), 7–11 (2018). https://doi.org/10.1177/0018578717738221
18. Martins, S., Costa, F.A.D., Caramona, M.: ImplementaÇÃo de cuidados farmacÊuticos em portugal, seis anos depois. Rev. Port. de Farmacoterapia **5**(4), 4–12 (2015). https://doi.org/10.25756/rpf.v5i4.38
19. Mercer, K., et al.: Physician and pharmacist medication decision-making in the time of electronic health records: mixed-methods study. JMIR Hum. Factors **5**(3), e24 (2018). https://doi.org/10.2196/humanfactors.9891
20. Mossialos, E., et al.: From retailers to health care providers: transforming the role of community pharmacists in chronic disease management. Health Policy **119**(5), 628–639 (2015). https://doi.org/10.1016/j.healthpol.2015.02.007
21. Mullins, A.K., et al.: Physicians and pharmacists use of my health record in the emergency department: results from a mixed-methods study. Health Inf. Sci. Syst. **9**(1), 1–10 (2021). https://doi.org/10.1007/s13755-021-00148-6
22. MURAL: Mural (2022). https://www.mural.co/
23. Murero, M.: E-prescribing: the rise of socio-tech-med micronetworks of care during the COVID-19 pandemic. Salute E SocietÀ **XX**(suppl. 2), 104–118 (2021). https://doi.org/10.3280/SES2021-002-S1007
24. Puspitasari, H.P., Aslani, P., Krass, I.: Challenges in the management of chronic noncommunicable diseases by indonesian community pharmacists. Pharm. pract. **13**(3), 578 (2015). https://doi.org/10.18549/PharmPract.2015.03.578
25. Rosenthal, M.M., Breault, R.R., Austin, Z., Tsuyuki, R.T.: Pharmacists self-perception of their professional role: insights into community pharmacy culture. J. Am. Pharm. Assoc. **51**(3), 363–368a (2011). https://doi.org/10.1331/JAPhA.2011.10034
26. Shane-McWhorter, L., et al.: Pharmacist-provided diabetes management and education via a telemonitoring program. J. Am. Pharm. Assoc. **55**(5), 516–526 (2015). https://doi.org/10.1331/JAPhA.2015.14285
27. Spradley, J.P.: Participant Observation. Holt, Rinehart and Winston (1980)
28. Storni, C.: Multiple forms of appropriation in self-monitoring technology: reflections on the role of evaluation in future self-care. Int. J. Hum. Comput. Interact. **26**(5), 537–561 (2010). https://doi.org/10.1080/10447311003720001
29. van de Pol, J.M., Geljon, J.G., Belitser, S.V., Frederix, G.W., Hövels, A.M., Bouvy, M.L.: Pharmacy in transition: a work sampling study of community pharmacists using smartphone technology. Res. Social Adm. Pharm. **15**(1), 70–76 (2019). https://doi.org/10.1016/j.sapharm.2018.03.004
30. Waszyk-Nowaczyk, M., et al.: Cooperation between pharmacists and physicians - whether it was before and is it still ongoing during the pandemic? J. Multidiscip. Healthc. **14**, 2101–2110 (2021). https://doi.org/10.2147/jmdh.s318480

Acknowledgements. This study was funded by the project ConnectedHealth (n.º 46858), supported by the Competitiveness and Internationalisation Operational Programme (POCI) and Lisbon Regional Operational Programme (LISBOA 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

References

1. Anderson, S.: The state of the world's pharmacy: a portrait of the pharmacy profession. J. Interprof. Care **16**(4), 391–404 (2002). https://doi.org/10.1080/1356182021000008337
2. Aungst, T.D., Miranda, A.C., Serag-Bolos, E.S.: How mobile devices are changing pharmacy practice. Am. J. Health-Syst. Pharm. **72**(6), 494–500 (2015). https://doi.org/10.2146/ajhp140139
3. Baldo, D., Benelli, G., Pozzebon, A., Sesto, R.: The fides project: a pharmacy toolbox to allow healthcare decentralization. In: 2011 E-Health and Bioengineering Conference (EHB), pp. 1–5 (2011)
4. Barlow, J., Wright, C., Sheasby, J., Turner, A., Hainsworth, J.: Self-management approaches for people with chronic conditions: a review. Patient Educ. Couns. **48**(2), 177–187 (2002). https://doi.org/10.1016/S0738-3991(02)00032-0
5. Burgess, H., et al.: The sticky notes method: adapting interpretive description methodology for team-based qualitative analysis in community-based participatory research. Qual. Health Res. **31**(7), 1335–1344 (2021)
6. Craddock, D.S., Hall, R.G.: Pharmacists without access to the EHR: practicing with one hand tied behind our backs. Innovations pharm. **12**(3), 16 (2021). https://doi.org/10.24926/iip.v12i3.4141
7. Crilly, P., Kayyali, R.: A systematic review of randomized controlled trials of telehealth and digital technology use by community pharmacists to improve public health. Pharmacy **8**(3), 137 (2020). https://doi.org/10.3390/pharmacy8030137
8. Fitzpatrick, G., Ellingsen, G.: A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas. Comput. Support. Coop. Work (CSCW) **22**(4), 609–665 (2012). https://doi.org/10.1007/s10606-012-9168-0
9. van der Gaag, M., Heijmans, M., Spoiala, C., Rademakers, J.: The importance of health literacy for self-management: a scoping review of reviews. Chronic Illn. **18**(2), 234–254 (2022). https://doi.org/10.1177/17423953211035472
10. Gheorghiu, B., Hagens, S.: Measuring interoperable EHR adoption and maturity: a canadian example. BMC Med. Inf. Decis. Making **16**(1), 1–7 (2016). https://doi.org/10.1186/s12911-016-0247-x
11. Gregório, J., Lapão, L.V.: Uso de cenários estratégicos para planeamento de recursos humanos em saúde: o caso dos farmacêuticos comunitários em portugal 2010–2020. Revista Portuguesa de Saúde Pública **30**(2), 125–142 (2012). https://doi.org/10.1016/j.rpsp.2012.12.003
12. Howard, J., et al.: Exploring the barriers to using assistive technology for individuals with chronic conditions: a meta-synthesis review. Disabil. Rehabil. Assist. Technol. **17**(4), 390–408 (2022). https://doi.org/10.1080/17483107.2020.1788181
13. Hughes, C.A., Guirguis, L.M., Wong, T., Ng, K., Ing, L., Fisher, K.: Influence of pharmacy practice on community pharmacists integration of medication and lab value information from electronic health records. J. Am. Pharm. Assoc. **51**(5), 591–598 (2011). https://doi.org/10.1331/JAPhA.2011.10085

their patients and manage medication data confidentially, so having a profile for patients would be a logical next step.

Provide a direct communication channel between pharmacists and clinicians. The fieldwork showed issues in communication between pharmacists and clinicians. Without a direct communication channel, patients had to wait or come back another time to the pharmacy, each time there seemed to be a medication interaction or an error in prescription. Collaboration between pharmacist and clinician was also hindered, and there was no way for clinicians to rely on the pharmacist besides what they could intuit from a prescription. With a direct software communication channel, pharmacists would be able to: (i) support patient education and training, (ii) share observations with clinicians about patient adherence or (side)effects, (iii) recommend adjustments in medication taking into consideration health issues or the daily habits of the patient, and (iv) ask for alternative medications, e.g., when a medication is out of stock. While clearly important for patient care, having this effective communication can also be valuable in defining the role of the pharmacist as a support for doctors, helping them support patient self-care and triage of acute illness episodes.

Enable measurements made at the pharmacy to be shared with clinicians. It was clear from the fieldwork that pharmacists take a number of precautions to reach measurements with clinical quality. Currently, values measured at the pharmacy are shared with clinicians using paper slips (Fig. 1), which can be easily lost and even require additional note-taking at the clinician's office. In case the measurement process is digitised, the measurements could be uploaded directly to the patient profile or an alternative option that can be easily shared with the clinician. Having these measurements made at the pharmacy and with the support of the pharmacist ensures measurement quality, avoiding false concerns and doctor visits, or delay in care.

6 Conclusion

This study was conducted to understand current practices and challenges faced by pharmacists to inform the design of solutions that could play an important role in supporting their chronic patients. Our findings indicate that while pharmacists are a key actor in helping chronic patients manage their diseases, they lack the information and communication with clinicians to better support their decisions and provide more services and better care to their patients. Technology can be seen as a potential solution, but it needs to address these professionals' real needs, and some key features must not be left out. The derived implications for design intend to summarise these action points, which should be implemented with Participatory Design projects that involve all relevant stakeholders.

Moreover, there is a need for a change of pharmacists role in healthcare and their relationship with clinicians, patients and in particular chronic patients. While this change seems to be clear for our participants, other policy, legal and organisational aspects should be approached and discussed in future studies.

[20,24]. Overall, pharmacists consider they should be able to have more information about the patients and their conditions, as it is their role not only to provide medication but also to check for medication interactions and support medication adherence.

Since pharmacists have a close relationship with patients, they could play an important role in communicating symptoms' evolution and adherence data to clinicians. As identified in prior work [19], clinicians often lack information that could allow them to better monitor patients, even without direct contact. Pharmacists dedicate their efforts to ensure that chronic patients have the opportunity to monitor their medication and other health parameters not only onsite at the pharmacy but also when helping patients that have difficulties performing their measurements at home an issue previously identified in prior work [28].

It became clear during this study that, as observed before [19], one of the most prominent challenges currently faced by pharmacists is the lack of direct communication with clinicians. Pharmacists envision a closer collaboration with clinicians, to support a better understanding and common agreement of the roles each will play in patient care, with doctors being in charge of the diagnosis, with pharmacists being involved in prescription and treatment adjustments.

It should be noted that our work is limited due to the localised nature of the fieldwork. We involved pharmacies from one city, and it is possible that other pharmacies in different locations (e.g. rural settings) report different practices. We compared our findings with the literature to overcome the localisation issue, but fieldwork in other locations would be required to further validate our findings.

5.1 Implications for Design

According to our fieldwork, technology can support collaboration between pharmacists, patients and clinicians to improve chronic care. As such, we derived a set of implications for the design of such technologies to ensure that they answer to the current challenges faced by pharmacists.

Enable pharmacies to keep a personalized profile of their patients. Pharmacies play a very important role in the measurement and monitoring of patients' measurements. By providing expert feedback on patients' values, pharmacists not only advise patients but also support screening of aggravations or further health issues. Having the possibility to keep track of measurements should enable pharmacists to more efficiently know the baseline of values for a specific patient, enabling them to detect minor issues before they aggravate. During the fieldwork, we also understood that pharmacists have access to the record of client purchases and use this list to screen for medication interactions. However, the list of purchases for each client may include products bought for others, which can complicate this process. With this in mind, pharmacists suggested having an individual profile for each patient, especially chronic patients taking more medication simultaneously. The data management of patient profiles will need to be carefully designed, but pharmacists already have an ethical duty to

> **Pharmacist 1**: "There could be a triangulation between clinicians, pharmacists, and patients, thinking about the benefits that those interactions could bring, even though dealing with the challenges of sensitive data protection could occur".

A more appropriate approach would be, as envisioned by Pharmacist 1, to have a more direct communication channel between clinicians, pharmacists and patients, where information could be safely shared. For this pharmacist, there are clear benefits in having access to this information despite its sensitivity and more importantly, it can avoid the long waits or misunderstandings that currently occur. As an example, they mention a mobile service connecting doctors, pharmacists and patients. Another participant referred that, in the case of patients with yearly medication subscriptions, the pharmacist could report observations about how patients are taking or reacting to the medication, which could inform medication adjustments. Another pharmacist described that, ideally, there should be an initial appointment in the pharmacy, where the pharmacist would get to know the patient's clinical conditions, medications, and medical recommendations. Currently, pharmacists only see patients when they purchase medication or perform measurements, which does not allow them to perform the role they envision. Another pharmacist emphasised that being able to accompany patients more would be ideal because pharmacists are the first healthcare professional which patients appeal to.

5 Discussion

The main goal of this study was to understand how pharmacists support patients with chronic diseases in their management, and the interactions with healthcare professionals, to inform future technology design. While performing observations in pharmacies, we noted that, in accordance to previous studies [14,29], the work carried out by pharmacists exceeds medication dispensing and includes not only contact with the public but also a significant amount of time spent with administrative tasks. It became clear that medication dispensing is a far more complex task than simply following a prescription. The process includes instructing patients, making sure they clearly understand the instructions and have the conditions to follow them while screening for possible medication interactions. As such, even when reducing the role of pharmacists as medication dispensers, as mentioned by [25], it is important to acknowledge the amount of expertise entailed.

With their current lack of involvement in the prescription phase, pharmacists are left in a position where they can provide the prescribed medication but are often unsure if they are providing the best available option for the patient, as also reported by [6]. In the opinion of our participants, there should be a clear legal definition of the pharmacist role, and a cultural shift in health services where collaboration between different services and professionals is valued in practice. These barriers and necessary changes are also mentioned in previous literature

closely accompany. Additionally, a pharmacist shared that when communicating with emergency services, their knowledge allows them to explain patients' condition better than the patients themselves or other person outside the healthcare field. This shows a strong commitment to patients, and an active role in enabling the healthcare system to work.

Interviewed pharmacists considered that their measurement devices should save patients' values, similar to a prior study [33]. Frequently, sometimes twice a day (morning and afternoon), patients go to the pharmacy to measure blood pressure due to medical instructions and preserve a piece of paper (Fig. 1) to present to the doctor to monitor their status. However, if the machine could record and save patients' values, it could facilitate communication or collaboration with clinicians. It could also enable pharmacists to know the usual values for a specific patient. According to our participants, pharmacies could have regular patients' informed consent, for example, to save and share data within a cross-disciplinary platform accessible to both pharmacists and clinicians.

4.3 Discussing Prescription Issues with Clinicians

In some rare occasions, pharmacists spot issues with the prescription that need to be discussed with clinicians. Examples include potential interactions between medication, very high doses prescribed, or unconsidered patient characteristics that might make the prescription unsafe. In these situations, pharmacists may contact clinicians to discuss potential issues and devise an adjusted medication plan, however, communication channels are far from ideal.

> **Pharmacist 1**: "The patient needs to wait one or two hours at the pharmacy before I can get information that I need from the clinician".

Since pharmacists do not usually have direct phone numbers to healthcare units, they are usually left with general phone numbers, which may not prioritise their question. This means that, as stated by Pharmacist 1, waiting periods can be long before the pharmacist can contact the clinician, which is not desirable for neither the patient nor the pharmacist. When direct communication is not possible, pharmacists mainly give the information to their patients so they can discuss it with their clinicians. Still, pharmacists find this compromise unsatisfactory because some information can be lost, misunderstood or misinterpreted.

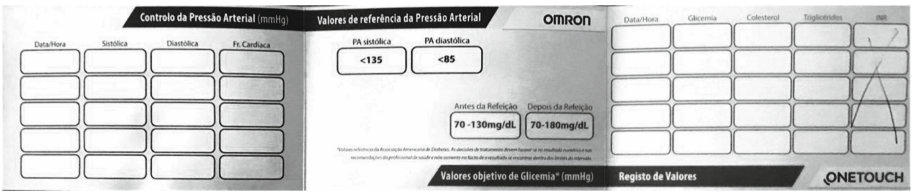


Fig. 1. Paper slip for recording blood pressure measurements currently in use in one of the pharmacies observed.

was correct, or because they had abnormal values and wanted to know what they could do about them. To illustrate the measurement practices occurring in pharmacies, we present a vignette of measuring blood pressure in a "self-service" blood pressure monitor machine at Pharmacy 1.

> *Pharmacist positions the patient arm in the machine and adjusts their back. Pharmacist enters the coins in the machine. Pharmacist and patient await the measurement. When the machine ends, pharmacist picks up the receipt paper with the measurement.*
> **Pharmacist 2**: "Everything is fine. Systolic is at 16; lets see if it keeps like that [or if it lowers]. Come back in 2-3 days again to see".
> **Patient**: "Is drinking coffee [before coming] bad?"
> **Pharmacist 2**: "Try avoiding coffee and then we test [to see if values change]".

As shown in the example above, pharmacists were the ones responsible for the setup of the blood pressure monitor device at pharmacy. They adjusted the position of the bench to fit the patient, they helped place the patient's arm into the inflatable cuff, and even operated the device's software. After choosing whether to perform a blood pressure measure, a weight measure, or both, pharmacists also entered the coins of the patient into the device. When asked about why pharmacists took such an active role in the setup of the blood pressure monitor device, participants explained that it was the way to obtain quality measurements. Patients were likely to have elevated blood pressure if they saw an error in the machine, so pharmacists were careful with the setup, to make sure the chances of errors by inappropriate measurement conditions stayed minimal. At the peak of COVID-19, pharmacists supported patients in performing a second measurement, adjusting the arm in the cuff or any other condition that could have caused the error in their perspective, but prior to COVID-19 they would take the patient to a separate room, sit them down, give them some minutes to relax, and only then would they perform a manual blood pressure measurement. These conditions supported the acquisition of measurements that were faithful to the patients' state and thus had clinical value.

When the machine performs the measurement, pharmacists stay close to the patient for noticing errors, might they arise. Once measurements are finished, pharmacists picked up the paper slip from the machine, read the value out loud to the patient, and explained to the patient if it was positive or negative. In case of abnormalities, pharmacists inquired patients about circumstances that could have elevated their blood pressure, including anxiety, salty food, or the recent start of a new medication. If the values obtained were concerning, they might ask the patient to return later to make a new measurement or suggest them to go see their doctor. Moreover, if values deviated considerably from standard values, they would call patients' relatives or emergency services. When asked if calling family members or emergency services was part of their responsibilities, one pharmacist explained that they had an ethical duty as healthcare professionals to care for their population, most especially in acute cases of regular clients they

lack complete information about their condition and treatments. Looking at prescriptions, pharmacists can get an idea of the reason for taking the medication, but it is not guaranteed, because, for example, a medication for diabetes could be used for losing weight. Another tool they use is the records available in the information system that can show them medication previously purchased by patients but this information is often unreliable: a patient can have on record medications that were purchased by them but not for them or they may have purchased medications in another pharmacy. The absence of data enables a lack of trust and extends the process of achieving trustworthy information, but participants argue that if they could access more data, they would have more confidence in the credibility of the information they convey to patients.

Pharmacists also supported their patients in successfully taking medication, mobilising resources that could be useful to them. One example we observed while at the pharmacy was how a pharmacist offered the patient to take a pill box for putting their medication. Noticing that the patient was having difficulties in remembering the medication they took, the pharmacist thought about potential solutions for the issue and decided to suggest the patient to try using a pill box. Even though thinking about where the patient would place the medication exceeded the pharmacist's role, she intuited that the patient would benefit from having a medication box which they could fill in daily and know when they had taken a certain medication.

The work of pharmacists is further supported by long-term relationships with their patients.

> **Pharmacist 3**: "There are affinities with certain clients. (...) We [pharmacists] create a friendly relationship, some [pharmacists] with more, others with less, but happens with all colleagues."

A trusted pharmacist will often become the preferred professional for a specific patient, and be the person with who they share illness episodes, questions, or even the news and pictures from their family. While pharmacists try to avoid people being attended only by one pharmacist, when pharmacists have close relationships with patients they are able to pay more attention to the patient's general health and detect acute illness episodes, which can be very useful.

4.2 Performing Measurements at the Pharmacy

Pharmacies are important health parameter measurement sites for patients with chronic conditions. They possess the devices to measure blood pressure, weight, and, in some cases, blood testing equipment that can serve to understand the state of the patient's cholesterol or diabetes. In addition to equipment, making measurements at the pharmacy has the added benefit of having the pharmacist operate measurement devices or, at least, provide feedback on the values. Most patients that made measurements at the pharmacy did not have a device for performing the measurement at home. Patients who had their own device sometimes went to the pharmacy to check if the measurement of their device

Technology-wise, Portuguese pharmacies currently have access to software that registers the medication sold to customers, and identifies potential medication interactions. Despite its recognized usefulness to pharmacists, the system does not take into account that customers frequently buy products for members of their household, registering these under the same customer id. As such, medication interactions or dosage alerts may result from wrongful information. Moreover, customer data is separated for different pharmacies, which means that pharmacists can only act on purchase information from their own pharmacy.

4 Results: Current Practices of Managing Chronic Patients

Pharmacists are highly sought-after mediators, who repeat medical advice, listen to patients' concerns, and help them reflect on medication side-effects and interactions. They do not replace primary care services, but have a fundamental role in this ecosystem, which is sometimes underestimated. As we observed, pharmacists' role entails teaching the treatment (ongoing or about to start), explaining posology, following medical attention, considering medication interactions and medical exams, and giving advice. To exemplify the role of pharmacists, one technical director said that pharmacists often teach asthmatics how to use the expansion chamber (asthma inhaler), advise patients about a balanced diet and associated medication, or in cases of constipation to "drink water as treatment".

4.1 Support Medication Starting and Correct Usage

Pharmacists spend a considerable amount of time dispensing medication to both the general public and chronic patients. The first step is usually to understand the products that patients want to take. In a country that has adopted electronic prescription, it is common for patients to hand their smartphones to pharmacists so they can access the prescription dispensing codes (similar practices were described in [23]). We also observed patients showing medication packages or photographs of medication packages to indicate their preferred medication brand.

The most relevant chunk of the time in dispensing medication is invested in explaining how to properly take the medication. According to our participants, patients have many doubts about their medication, as doctors spend less time explaining how to take medication. As a result, patients resort to the support of their pharmacist, who seems to have more time or availability to address their doubts. An additional issue has to do with the health literacy of the patients. Clinicians often use concepts that patients do not understand which can lead to ignoring important information. For example, one pharmacist referred that clinicians sometimes ask patients to avoid anti-inflammatory medication, but that it is extremely common for patients not to know what those medications are, and thus pharmacists provide this information to support patients.

Another important task of pharmacists while dispensing medication is to screen potential medication interactions. During this search, pharmacists perform a great deal of "guessing work", as patients, which usually are older adults,

with patients. We chose to involve pharmacists with different levels of experience and from different settings, to gain access to diverse experiences and backgrounds. The interview guide touched on three main topics: (1) Interactions with chronic patients, (2) Dealing with regular clients, and (3) Role of the pharmacy and pharmacist in the healthcare system. We also inquired participants about demographics, formal education, and previous experience. Before starting the interview, participants received information about the study and data privacy.

In total, 11 participants were interviewed (7 female, 4 male). Three interviews were conducted via videoconference and recorded. The remaining interviews were conducted face-to-face, inside the pharmacies, and were not recorded. Each interview lasted between 30 and 60 min, and the participants were all pharmacists, with responsibilities of customer attendance and/or technical direction of the pharmacy. Experience varied from 1 to over 30 years of experience.

3.3 Recruitment and Ethics

The four pharmacies were recruited through Associação Nacional de Farmácias, the portuguese association of pharmacies, who called potential pharmacies from a convenience sample, considering variety in terms of size, context, and innovation attitude. All pharmacies were based in Porto, Portugal, and one pharmacy had more than one physical site. Once pharmacies agreed to take part in the study, researchers called pharmacies' technical directors for arranging visits. Technical directors introduced the research team to the pharmacists, be them on site or online. In some instances, pharmacists were asked to indicate colleagues from their pharmacy to participate in the interviews. Participants were all volunteers and received no monetary compensation.

All participants provided informed consent after receiving information about the project, goals of the study, data management and security.

3.4 Portuguese Healthcare Context

In Portugal there is a national health service - Serviço Nacional de Saúde (SNS) – that is the main health service, based on universal and equal health access for people living in the country. Community pharmacies however are not part of the SNS; they are privately owned, subject to government-issued requirements (e.g. staff must be comprised of at least two pharmacists) but not directly connected to SNS entities such as hospitals or primary care clinics. There is also a national pharmacy association – Associação Nacional de Farmácias (ANF) that represents pharmacy owners, and whose mission is to support pharmacies and initiatives that value their services. In 1999, ANF created a department to develop pharmaceutical care programs that promoted the integration of pharmacists in patient care, monitoring and follow-up, thus enriching pharmacists' role. The first pilot was launched in 2001, focused on supporting the care of patients with diabetes, hypertension and COPD, and since then, more pharmacies have adhered to these programs [18].

to explore and understand the characteristics these solutions must exhibit to successfully implement them in a useful and sustainable way.

Pharmacists can be important in chronic care management, but little is known about how and which tools can support this. With our study we expect to better understand how pharmacists support chronic patients within the Portuguese context and derive the necessary recommendations for developing high-impact solutions that contribute to the enrichment of pharmacists' role and facilitate teamwork with physicians, ensuring the best care for their patients.

3 Methods

We conducted an observation and interview study to understand pharmacies' role in managing patients with chronic conditions. The ethnographic fieldwork was conducted between May and July 2021 by six researchers, who were grouped in pairs to observe pharmacies and interact with pharmacy workers (pharmacists and pharmacy assistants) in different settings. Notes, photographs, drawings, and interview transcripts were shared and discussed among the research team. The analysis was supported by the Affinity Mapping method [5], whereby six researchers in the team summarised, grouped, and discussed the main insights of the study individually as well as in three group sessions, around a digital whiteboard supporter by Mural [22] software.

3.1 Observations

We used non-participant observation at pharmacies [27], complemented with informal interviews with pharmacists. Observations had three main *foci*: pharmacist-patient interactions, health parameter measurement, and interactions with the existing software. Researchers did not directly interact with pharmacy clients.

The observation sessions, which ranged from 1 h to 3 h, were always conducted by two researchers simultaneously and took place in five separate locations. During observations, researchers chose different locations in the pharmacy, including being next to pharmacists, behind the counter, or standing next to shelves or near clients being attended to. In total, researchers spent 14 h in observation sessions, with some sites receiving multiple observation sessions. Data from observations were collected mainly using fieldnotes, occasionally complemented with photographs and drawings.

3.2 In-Depth Interviews

To understand pharmacists' perspectives about their role in supporting patients with chronic conditions, we conducted in-depth interviews [16]. The interviews were qualitative, and semi-structured, to touch on specific topics while giving space for participants to bring other topics to the table. We recruited pharmacists, pharmacist assistants, and technical directors, as they contacted directly

dispensed prescriptions, preparing prescriptions for reimbursement issues, and meetings with vendors and salespersons [14]. A study conducted in the Netherlands [29] found similar results. The time spent with secondary tasks has burdened pharmacists and prevented them from expanding their healthcare services. Technological solutions for medicine dispensing could optimise the process thus reducing pharmacists' effort and increasing available time [14].

2.2 Technology in Community Pharmacies

The community has created several technologies for pharmacies (for a review see [7,31]). Mobile applications can be implemented in pharmacy practice for varied purposes such as clinical references, order processing, communication or patient engagement [2]. Patients with chronic conditions are one of the groups that could benefit the most from the use of these technologies and a closer relationship with pharmacists, particularly with the use of condition monitoring devices (e.g. glucometers, blood pressure monitors, etc.) that can provide valuable information to equip pharmacists to assist better patients with diabetes, COPD, or congestive heart failure [17]. Different approaches to chronic disease monitoring have been successfully explored such as telemonitoring for diabetes management and education [26], or a platform with a set of devices measuring health parameters to be positioned inside the pharmacies [3]. Another mobile solution was used for diabetic and hypertension patients connected to monitoring devices controlled by the pharmacist that store measurements taken at the pharmacy and provide patients with relevant information to manage their disease while facilitating communication between patients and pharmacists [33].

While there seems to be a consensus that a technological approach to the relationship between pharmacists and patients could benefit both, having a stronger focus on the patient and their clinical status also creates an urgent need for more and better interactions between pharmacists and physicians. Both pharmacists and physicians confirm that multidisciplinary teams can improve patient care and treatment efficacy, but there is still a need to modify infrastructures, agree upon goals and educate healthcare workers to fully take advantage of such partnership [30]. A Canadian study with 19 pharmacies and nine medical clinics observed limited communication and collaboration between primary care doctors and pharmacists, with pharmacists missing prescription data and physicians missing data on adherence [19]. Even in hospitals, where pharmacists and clinicians collaborate regularly, professionals lack agreement about their specific roles and responsibilities in the medication reconciliation process, resulting in incomplete, inefficient, and duplicate work around medication regimens [15].

Technological solutions, particularly Electronic health records (EHR) can be a potential tool to aid communication between clinicians and pharmacists [19]. Countries like Canada [10] and Australia [21] are already using these systems but there are others such as the United States where the only information available to most pharmacists when dispensing medication is essentially the prescription, which is not sufficient to make informed decisions for patients [6]. With all the advantages that EHRs could bring to pharmacy practice [13] it is then paramount

Despite the recognised value that Information and Communication Technologies (ICT) bring to healthcare, there is still a gap when it comes to studies on the use of ICT in pharmacies [8]. Our work aims to contribute to understanding community pharmacists' interactions with chronic patients, including what tools they have access to, how they gather important disease-related data, and how they intervene in patient care. To this end, we conducted an ethnographic informed study in four pharmacies, performing observations and interviews with eleven pharmacists with different roles and experience levels. This approach allowed us to gain insights into the role of the pharmacy in the community and in the healthcare system, as well as the role of the pharmacist in supporting chronic care management.

This paper reports on the outcomes of this qualitative study and the derived design implications. Our results emphasise the importance of pharmacists in chronic care and detail how their role in medication management and measurements is essential to chronic patients. Moreover, it is also clear that pharmacists need improved tools to support their work and our results may better inform the development of solutions targeting pharmacies and their patients.

2 Background

2.1 The Role of Community Pharmacists

Pharmacists play a vital role in the healthcare system due to their close proximity to patients. They are experts in medication and are responsible for ensuring the safe, effective and rational use of drugs. The connection of pharmacists to medication development, supply and management is widely recognised, however, a shift in pharmacists' responsibilities is already taking place and evolving towards a more patient-centric approach [1]. In some countries, pharmacists already take roles previously exclusive to nurses or doctors, including supporting blood pressure, glucose and cholesterol measurement, pregnancy testing, or providing smoking cessation advice and diabetes guidance [1]. This shift was motivated by population ageing, an increase in number of chronic patients, shortages in healthcare professionals, COVID-19 demands, but it was also a necessary step to ensure the sustainability of the profession itself because medication dispensing can be automated [11].

Existing barriers to more patient-centred pharmacists include pharmacists' self-perception as "dispensers of medication" and not patient-centred practitioners coupled with the business-driven culture of pharmacies [25]. Other studies indicate that pharmacists are not used as public health professionals because of a negative attitude towards pharmacists' role in patient care, pharmacy education, standards, government policies [24], lack of interprofessional care, inadequate compensation models, and lack of a shared vision for pharmacy services [20]. Results from a Portuguese study with four pharmacies suggest that time management can also be a barrier to the optimal use of pharmacists' skills: while 50% of pharmacists' time was used in interactions with customers, close to 38% was spent ordering and storage of medicines, checking for errors in the

Patient-Pharmacist Interactions in Chronic Care: A Qualitative Study and Implications for Design

Ana Vasconcelos, Joana Couto Silva, Ruben Moutinho, Fernando Ricaldoni, Ana Correia de Barros, and Francisco Nunes(✉)

Fraunhofer Portugal AICOS, R. Alfredo Allen 455/461, 4200-135 Porto, Portugal
{Ana.Vasconcelos,Joana.Silva,Ruben.Moutinho,Fernando.Ricaldoni,
Ana.Barros,Francisco.Nunes}@fraunhofer.pt

Abstract. Chronic patients are often asked to perform measurements as part of their self-care. Some patients make measurements at home, but others resort to their local pharmacy for information and support. However, there is a shallow understanding of the role of pharmacists and pharmacies in chronic care management, which may hinder the development of tools to support patient care. To better understand the work carried out at community pharmacies for chronic care, and inform the design of these systems, we conducted an ethnographic informed study. We observed four community pharmacies and interviewed eleven pharmacists. Results show that pharmacists are essential in providing patients with information regarding their medication and support in health measurements. However, their work is restricted by a general lack of information about the patient and limited collaboration with other clinicians. Drawing on the insights from this work, we derived three implications for the design, including developing software for pharmacies that keeps track of patient measurements and shares them with doctors, and creating a pharmacist-doctor communication channel for enabling medication adjustments.

Keywords: chronic care · health measurements · pharmacy · pharmacists · chronic patients · observations · interviews

1 Introduction

Chronic conditions such as Diabetes, Hypertension, or Chronic Obstructive Pulmonary Disease (COPD) are the main causes of mortality, representing 71% of all deaths [32]. Chronic patients need to frequently monitor their condition and engage in self-care [4,9,12], however, some patients might experience difficulties transitioning to or managing autonomously, requiring the help of healthcare professionals. Given community pharmacies' proximity to patients, pharmacists can be crucial in monitoring chronic patients, releasing the burden from often overworked doctors or nurses.

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 151–165, 2024.
https://doi.org/10.1007/978-3-031-59717-6_11

Pervasive Health for Carers

19. Cochran, A.R., Raub, K.M., Murphy, K.J., Iannitti, D.A., Vrochides, D.: Novel use of REDCap to develop an advanced platform to display predictive analytics and track compliance with enhanced recovery after surgery for pancreaticoduodenectomy. Int. J. Med. Inform. **119**, 54–60 (2018). https://doi.org/10.1016/j.ijmedinf.2018.09.001
20. Côté, M., Lamarche, B.: Artificial intelligence in nutrition research: perspectives on current and future applications. App. Physiol. Nutr. Metab. (2021). https://doi.org/10.1139/apnm-2021-0448
21. Schröder, H., et al.: A Short screener is valid for assessing mediterranean diet adherence among older Spanish men and women. J. Nutr. **141**(6), 1140–1145 (2011). https://doi.org/10.3945/jn.110.135566
22. Martínez-Larrad, M.T., et al.: Prevalencia del síndrome metabólico (criterios del ATP-III). Estudio de base poblacional en áreas rural y urbana de la provincia de Segovia. Med. Clin. (Barc.) **125**(13), 481–486 (2005). https://doi.org/10.1157/13080210
23. Brauer, P., et al.: Key process features of personalized diet counselling in metabolic syndrome: secondary analysis of feasibility study in primary care. BMC Nutr. **8**(1) (2022). https://doi.org/10.1186/s40795-022-00540-9
24. Tang, H., et al.: Randomised, double-blind, placebo-controlled trial of Probiotics to Eliminate COVID-19 Transmission in Exposed Household Contacts (PROTECT-EHC): a clinical trial protocol. BMJ Open **11** (2021). https://doi.org/10.1136/bmjopen-2020-047069
25. Nourani, A., Ayatollahi, H., Solaymani-Dodaran, M.: A clinical data management system for diabetes clinical trials. J. Healthc. Eng. **2022** (2022). https://doi.org/10.1155/2022/8421529
26. Odukoya, O., et al.: Application of the research electronic data capture (REDCap) system in a low- and middle income country– experiences, lessons, and challenges. Health Technol (Berl) **11**(6), 1297–1304 (2021). https://doi.org/10.1007/s12553-021-00600-3

2. Bush, C.L., et al.: Toward the definition of personalized nutrition: a proposal by the American nutrition association. J. Am. Coll. Nutr. **39**(1), 5–15 (2020). https://doi.org/10.1080/07315724.2019.1685332
3. Kawamura, A., et al.: Dietary adherence, self-efficacy, and health behavior change of WASHOKU-modified DASH diet: a sub-analysis of the DASH-JUMP study. Curr. Hypertens. Rev. **16**(2), 128–137 (2019). https://doi.org/10.2174/1573402115666190318125006
4. Eduardo. Sabaté and World Health Organization., Adherence to long-term therapies: evidence for action. World Health Organization (2003)
5. Jimmy, B., Jose, J.: Patient medication adherence: measures in daily practice (2011)
6. Landa-Anell, M.V., Melgarejo-Hernández, M.A., García-Ulloa, A.C., Del Razo-Olvera, F.M., Velázquez-Jurado, H.R., Hernández-Jiménez, S.: Barriers to adherence to a nutritional plan and strategies to overcome them in patients with type 2 diabetes mellitus; results after two years of follow-up. Endocrinol. Diabetes Nutr. **67**(1), 4–12 (2020). https://doi.org/10.1016/j.endinu.2019.05.007
7. Hughes, R.L., Marco, M.L., Hughes, J.P., Keim, N.L., Kable, M.E.: The role of the gut microbiome in predicting response to diet and the development of precision nutrition models-Part I: overview of current methods. Adv. Nutr. **10**(6), 953–978 (2019). https://doi.org/10.1093/advances/nmz022
8. Tay, W., Kaur, B., Quek, R., Lim, J., Henry, C.J.: Current developments in digital quantitative volume estimation for the optimisation of dietary assessment. Nutrients **12**(4), 8–15 (2020). https://doi.org/10.3390/nu12041167
9. Kashyap, P.C., Chia, N., Nelson, H., Segal, E., Elinav, E.: Microbiome at the frontier of personalized medicine. Mayo Clin. Proc. **92**(12), 1855–1864 (2017). https://doi.org/10.1016/j.mayocp.2017.10.004
10. Wilkinson, M.D., et al.: Comment: The FAIR guiding principles for scientific data management and stewardship. Sci Data **3** (2016). https://doi.org/10.1038/sdata.2016.18
11. Ismail, L., Materwala, H., Karduck, A.P., Adem, A.: Requirements of health data management systems for biomedical care and research: scoping review. J. Med. Internet Res. **22**(7) (2020). JMIR Publications Inc. https://doi.org/10.2196/17508
12. Samra, H., Li, A., Soh, B.: G3DMS: design and implementation of a data management system for the diagnosis of genetic disorders. Healthcare (Switzerland) **8** (3) (2020). https://doi.org/10.3390/healthcare8030196
13. Olsen, I.C., Haavardsholm, E.A., Moholt, E., Kvien, T.K., Lie, E.: NOR-DMARD data management implementation of data capture from EHR. Clin. Exp. Rheumatol. **32**(5), S158–S162 (2014)
14. Folador, J.P., Vieira, M.F., Pereira, A.A., Andrade, A.D.O.: Open-source data management system for Parkinson's disease follow-up. Peer J. Comput. Sci. **7**, 1–23 (2021). https://doi.org/10.7717/peerj-cs.396
15. Zhao, X., Xu, X., Li, X., He, X., Yang, Y., Zhu, S.: Emerging trends of technology-based dietary assessment: a perspective study. Eur. J. Clin. Nutr. **75**(4), 582–587 (2021). https://doi.org/10.1038/s41430-020-00779-0
16. Berger, M.M., et al.: Impact of a computerized information system on quality of nutritional support in the ICU. Nutrition **22**(3), 221–229 (2006). https://doi.org/10.1016/j.nut.2005.04.017
17. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G.: Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. **42**(2), 377–381 (2009). https://doi.org/10.1016/j.jbi.2008.08.010
18. Charles Vesteghem, A., et al.: Implementing a data infrastructure for precision oncology projects leveraging REDCap Charles Vesteghem. medRxiv preprint (2022). https://doi.org/10.1101/2022.05.09.22274599

electronic data capture systems in clinical research and their ability to improve and facilitate the data management process.

Having access to data with real-time monitoring and reporting can facilitate the process of informed decision-making making, therefore be valuable for (1) achieving a proper assessment by identifying patient inputs that will help to gain a better understanding of the individual's behavior and distinctive characteristics ranging from general host data, biological data to the more advanced assessment of complex data such as gut microbiome, (2) interpreting the analysis of the patient's data gathered to derive actionable information from them and identify patterns and relationships between the different factors, (3) producing actionable interventions tailored to the specific patient's needs and goals, and (4) monitoring and evaluating to track progress and adjust recommendations as needed throughout the time, in an iterative process, and guarantee further refining in the intervention, enhancing thus patient's adherence and engagement to the personalized nutrition plan. One interesting study [26] reported the substantial contributions to patient care from using REDCap, where demographic, epidemiologic, and clinical data of HIV-positive and negative patients with or without liver and cervical cancer were accessible in REDCap, and have helped healthcare professionals in providing more personalized care, as well as promoting patients' involvement in their health care.

Overall, the data-driven methodology and workflow process leveraging REDCap has been deployed to structure the collection of data relevant for coordinating and implementing personalized nutrition in clinical practice, emphasizing the importance and value of the insights gained from data in providing efficient solutions. The digitization process and integration offer real-time accessible information to healthcare providers to inform clinical decisions for personalized nutrition purposes and improve individual health outcomes by leveraging an approach of patient-centered care, with a special focus on improving adherence to nutritional treatment plans. This study, however, has some limitations due to its design (i.e., observational) and due to the individuals lost in the follow-up assessments. Nevertheless, death causes will be available for the 20-year period of follow-up, as well as some clinical information of those subjects that were not revised during the first and second follow-up waves which will be extracted from medical registries in the coming months. Future work will eventually emphasize on making available sequenced data for genome-wide association studies (GWAS) to help achieve more targeted interventions and recommendations.

Acknowledgments. The authors wish to acknowledge the partners of the BEAMER project. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 101034369. The JU receives support from the European Union's Horizon 2020 Research and Innovation Programme, the European Federation of Pharmaceutical Industries and Associations (EFPIA), and Link2Trials BV.

References

1. van Ommen, B., et al.: Systems biology of personalized nutrition. Nutr. Rev. **75**(8), 579–599 (2017). https://doi.org/10.1093/nutrit/nux029

3.3 Impact of Digitization in the Clinical Setting

The digitization of the available data contributed to having an efficient process, compared to paper-based data collection, to limit the number of manual tasks and automatically capture the relevant data from capturing individual dietary records to integrating omics data and calculating the score for adherence to the Mediterranean diet. The use of the REDCap platform ensured a standardized collection of patient data, minimizing missing data and ensuring high-quality data collection. It offers hospital practitioners a practical option to easily export the data, generate reports, statistics, and charts that can help in comprehensively presenting data, as well as overviewing the included patients and the availability of their data, combining clinical and omics data, offering a support tool for clinicians to inform clinical decision making related to providing personalized nutrition strategy in a novel way that could be adopted in a clinical setting.

4 Discussion and Conclusion

The study associated with this work, the Segovia study, is a longitudinal population-based study that was at the beginning a study representative of the Segovia province resident population in urban and rural areas aiming at estimating the prevalence of the Metabolic Syndrome [22]. The present paper focuses on the sub-cohort of patients with a median follow-up of 20 years (i.e., third visit) that aims at studying the Metabolic Syndrome, diet and characterization of the intestinal microbiota. Given the importance of digital health, and to leverage the use of digital tools in clinical settings like REDCap and its advanced data storage features, we were able to create a detailed patient profile by integrating multiple data types, from dietary, and clinical data to microbial profiles. This comprehensive patient profile is important for designing nutritional recommendations that are more likely to be adhered to. The microbiome profile of the patient is provided by the relative abundance of the microorganisms at the phylum level. It gives us information on how many proportions of the microbiome are made up of bacterial taxa at the phylum level. This allows us to evaluate further if the relative abundance of the bacteria is associated with one of the variables of interest such as dietary pattern or adherence to the Mediterranean diet.

Ensuring efficient data collection and enhancing data accessibility can be relevant for supporting actionable nutritional recommendations. Tools for data management systems like REDCap have a valuable contribution and have found larger use in several domains including research projects and clinical environments. As in the study of Brauer et al., [23] where they aimed to assess whether personalized nutrition in metabolic syndrome can be associated with diet quality changes, the data capture system was used to enter the nutrition process data, including data restrictions and real-time data integrity checks. Moreover, REDCap [24] has found its use in clinical trials where included subjects can complete an online self-screening form and a survey in the application, and where the completed case report forms and demographic information will be stored and updated. Related to that, researchers [25] have designed and implemented a web-based data management system for diabetes clinical trials that had a good rating among researchers using it, showing that electronic systems can facilitate the clinical data management process in diabetes and endocrinology research. These support the extensive use of

Table 1. (*continued*)

Analytical parameters
Diabetes diagnostic *Basal capillary glycemia (fasting glycemia)* *Capillary glycemia 2 h after glucose tolerance test* *Glycated hemoglobin HbA1c* Lipids profile: total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides Kidney and hepatic functions Urine: microalbuminuria
Gut microbiota data (relative abundance by phylum)
Overall characterization of 26 phyla; relative abundance of each phylum for each patient

3.2 Clinical REDCap Workflow

After designing and creating the environment on REDCap, one can either import the data using the data import tool option or add records directly via the record dashboards as the surveys and the required components have been created to facilitate the capture of data. By adding a new record, we are creating a new patient profile with its personal ID. We can then start capturing the demographic data, the phenotype-based information data, including anthropometrics, clinical information and clinical biomarkers, and nutritional information by filling the dietary record related to the FFQ, as well as the survey related to the Mediterranean questionnaire that will automatically retrieve the total score of the survey. Finally, the microbiome profile will be integrated as the relative abundance of each phylum present in each sample.

Figure (1) briefly shows the clinical REDCap workflow setup to enable the deployment of the digitization of the needed data for a personalized nutrition strategy in clinical care. The data management system created on REDCap allows us to capture the data directly by filling in the information in the appropriate fields. Once capturing data is completed, reports and statistics can be accessed, as well as initiating the data export.

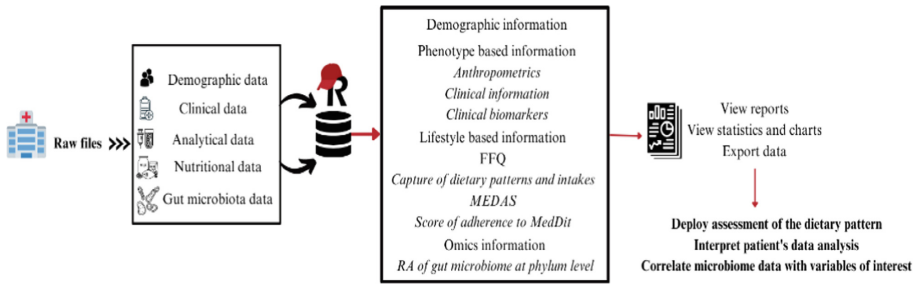


Figure. 1. Clinical REDCap workflow for supporting the implementation of personalized microbiome-based nutrition approaches in clinical research. *Abbreviation: RA. Relative abundance*

3 Results

3.1 Defining the Block of Variables and Integrating Data

Our data source included 113 patients (58,4% women and 41,6% men). Overall, after the selection of variables, we were able to define five blocks of variables that represent the instruments to be created on REDCap, with their associated metadata. We integrated multiple data types, including *(1) demographic data, (2) clinical data, (3) dietary records and nutritional data, (4) analytical data, and (4) omics data*, as detailed in Table 1.

Table 1. Block of variables retrieved with the associated metadata

Demographic data
Personal ID
Gender
Clinical center
Date of interview
Age
Year of revision
The starting time of the questionnaire
Marital status
Place of residence
Professional activity
Clinical data
Medical history, including *history of diabetes, hypertension, cholesterol, cardiovascular and intestinal diseases, other diseases, birth delivery, and physical examination*
Anthropometric measurements
Weight, Height, Waist circumference, Hip circumference
Blood pressure measurements
Electrocardiogram measurements
Nutritional data
14 items questionnaire of MEDAS
Food Frequency Questionnaire
Macronutrients intake (per day)
Micronutrients intake (per day)
Vitamins
Minerals
Total intake by food groups (g/day)

(*continued*)

composition: Operational Taxonomic Units (OTUs) table. OTUs table summarizes the composition of the microbial communities present in each sample, where each column represents a different sample, and each row represents the taxonomic identification of the bacterial taxa from the level of phylum, class, order, family, genus to species. The abundance of each taxonomic unit in each sample is represented by the relative abundances in the cells.

2.4 Setting up the REDCap Database and Data Entry Workflow

The collected patient data was transferred to a REDCap database created on the REDCap web-based application accessed through the hospital network, with only the research team having user rights. The digitization workflow included:

Selecting the relevant variables for defining the block of variables.

Patient clinical information gathered from the general questionnaire consisted of extensive information on patient health where several types of data were collected. For now, and based on the purpose of the research team and the study, only the data of interest was captured and digitized. The selection was made by the research team based on a mapping of common variables from the different questionnaires of each of the visits at the three points of time, i.e., 2000, 2008, and 2021. The interest behind this procedure is to be able later to investigate and establish the relationship between these variables over time, in a longitudinal way, as the data of the two previous studies projects is stored in a separate REDCap repository. In the end, only the relevant data has been selected, and the block of variables has been defined and captured in REDCap, as presented in the results part.

The nutritional information to be digitized covered on one hand, the data gathered from the answers to the 14 questions of the Mediterranean Diet Adherence Screener and its total score, and on the other, the data gathered from the Food Frequency Questionnaire, that is the result of the automatic calculations of estimated daily intakes of different nutrients and food groups from the respondent-reported information.

Finally, the gut microbiome information, given the large size of the initial OTUs table and the complexity of gut microbiome data, data will be captured as the relative abundance at each phylum level for every sample (i.e., patient) to enable simple and efficient data capture.

Completing the required metadata information

The worksheet with all the defined data elements is used to complete the requirements for data entry, by filling in the specific information related to *the type of variable (field type), the field label, and the variable name in REDCap* to prepare the data entry in the software for building the environment.

Creating the environment via the Online Designer

The online designer in REDCap allows the creation of the environment via the instrument collection page, where data collection instruments, referring to the block of variables defined, can be created and the variables can be added.

Capturing the data

Once the environment has been added to REDCap, it is possible to upload the data directly into REDCap using the Data Import Tool.

2 Materials and Methods

2.1 Ethical Statement

The study protocol was approved by the Ethics Committee of Reference of the Regional Health Service of Segovia, Spain, as well as the Ethics Committee from Hospital Clínico San Carlos of Madrid, Spain, which also approved the related data management processes (17/183-E-BS and 19/409-E).

2.2 Data Source

The Hospital Clínico San Carlos of Madrid, together with the Primary Care Centre of the province of Segovia (Autonomous Community of Castilla y León) in Spain are carrying out a collaborative study to investigate the metabolic syndrome and cardiovascular risk factors in a cohort of patients: the SEGOVIA cohort study. This study is a longitudinal population-based study with a long follow-up of 20 years involving a cohort of 809 subjects aged between 35 and 74 years, enrolled in the study between January 2000 and January 2003. Assessments were carried out at three points in time, in which the study variables are collected: a baseline visit from 2000, a second visit with a median follow-up of 7 years in 2008, and, finally, a third visit with a median follow-up of 20 years in between 2021–2023. The work presented in this paper focuses on the sub-cohort of patients with a median follow-up of 20 years, where a sub-study of metabolic syndrome, diet and characterization of the intestinal microbiota is being conducted.

2.3 Data Collection

The patient clinical information was gathered by means of a written questionnaire in a face-to-face interview with an interviewer, where an individual code was assigned to each patient. The report form includes extended questions to gather the necessary information about the patient, related among others, to his personal data, medical history, clinical data, and daily habits.

The nutritional assessment and the dietary data collection of patients were carried out using two types of evaluation tools: *(1)* a 14-item Mediterranean Diet Adherence Screener (MEDAS) [21] integrated into the general questionnaire, where the level of adherence to the Mediterranean diet is assessed using a 14-item questionnaire (12 questions on food consumption frequency and 2 questions on food intake habits considered characteristic of the Spanish Mediterranean diet), and *(2)* a Food Frequency Questionnaire (FFQ) from the University of Navarra (Spain) that consists of a list of foods and beverages with categories of response to estimate the frequency of consumption over a specified period of time. Both the patient clinical information and the nutritional data were given in a Microsoft Excel spreadsheet file in an anonymized way, with the patient ID as the identifier.

Finally, the characterization of the composition of the gut microbiota was done after the collection of feces samples provided by the patients and their analysis by the Microbiota Laboratory of the Hospital of Madrid using Next Generation Sequencing (NGC) technologies. The results were given in a tabular representation of the gut microbiome

of health data management systems to provide to a greater extent accurate and better patient care [11]. Researchers have designed and implemented a data management system to manage patient data and support clinicians in their decision-making to diagnose genetic diseases [12]. Some worked on capturing data related to patients with inflammatory joint diseases directly from an electronic health record system and transferring them into an electronic data capture system, helping in transitioning from paper format to electronic system [13]. Another study proposed an integrated data management system to support and manage data of patients with Parkinson's disease, preventing, therefore, data loss, and offering patients clinical follow-up and monitoring [14]. In nutrition care, dietary assessment and nutritional monitoring can be challenging for healthcare professionals and nutritionists as the available methods are time-consuming and susceptible to human errors [15]. Leveraging digital health technology may now offer a new way to provide medical nutritional therapy on a more accurate, personal, and accessible level. There is evidence for example that computerized patients' data management systems can improve nutritional care and monitoring, with better data visibility and adequate nutrition delivery [16].

One of the easily accessible data collection and management systems is the Research Electronic Data Capture (REDCap) [17]. REDCap is a secure, web-based application designed for research teams as a tool to collect, manage, and store research data in a secure environment. Because of its user-friendly interface, it has been used in several domains, and more importantly in clinical research as a tool to support precision medicine in oncology [18] and support clinicians in assessing the probability of patient outcomes after surgery for pancreaticoduodenectomy [19]. The field of nutrition research is known for its huge amount and complex data, which makes it one of the healthcare fields that are advancing in the use and application of computational techniques to its important data [20]. REDCap can be an important part of data-driven projects related to nutrition, as data collected and managed using REDCap can be later processed and analyzed using various computational techniques, including machine learning (ML) and artificial intelligence (AI), evolving the field of personalized nutrition.

Therefore, given the close relationship of the microbiome with nutrition and its key role in modulating health and disease, it is important to integrate the gut microbiome as a component of a personalized nutrition intervention along with other individuals' information. With the ability to modulate the gut with diet, it is appealing to target the gut microbiome with diet-based strategies and to harness digital technologies, such as REDCap's data management features to efficiently collect and integrate data, leading to more tailored dietary interventions, which can be of interest for both healthcare professionals and patients for providing targeted and actionable nutritional approaches and receiving recommendations for achieving sustainable results, respectively. To this end, this paper aims to describe the methodology and workflow process deployed to manage and integrate data at different scales using digital health (REDCap) to advance the field of personalized nutrition and facilitate the implementation of personalized microbiome-based nutrition approaches in clinical research.

1 Introduction

As stated by Van Ommen et al. [1] "Personalized nutrition tailors dietary recommendations to specific biological requirements based on a person's health status and goals". The personalized nutrition care model [2] reflects the different aspects used to allow patients to benefit from tailored interventions, with regular and continuous monitoring to reach specific outcomes. It includes: *(1)* The assessment with quantitative and qualitative host data such as diet, biochemistry, metagenomics, etc., *(2)* The interpretation of the personalized data through scientific evidence, *(3)* The intervention developed using guidance and therapeutics to design actionable interventions such as changes to diet and lifestyle factors, and *(4)* The ongoing monitoring and evaluation along the care process for regular feedback, and therapeutic interventions refinement to achieve self-efficacy and behavior change. The therapeutic intervention adherence, represented here by the adherence to an adequate diet, is an important indicator of self-efficacy and health behavior change [3]. Adherence is defined by the World Health Organization (WHO) as the degree to which a person complies with agreed recommendations from a healthcare practitioner, whether taking medication, adhering to a diet, or implementing other lifestyle changes [4]. Across all therapeutic areas, patient non-adherence represents an issue. The lack of patient treatment adherence can be associated with poorer health outcomes, lower quality of life, death, and a burden on healthcare costs [5]. In Medical nutritional therapy (MNT), we understand adherence issues as mainly poor adherence to the nutritional plan. Research indicates that individuals facing different barriers are less likely to comply with a long-term dietary plan [6]. With tailored nutritional recommendations that address patient's needs and barriers, considering the perspectives of personalized nutrition can be a great strategy for facilitating and improving nutrition adherence, and as variations in how each individual responds to diet are always present [7], personalized nutrition engages the idea of going from delivering lifestyle and nutritional recommendations from population to individual level, allowing thus a better adherence and achievement of nutritional goals, and effective behavior changes [8].

For this reason, richer data can help to achieve more targeted interventions and recommendations: from nutritional intake assessment, routine lab testing, and targeted lab testing to omics analysis, the tailored evidence-based strategies and interventions will go from generalized to more personalized for individuals [2]. As an example of the gut microbiome that is unique to each individual and can be influenced by several factors, our diet greatly impacts it, therefore, manipulating this latter with dietary approaches will consist of using gut microbiome markers to optimize dietary interventions, to modulate diet and using diet to modulate the gut microbiome [9]. Personalized nutrition is therefore considered one of the greatest advances in modern medicine, especially with the development of omics and digital technologies.

However, shifting recommendations from the population to the individual level to achieve personalized nutrition requires extensive clinical and omics data, which comes with the need for a robust data collection and management system. As defined in the FAIR (Findable, Accessible, Interoperable, Reusable) data guiding principles [10], making the data collected easily searchable, accessible, integrated, and used in combination with other data and reusable is also a crucial element for supporting personalized nutrition. Moreover, the advancement in information technologies (IT) has led to the development

A Data-Driven Methodology and Workflow Process Leveraging Research Electronic Data Capture (REDCap) to Coordinate and Accelerate the Implementation of Personalized Microbiome-Based Nutrition Approaches in Clinical Research

Hania Tourab[1], Macarena Torrego Ellacuría[2], Laura Llorente Sanz[2], Arturo Corbatón Anchuelo[2], Dulcenombre Gómez-Garre[2], Silvia Sánchez González[2], María Luaces Méndez[2], Beatriz Merino-Barbancho[1], Julio Mayol[2,3], María Fernanda Cabrera[1], María Teresa Arredondo[1], and Giuseppe Fico[1,3(✉)]

[1] Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain
gfico@lst.tfo.upm.es
[2] Instituto de Investigación Sanitaria, Hospital Clínico San Carlos, (IdISSC-HCSC), Madrid, Spain
[3] Departamento de Cirugía, Facultad de Medicina, Universidad Complutense de Madrid, Madrid, Spain

Abstract. In the rapidly evolving field of precision medicine, personalized nutrition is taking over as well. As individuals respond differently to diet, personalized nutrition engages the idea of going from delivering lifestyle and nutritional recommendations from the population to the individual level, allowing thus a better adherence to the diet, an achievement of nutritional goals, and an effective behavior change. The main factor contributing to the development of personalized nutrition in healthcare is the increase in the availability of large patient data that offers the opportunity to explore and investigate the relationship between various patient features, going from integrating simple data host to metagenomics data. Therefore, using microbiome data as a component of personalized nutrition can be substantial given the close relationship of the gut microbiome with nutrition and the host's health. However, shifting recommendations from the population to the individual level requires a robust data collection and management strategy. In this paper, we aim to describe the methodology and workflow process that uses Research Electronic Data Capture (REDCap) to facilitate the implementation of personalized microbiome-based nutrition approaches in clinical research.

Keywords: personalized nutrition · microbiome data · clinical research · electronic data capture

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 137–147, 2024.
https://doi.org/10.1007/978-3-031-59717-6_10

8. Kang, M.J., Rossetti, S.C., Knaplund, C., et al.: Nursing documentation variation across different medical facilities within an integrated health care system. Comput. Inf. Nurs. **39**, 845 (2021)
9. Rossetti, S.C., Dykes, P.C., Knaplund, C., et al.: The communicating narrative concerns entered by registered nurses (CONCERN) clinical decision support early warning system: protocol for a cluster randomized pragmatic clinical trial. JMIR Res. Protoc. **10**, e30238 (2021)
10. Xu, W., Xu, D., Alatawi, A., et al.: Statistical unigram analysis for source code repository. Int. J. Semant. Comput. **12**, 237–60 (2018)
11. Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson/ Prentice Hall, New Jersey (2009)
12. Afzal, Z., Schuemie, M.J., van Blijderveen, J.C., et al.: Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med. Inform. Decis. Mak. **13**, 1–11 (2013)
13. Zuccon, G., Wagholikar, A.S., Nguyen, A.N., et al.: Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the SNOMED CT ontology. AMIA Summits Transl. Sci. Proc. **2013**, 300 (2013)
14. Wrenn, J.O., Stetson, P.D., Johnson, S.B.: An unsupervised machine learning approach to segmentation of clinician-entered free text. In: AMIA Annual Symposium Proceedings. vol. 2007, p. 811. American Medical Informatics Association (2007)
15. Koleck, T.A., Dreisbach, C., Bourne, P.E., et al.: Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J. Am. Med. Inform. Assoc. **26**, 364–79 (2019)
16. Jensen, K., Soguero-Ruiz, C., Oyvind Mikalsen, K., et al.: Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci. Rep. **7**, 46226 (2017)
17. Ng, H.T., Lim, C.Y., Koo, J.L.T.: Learning to recognize tables in free text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 443–450 (1999)
18. Josefsson, S.: The base16, base32, and base64 data encodings. Tech. rep. (2006)
19. Moy, A.J., Schwartz, J.M., Chen, R., et al.: Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. J. Am. Med. Inform. Assoc. **28**, 998–1008 (2021)
20. Bakken, S., Dykes, P.C., Rossetti, S.C., et al.: Patient-Centered Care Systems, pp. 575–612. Computer Applications in Health Care and Biomedicine. Springer, Biomedical Informatics (2021)
21. Tran, B., Lenhart, A., Ross, R., et al.: Burnout and EHR use among academic primary care physicians with varied clinical workloads. AMIA Summits Transl. Sci. Proc. **2019**, 136 (2019)
22. Gregório, J., Cavaco, A.M., Lapao, L.V.: How to best manage time interaction with patients? Community pharmacist workload and service provision analysis. Res. Soc. Adm. Pharm. **13**(1), 133-47 (2017)
23. Morris, R., MacNeela, P., Scott, A., et al.: Reconsidering the conceptualization of nursing workload: literature review. J. Adv. Nurs. **57**, 463–71 (2007)
24. Bokhari, S.M.A., Basharat, I., Khan, S.A., Qureshi, A.W., Ahmed, B.: A framework for clustering dental patients' records using unsupervised learning techniques. In: 2015 Science and Information Conference (SAI), pp. 386–394. IEEE (2015)
25. Bokhari, S.M.A., Khan, S.A.: Applying supervised and unsupervised learning techniques on dental patients' records. In: Emerging Trends and Advanced Technologies for Computational Intelligence: Extended and Selected Results from the Science and Information Conference 2015, pp. 83–102. Springer (2016)

that were only recorded in the free text by nurses. The choice of nurses to exclusively document this information in the free text could be attributed to several hypotheses. It could potentially result from usability concerns or limitations with the granularity of data accommodated by structured forms. Alternatively, it may reflect a preference for narrative documentation when conveying specific clinical phenomena. Further research is needed to understand the characteristics and implications of terms present in either free text or structured data from nurses' notes. Typically, free text notes give a summarized and up-to-date picture of a patient's current state. Such free text data may be used in EWS to predict health deterioration early before changes in vital signs appear [1].

To the best of our knowledge, this is a unique contribution to the NLP literature, namely, to extract free text from the primary formats of nursing documentation (structure, semi-structured, free text) and subsequently use unigram analyses to attain deeper insights into the free text. The HTML format notes used for this research are coming from Epic which itself is a widely used system in US hospitals, implying a common HTML format. We understand the limitations of the heuristic-based approaches though; however, we see the problem as a text classification problem, which relies on annotated datasets for training purposes. Our heuristic-based approach helped annotate the data to train ML algorithms in the future for a more scalable solution CONCERN EWS system at other hospitals.

Acknowledgments. This work is supported by the National Institute for Nursing Research (NINR) CONCERN Study #1R01NR016941 and the American Nurses Foundation Reimagining Nursing Initiative (RN Initiative). JW is a postdoctoral research fellow supported by the Reducing Health Disparities through Informatics training grant (T32NR007969).

References

1. Rossetti, S.C., Knaplund, C., Albers, D., et al.: Healthcare process modeling to phenotype clinician behaviors for exploiting the signal gain of clinical expertise (HPM-ExpertSignals): development and evaluation of a conceptual framework. J. Am. Med. Inform. Assoc. **28**, 1242–51 (2021)
2. Merchant, R.M., Yang, L., Becker, L.B., et al.: Incidence of treated cardiac arrest in hospitalized patients in the United States. Crit. Care Med. **39**(11), 2401–2406 (2011)
3. Liu, V., Escobar, G.J., Greene, J.D., et al.: Hospital deaths in patients with sepsis from 2 independent cohorts. JAMA **312**, 90–2 (2014)
4. Collins, S.A., Vawdrey, D.K.: Reading between the lines of flowsheet data: nurses optional documentation associated with cardiac arrest outcomes. Appl. Nurs. Res.: ANR **25**(4), 251 (2012)
5. Collins, S.A., Fred, M., Wilcox, L., et al.: Workarounds used by nurses to overcome design constraints of electronic health records. In: NI 2012: 11th International Congress on Nursing Informatics, June 23-27, 2012, Montreal, Canada, vol. 2012. American Medical Informatics Association (2012)
6. Collins, S.A., Bakken, S., Vawdrey, D.K., et al.: Agreement between common goals discussed and documented in the ICU. J. Am. Med. Inform. Assoc. **18**, 45–50 (2011)
7. Collins, S., Hurley, A.C., Chang, F.Y., et al.: Content and functional specifications for a standards-based multidisciplinary rounding tool to maintain continuity across acute and critical care. J. Am. Med. Inform. Assoc. **21**, 438–47 (2014)

Table 2. Unigram analysis of the terms exclusive to structured data

Rank	Structured Text Word	Word Freq.	Document Freq.
1	vu [verbalized understanding]	192822	758
2	discipline	4632	772
3	latino	3972	1324
4	solving	3747	1100
5	element	3474	526
6	grass	3357	767
7	implement	2946	245
8	opium	2457	763
9	hydrocodone	2451	753
10	mushrooms	2394	758
11	hallucinogens	2394	758
12	ecstasy	2394	758
13	stimulants	2373	759
14	sedatives	2370	758
15	dexedrine	2361	759
16	concerta	2361	759
17	ritalin	2361	759
18	ghb	2358	758
19	serepax	2358	758
20	introductions	2352	778

4 Discussion

Literature suggests that when nurses optionally decide to write free text the contents may be a strong signal for information that the nurse wants to communicate to the rest of the healthcare team [1,9]. In this regard, this research analyzed over 200K EHR notes and extracted 40,000 free text notes from them. The problem is that such free text is often found embedded in large datasets, which are hard to retrieve given a lack of clear distinctions between the data. Furthermore, it was challenging to extract such data because of their structural diversities.

This paper describes a heuristic-based extraction and unigram analysis approach to identify as well as understand free text residing in larger EHR nursing notes. We analyzed the data by identifying the unigrams unique to free text data to determine the difference between the two datasets (structured and free text documentations). Because if there were no major differences between the two texts then it would be harder to detect such texts dynamically as both could be labeled essentially the same. Our research found the difference between free text and structured data is statistically significant; there are many clinical terms

Table 1. Unigram analysis of the terms exclusive to narrative-free text data

Rank	Free Text Word	Word Freq.	Document Freq.
1	isol - [isolation]	599	598
2	flange	510	171
3	hollister	315	219
4	couplets	293	284
5	kpouch	279	114
6	peristomal	269	228
7	midabdominal	255	187
8	pacs [premature atrial contractions]	241	231
9	drainable	202	167
10	endo	197	177
11	budded	195	190
12	ceraring	180	159
13	urinal	171	152
14	padded	157	153
15	mf [multiform]	154	144
16	convexity	151	137
17	incont [incontinence]	149	124
18	sterility	138	138
19	phenylephrine	138	108
20	apcs [atrial premature complexes]	138	127

Also, we noticed that some aspects of the templates are not always relevant or useful in all patient cases. For instance, a prevalent term found was the word "element", which often appeared as part of a templated structured field as N/A, thereby indicating that the specific data element was not applicable. The frequent occurrences of the terms such as "element" being "N/A" in the documentation suggest that such information is continuously being recorded, even when a specific data element does not apply to the patient's situation. The need to document each aspect of patient care, even when certain data elements are not applicable, may contribute to the documentation burden specifically related to reviewing and synthesizing data, as well as "note bloat" [19,20]. This may impact the workload of clinicians, thereby affecting the time spent on direct patient care [21–23]. Furthermore, understanding the rationale of nurses regarding their decision to document certain aspects of clinical care in narrative free-text notes rather than structured flowsheet fields, could also be an area of future research. Use of our heuristic approach to detect and leverage concerning clinical concepts documented in narrative nursing notes, and subsequently incorporating this as a feature into the predictive model can help improve clinical deterioration prediction.

portions are generated from structured (templated) data entered by nurses elsewhere in the EHR. Therefore, our approach does not convert these words into their respective root words in order to preserve any grammatical structures that exist and may be helpful in the future to detect such text automatically using machine learning approaches.

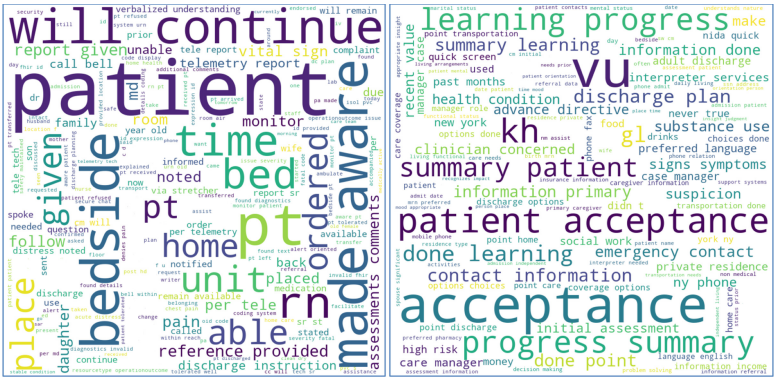


Fig. 5. Free text (left) and structured data (right) word clouds.

The free text unigrams in Table 1 revealed two primary clinical areas: 1) wound ostomy continence (WOC) care [associated terms in yellow] and 2) cardiac patient care [associated terms in cyan]. Interestingly, despite the existence of specific structured fields in the EHRs to capture concepts aligned with wound ostomy continence and cardiac care, we observed unique terms associated with these domains captured only in the free text. There may be several reasons that nurses choose to document this information in free text, ranging from the usability or data granularity issues with the structured forms to a preference for narrative forms of documentation to describe particular clinical phenomena. Future research should seek to explain the characteristics of terms found only in narrative text versus those found only in structured data.

As for the terms exclusive to structured text in Table 2, the most frequently appearing word was 'VU' which is short for "verbalized understanding". The documentation of patient verbalized understanding in structured fields indicates the nurses' evaluation of the level of comprehension regarding the patient's education. Besides, these unigram results for the structured portion comprise the terms primarily found in two nursing documentation templates, 1) care management assessment [associated terms in green], 2) substance use history screening [associated terms in pink] while to a lesser extent, the others belong to templates for nursing care plans and the associated action(s) to achieve the stated care planning goals [associated term in grey] and time-out documentation for the operating room [associated term in peach]. This may infer that these specific templates are more frequently used than others; further research should explore the portion of structured documentation that are focused on these aspects of patient care.

Algorithm 2:

```
1 func_check_no_label (all_divs):

2       check_no_label = True
3       for div in all_divs:
4               if (div not empty):
5                       if (check_label(div) = True):
6                               check_no_label = False
7                               break

8   return check_no_label
```

that identifies free text. To do so, we conducted a thematic analysis using the unigram language model [10] to identify and compare the recurring domains with the aim to understand the clinical context in which the notes were likely written. Two registered nurses (RL, JW) who have training in informatics research and clinical experience, served as the subject matter experts and individually interpreted the unigram results in Table 1 and 2 to gain more insight about the difference between the contents of free text and structured data. They then met with the primary author (SMAB) to iteratively discuss and reach a consensus on the interpretation of the results related to clinical context and nurse documentation workflow.

3 Results

In our analysis of over 200K documents, we retrieved (based on the pre-specified rules in the algorithm) 40K free text notes in total, out of which 33K were identified as all free text records and 7K free text records found embedded in structured data. We detached free text from the structured portion through our aforementioned heuristics-based approach. A large portion (160K) of the notes consists of only the structured data while the percentage of narrative free text in a note was found to be 1–3%. We found high levels of redundancy in the structured portion of the note as compared to the narrative portion. The same words/blocks of the structured portion of the note are repeated several times. The contents of the structured and free text differed sufficiently; we detected 15K unique words in the narrative text that are not present in the structured portion of the text and 7K unique words in the structured text that are not present in the narrative portion of the text. There were 14.5K overlapping words found.

Figure 5 shows the word clouds for the free text and structured portion indicating the difference between the two where the size of each word indicates its frequency. Table 1 shows the top 20 most frequently occurring free-text terms exclusive to narrative free text while Table 2 shows the top 20 most frequently found terms unique to structured data. Tables show the frequency each of these words appears across the entire dataset (word frequency) and the number of documents in which each of these terms appears (document frequency). Again, the free text is written narratively by registered nurses while the structured

Algorithm 1:

```
check_label (selected_div, div_count)
    boolean = False
    div_txt = selected_div.text.strip() # Remove leading and trailing empty spaces from div text

    # No 'plan of care' or discharge note'
    if (!(div_txt.lower().contains("plan of care" or "discharge note"))):
        words = div_txt.split()
        for w in words[:5] # Check the first five words to check if they are heading/title
            if ((w.istitle() or w.isuper()) and (w.endswith(":"))):
                if (!w.equals("Comment" or "Comments")):
                    boolean = True
                    break
                else:
                    # Only consider comments from the later divs of the document
                    if ((selected_div.index()/ div_count)>0.95):
                        boolean = True
                        break
    return boolean
```

to be all free text. We depict this in Fig. 4. Algorithm 2 detects all free text by checking if there are no headings in the HTML file. Overall, the algorithm works in this way if there exist relevant tokens, headings, and other indicators, then the document is a mix of structured and free text. If no tokens are present, then the document is likely all free text with no structured portions in it and we annotate it as all free text notes accordingly. If the relevant indicators (tokens) exist in a document and the pre-specified (static) rules are met, the algorithm identifies and extracts free text from relevant locations and what is left behind is merely the structured portion.

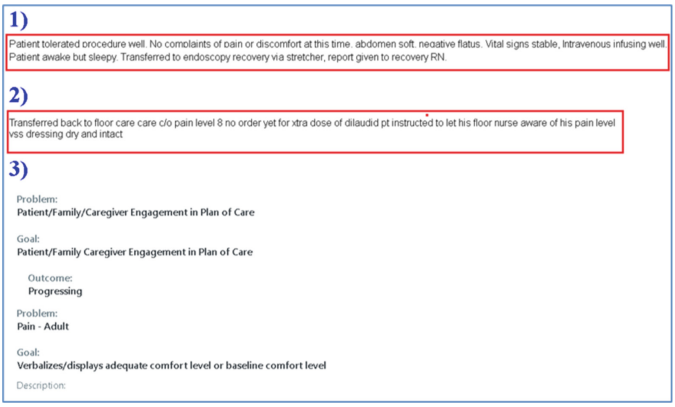


Fig. 4. Free-text examples

In this way, we extracted the relevant information and then categorized the identified free texts into different categories based on the identified tokens. For efficient analysis, the extracted information was stored in a JSON. In addition, we aimed to understand the unique characteristics of free text compared to structured text in order to inform the creation of an automated dynamic system

tokens, e.g., 'Assessments/Comments', 'Additional Comments', 'Comments', 'Other Comments', 'Comment', 'Nursing Note', 'Progress Note', 'Treatment Note', and 'Note'. Algorithm 1 retrieves the text from the selected divs of the HTML and removes leading and trailing spaces to check if the first few words contain a heading, i.e., a title containing a colon. Having traversed through all the divs in a document, only relevant divs are selected based on the aforementioned free text tokens.

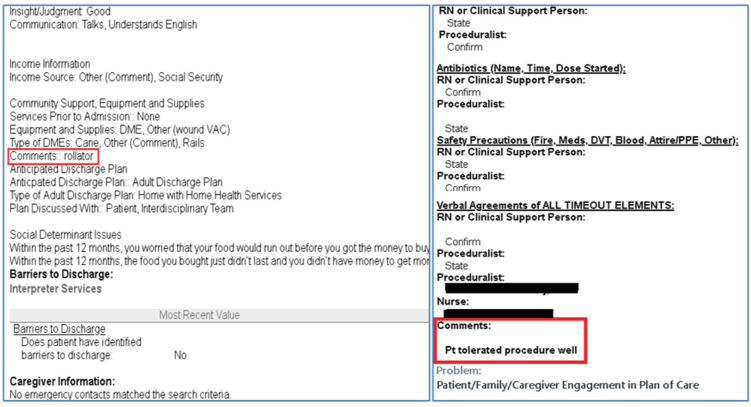


Fig. 3. Example of overlapping token: structured text (left) and free text (right)

If a relevant (containing tokens) heading is found in a div (in certain locations), the algorithm extracts the surrounding text of other divs as free text, otherwise ignores it. If the heading contains specific tokens such as 'comment' and 'comments' then it checks the location of the div within the HTML since not all comments are free text, but the comments in the later part of the documents are likely to be free text. We identified the free text nursing notes from the respective divs based on such identified static rules to build our approximate free text dataset. In the process, we ignore the divs containing the plan of care or discharge notes. The plan of care notes primarily contain structured text reflecting future plans, as opposed to immediate concerns about a patient's state. Discharge notes are documented at the end of patients' hospital stay and therefore would not be available to our algorithm because we are interested in predicting deterioration during a patient's hospital stay in real-time.

2.3 Identify and Retrieve All Narrative Notes Data

Upon examination, we observed that if no heading exists in a document, the structured text is unlikely to be present, rather, the document content is likely

to locate the relevant divs, 3) how to differentiate the relevant divs from the other div information, and 4) where to cut and extract the information given that no nested divs exist to indicate the ideal spot to cut. Again, there were no explicit labels to differentiate parts of the notes such as document header, div name, or any other metadata, as well as to differentiate between different divs and analyze div text accordingly. Importantly, there can be hundreds of lines in a note while there may be only a few free text words present in the note.

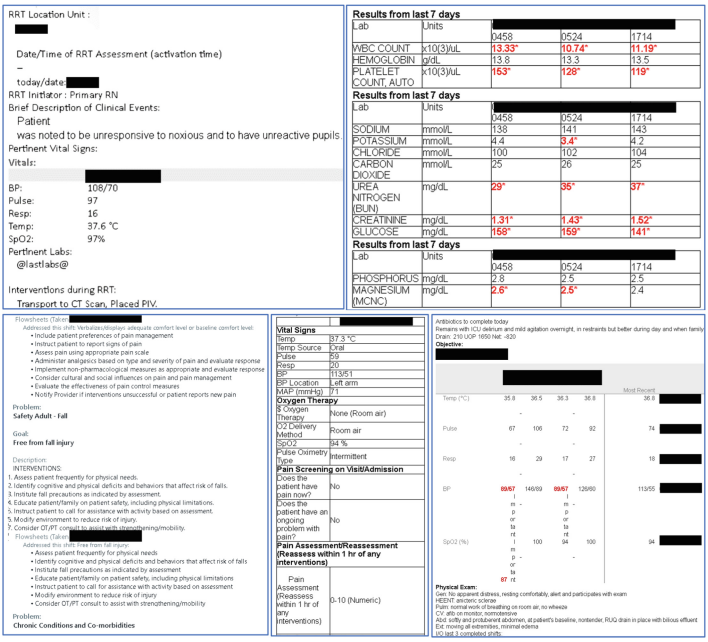


Fig. 2. Example formats of nursing notes originated from the Epic© EHR system

The approach we used was to manually review thousands of files to develop a heuristic-based algorithm, based on the identified static (prespecified) rules to retrieve the relevant portions of the data from HTML tags. For instance, we observed that if certain indicators such as tokens, formats, and headings, exist in certain locations, the data are likely to be free text. Also, it is important to determine which are the relevant and also the irrelevant tokens since token names may overlap between free text and structured portions of a note (see Fig. 3). In this case, we take into consideration other indicators to determine the relevant information, such as the location of the token in the document. The ultimate goal of this approach is to help build a free text dataset that can be used to identify such narrative texts automatically independent of the syntactical differences, which, as aforementioned, is important for the scalability of the system.

Therefore, we traverse through all the files and their respective divs and select only those divs which are relevant, i.e., div text contains specific free text

notes files. HTML notes files contain both free texts, as well as structured data. Figure 1 shows different stages of our dataset: notes SQL data in base64 [18] encoded format, decoded HTML notes files, retrieved text from HTML notes, and HTML text transformed into JSON documents.

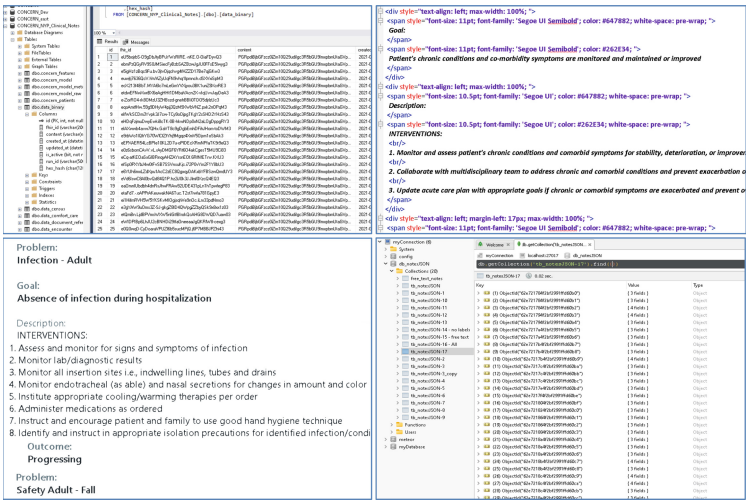


Fig. 1. Example notes originated from the Epic© EHR system: Encoded SQL notes data (top-left), HTML notes file (top-right), notes text (bottom-left), and notes JSON documents (bottom-right).

We distribute our dataset into two parts: 1) sections with structured text and 2) free text. The task is to differentiate and extract the free text. And it is important to be aware of the structural and content-level differences in the data to identify and detach free text from the structured portion. Structural differences refer to the way specific data are stored, which is important for retrieval of relevant data while the content level differences are important to uniquely identify relevant texts, which is crucial for the dynamicity of the solution.

2.2 Segregate and Retrieve the Narrative Embedded in the Nursing Notes

We started by traversing through the HTML nursing notes. We found significant variability in the formats, some of the examples are shown in Fig. 2. The figure shows different layouts of the nursing clinical notes, including plain text, different tabular and other formats, demonstrating a high level of variability. However, we observed that the independent divs in HTML notes files were primarily the place where the narrative data were stored. The 'div' tag defines a division or a section in an HTML document. In automating the extraction process, it is important to determine: 1) which div contained the relevant information (free texts), 2) how

regard, this research uses a heuristic-based approach to extract free text data and utilizes unigram analysis to gain deeper insights into the nursing free text. Unigrams are the elementary subset of the n-gram language models, which is a subfield in natural language processing [10,11]. Based on our prior work we know that nursing free text has signals of nurses 'concerns about a patient. Such concerns can help detect patient health deterioration early even before the vital signs start to appear [1].

While existing research [12–16] has applied machine learning and NLP algorithms directly to free text datasets, and there has been an attempt [17] to recognize tables within free text data, our research uniquely focuses on first establishing the ground truth regarding the location and nature of free text to build a training dataset. This training dataset can be used in the future by machine learning algorithms to identify the free texts dynamically independent of the heuristic-based approach, which is extremely important for the scalability of the system given the potential variation in syntax across different sites. Since we see the problem as a classification task; a training set is required, which aims to establish the foundation to use machine learning approaches for predictive modeling in the future. Moreover, the fraction of the free text in the structured portions can be less than 1%, in addition, there is no metadata to differentiate. Without an annotated dataset, machines may struggle to distinguish the relevant portions. Since the relevant free text portions are so small and as an embedded part (free text) in the structured text, all look the same. Therefore, our heuristic approach is useful for creating a training set for the supervised learning approaches to make the solution heuristic-independent in the future for scalability purposes. Moreover, this HTML format is coming from the Epic EHR which is a widely used system in many hospitals within the US which also makes our heuristic-based approach potentially generalizable across hospitals that use the Epic EHR. Furthermore, unigram analysis of both structured and free text data in this research gives us more insights into the difference in the nursing free text compared to the structured data.

This research is foundational to developing an automated framework for identifying free text containing nurses' concerns from nursing notes through machine learning approaches in the future. In addition, our framework also needs to be a scalable component of the CONCERN EWS that is already being spread to multiple sites.

2 Methods

2.1 Description of Data

In our study, we used more than 200K nursing clinical notes that originated from the Epic©EHR system. Epic is one of the most widely used EHRs in the United States. The notes were retrieved using the Fast Healthcare Interoperability Resources' standard (FHIR) document service. FHIR is a set of rules and specifications for exchanging electronic healthcare data. The notes data was stored in SQL in base64 [18] encoded format, which was decoded into HTML

model implemented as clinical decision support tools in the inpatient setting to identify patients at risk of deterioration, including from events such as cardiac arrest and sepsis which impact approximately 330,000 inpatients per year [2,3]. Early identification of patient deterioration can allow for faster treatment and escalation of care to prevent harmful outcomes, such as inpatient mortality. EWS have had limited impact on clinical outcomes likely due to their primary reliance on vital signs, a late indicator of patient deterioration. When nurses are concerned about the potential for patient deterioration they increase surveillance of the patient and their respective nursing notes documentation in electronic health records (EHRs) [4–7]. Our team has developed an EWS named CONCERN (COmmunicating Narrative Concerns Entered by Registered Nurses) that leverages nursing surveillance and documentation patterns that reflect how nurses observe and monitor subtle changes in patients before deterioration is noted in their physiological conditions' parameters [1]. CONCERN is currently in production at 2 academic medical centers with implementation in progress at 2 more health systems [1].

The data from nursing documentation are large in volume and are structured, semi-structured, and time-varying. The large templated documentation from nursing notes also contains free text data written by nurses. These free text data can be useful as features in EWSs to predict patient health deterioration [1]. These free text data will act as an important feature in our CONCERN EWS [1]. Leveraging free text data can be challenging because of their large volume and clinical diversity [8,24,25]. Nursing EHR data are time-varying, semi-structured, and variable on a content level, which make the identification of the free text portion of notes a cumbersome task.

Nursing notes include: 1) templated documentation, which are structured data entered by nurses elsewhere in the chart, and 2) narrative (free text) information written by nurses in their own words. The free text may represent nurses' concerns about patients and can be useful in predictive modeling [1]. However, to leverage information from the free text documentation by nurses we first need to be able to identify where this free text resides within semi-structured nurses' notes and how to retrieve it. Often the narrative free text can be found embedded in other relatively structured texts, which is difficult to detect. Such data are not explicitly labeled as free text and can often be found intertwined within relatively structured texts, thereby making the detection difficult. The absence of clear distinctions between documents' information such as document headers and metadata, further adds up to the problem. This ultimately poses challenges in the automatic extraction of the free text data, which may contain important signals for improved clinical decision-making [1,9].

This research study is focused on HTML-based nursing notes from an academic medical center in the Northeastern United States. The dataset contains more than 200K notes with all free text (with no structured data in it), structured data with no free text, and free text embedded in structured data. Our study aimed to 1) identify and retrieve all narrative (free text) notes data, and 2) distinguish and retrieve the free text embedded in the nursing notes. In this

Heuristic-Based Extraction and Unigram Analysis of Nursing Free Text Data Residing in Large EHR Clinical Notes

Syed Mohtashim Abbas Bokhari[1(✉)], Kriste Krstovski[2,3], Jennifer Withall[1], Rachel Lee[4], Patricia Dykes[5,6], Mai Tran[1], Kenrick Cato[7,8], and Sarah Rossetti[1,4]

[1] Department of Biomedical Informatics, Columbia University, New York, NY, USA
mohtashim_abbas@yahoo.com
[2] Data Science Institute, Columbia University, New York, NY, USA
[3] Columbia Business School, Columbia University, New York, NY, USA
[4] School of Nursing, Columbia University, New York, NY, USA
[5] Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA
[6] BWH Center for Patient Safety, Research and Practice, Boston, MA, USA
[7] University of Pennsylvania, Philadelphia, PA, USA
[8] Children's Hospital of Philadelphia, Philadelphia, PA, USA

Abstract. Free text in nurses' notes can play an important role in clinical decision-making; however, such information has not been explored to the fullest of its potential as it is hard to extract it from electronic health records (EHRs). Free text is a subset of the information recorded in nursing notes. Automated extraction of free text is challenging due to EHRs' size and structural diversity. Understanding these structural and content-level differences is essential for the extraction. Free text is embedded in other relatively structured texts, which are difficult to detect automatically. Moreover, there is no information indicating whether a note is a free text. As a first step in automating the extraction process, we explore heuristic-based algorithms with the goal of establishing a baseline and developing an annotated dataset, which could then be used for further machine learning-based extraction algorithms for a more scalable solution. In this research, we analyze over 200,000 EHR notes and extract 40,000 free text notes from them. Furthermore, we use the unigram language model to analyze the differences between free and structured texts to better understand the free text content.

Keywords: nursing documentation · health informatics · clinical notes · nursing notes · heuristics · natural language processing · information retrieval · unigram analysis

1 Introduction

Nursing documentation, including the concepts written in nursing notes, can play an integral role in healthcare prediction models to inform effective clinical decision-making [1]. Early warning scores (EWS) are one type of prediction

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 124–136, 2024.
https://doi.org/10.1007/978-3-031-59717-6_9

55. Kylliäinen, A., Jones, E.J.H., Gomot, M., Warreyn, P., Falck-Ytter, T.: Practical guidelines for studying young children with autism spectrum disorder in psychophysiological experiments. Rev. J. Autism Dev. Disord. **1**, 373–386 (2014)
56. McCarthy, C., Pradhan, N., Redpath, C., Adler, A.: Validation of the empatica E4 wristband. In: 2016 IEEE EMBS International Student Conference (ISC), pp. 1–4 (2016)
57. Charlton, P.H., et al.: Extraction of respiratory signals from the electrocardiogram and photoplethysmogram: technical and physiological determinants. Physiol. Meas. **38**, 669 (2017)
58. Bonnici, T., Orphanidou, C., Vallance, D., Darrell, A., Tarassenko, L.: Testing of werable monitors in a real-world hospital environment: what lessons can be learnt? In: 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, pp. 79–84 (2012)
59. Joshi, M., et al.: Perceptions on the use of wearable sensors and continuous monitoring in surgical patients: interview study among surgical staff. JMIR Form. Res. **6**, e27866 (2022)
60. Areia, C., et al.: Experiences of current vital signs monitoring practices and views of wearable monitoring: a qualitative study in patients and nurses. J. Adv. Nurs. **78**, 810–822 (2022)
61. Zhang, Q., Zhou, D., Zeng, X.: Highly wearable cuff-less blood pressure and heart rate monitoring with single-arm electrocardiogram and photoplethysmogram signals. Biomed. Eng. Online **16**, 23 (2017)
62. Nelson, B.W., Low, C.A., Jacobson, N., Areán, P., Torous, J., Allen, N.B.: Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. npj Dig. Med. **3**, 1–9 (2020)

33. Jenkins, J.L., Valacich, J.S., Williams, P.: Human-computer interaction movement indicators of response biases in online surveys. In: ICIS 2017 Proceedings, pp. 1–16. Association for Information Systems (2017)
34. Varni, G., et al.: Interactive sonification of synchronisation of motoric behaviour in social active listening to music with mobile devices. J. Multimodal User Interfaces **5**, 157–173 (2012)
35. Manikandan, M.S., Soman, K.P.: A novel method for detecting R-peaks in electrocardiogram (ECG) signal. Biomed. Signal Process. Control **7**, 118–128 (2012)
36. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. BME **32**, 230–236 (1985)
37. Christov, I.I.: Real time electrocardiogram QRS detection using combined adaptive threshold. Biomed. Eng. Online **3**, 28 (2004)
38. Elgendi, M., Eskofier, B., Dokos, S., Abbott, D.: Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems. PLoS ONE **9**, 1–18 (2014)
39. Qin, Q., Li, J., Yue, Y., Liu, C.: An adaptive and time-efficient ECG R-peak detection algorithm. J Healthc, Eng. **2017**, 5980541 (2017)
40. Feldman, M.: Hilbert transforms. In: Braun, S., Ewins, D., Rao, S.S. (eds.) Encyclopedia of Vibration, pp. 642–648. Elsevier Ltd., Amsterdam (2001)
41. Sunami, N.: Shannon energy R peak detection (2020). https://github.com/nsunami/Shannon-Energy-R-Peak-Detection
42. Goldberger, A.L., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**, E215–E220 (2000)
43. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag. **20**, 45–50 (2001)
44. Chen, L., Reisner, A.T., Reifman, J.: Automated beat onset and peak detection algorithm for field-collected photoplethysmograms. In: Paper presented at the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, 3 Sept 2009 (2009)
45. Shin, H.S., Lee, C., Lee, M.: Adaptive threshold method for the peak detection of photoplethysmographic waveform. Comput. Biol. Med. **39**, 1145–1152 (2009)
46. Kuntamalla, S., Reddy, L.R.G.: An efficient and automatic systolic peak detection algorithm for photoplethysmographic signals. Int. J. Comput. Appl. **97**, 1–6 (2014)
47. Mishra, V., et al.: Evaluating the reproducibility of physiological stress detection models. Proc. ACM Interact. Mobile Wearable Ubiq. Technol. **4**, 1–29 (2020)
48. Graham, F.K.: Constraints on measuring heart rate and period sequentially through real and cardiac time. Psychophysiology **15**, 492–495 (1978)
49. Tanaka, H., Monahan, K.D., Seals, D.R.: Age-predicted maximal heart rate revisited. J. Am. Coll. Cardiol. **37**, 153–156 (2001)
50. Palumbo, R.V., Marraccini, M.E., Weyandt, L.L., Wilder-Smith, O., Liu, S., Goodwin, M.S.: Interpersonal autonomic physiology: a systematic review of the literature. Pers. Soc. Psychol. Rev. **21**, 99–141 (2016)
51. Kristiansen, J., et al.: Comparison of two systems for long-term heart rate variability monitoring in free-living conditions: a pilot study. Biomed. Eng. **10**, 1–14 (2011)
52. Goldman, S., Wang, C., Salgado, M.W., Greene, P.E., Kim, M., Rapin, I.: Motor stereotypies in children with autism and other developmental disorders. Dev. Med. Child Neurol. **51**, 30–38 (2009)
53. Rice, C.E., et al.: Reported wandering behavior among children with autism spectrum disorder and/or intellectual disability. J. Pediatr. **174**, 232-239.e2 (2016)
54. Sinzig, J., Walter, D., Doepfner, M.: Attention deficit/hyperactivity disorder in children and adolescents with autism spectrum disorder: symptom or syndrome? J. Atten. Disord. **13**, 117–126 (2009)

13. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Paper presented at the 20th ACM International Conference on Multimodal Interaction (ICMI 2018), New York, NY, 16 October 2018 (2018)
14. Quesnel, P.X., Chan, A.D.C., Yang, H.: Real-time biosignal quality analysis of ambulatory ECG for detection of myocardial ischemia. In: 2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–5 (2013)
15. Bae, T.W., Kwon, K.K.: ECG PQRST complex detector and heart rate variability analysis using temporal characteristics of fiducial points. Biomed. Signal Process. Control **66**, 1–21 (2021)
16. Clifford, G., Behar, J., Li, Q., Rezek, I.: Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiol. Meas. **33**, 1419–1433 (2012)
17. Daluwatte, C., Johannesen, L., Galeotti, L., Vicente, J., Strauss, D.G., Scully, C.G.: Assessing ECG signal quality indices to discriminate ECGs with artefacts from pathologically different arrhythmic ECGs. Physiol. Meas. **37**, 1370–1382 (2016)
18. Sharma, L.N., Dandapat, S., Mahanta, A.: Kurtosis based multichannel ECG signal denoising and diagnostic distortion measures. In: TENCON 2009 - 2009 IEEE Region 10 Conference. pp. 1–5. IEEE, New York (2009)
19. Zaunseder, S., Vehkaoja, A., Fleischhauer, V., Hoog Antink, C.: Signal-to-noise ratio is more important than sampling rate in beat-to-beat interval estimation from optical sensors. Biomed. Signal Process. Control **74**, 103538 (2022)
20. Castiglioni, P., Parati, G., Faini, A.: Cepstral analysis for scoring the quality of electrocardiograms for heart rate variability. Front. Physiol. **13**, 1–13 (2022)
21. He, R., et al.: Reducing false arrhythmia alarms in the ICU using novel signal quality indices assessment method. Paper presented at the 2015 Computing in Cardiology Conference, New York, NY, 6 September 2015 (2015)
22. Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., Tarassenko, L.: Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. IEEE J. Biomed. Health Inf. **19**, 832–838 (2014)
23. Behar, J., Oster, J., Li, Q., Clifford, G.D.: A single channel ECG quality metric. In: 2012 Computing in Cardiology, pp. 381–384. IEEE, Kraków (2012)
24. Kazemi, K., Laitala, J., Azimi, I., Liljeberg, P., Rahmani, A.M.: Robust PPG peak detection using dilated convolutional neural networks. Sensors **22**, 1–22 (2022)
25. Bizzego, A., Battisti, A., Gabrieli, G., Esposito, G., Furlanello, C.: Pyphysio: a physiological signal processing library for data science approaches in physiology. SoftwareX **10**, 1–5 (2019)
26. Carreiras, C., Alves, A.P., Lourenço, A., Canento, F., Silva, H.P. da, Fred, A.: BioSPPy: Biosignal processing in Python (2015). https://biosppy.readthedocs.io/
27. Kramer, L., Menon, C., Elgendi, M.: ECGAssess: a python-based toolbox to assess ECG lead signal quality. Frontiers in Digital Health. **4**, 1–9 (2022)
28. Makowski, D., et al.: NeuroKit2: a python toolbox for neurophysiological signal processing. Behav. Res. Methods **53**, 1689–1696 (2021)
29. van Gent, P., Farah, H., van Nes, N., van Arem, B.: HeartPy: a novel heart rate algorithm for the analysis of noisy signals. Transport. Res. F: Traffic Psychol. Behav. **66**, 368–378 (2019)
30. Vollmer, M.: HRVTool—an open-source MATLAB toolbox for analyzing heart rate variability. In: 2019 Computing in Cardiology (CinC), pp. 1–4. IEEE, Singapore (2019)
31. Blechert, J., Peyk, P., Liedlgruber, M., Wilhelm, F.H.: ANSLAB: integrated multichannel peripheral biosignal processing in psychophysiological science. Behav. Res. Methods **48**, 1528–1545 (2016)
32. Plotly: Dash Python user guide. https://dash.plotly.com/

7 Conclusion

This paper presents HeartView, an extensible signal quality assessment pipeline with a web-based visualization dashboard for ambulatory cardiovascular data. We developed HeartView using open-source libraries and frameworks in Python. We assessed our pipeline using an ECG dataset collected from children with and without autism and the publicly available WESAD dataset. Our tool has a singular advantage over most extant cardiovascular signal pre-processing and quality assessment approaches. We offer an open-source pipeline with a user-friendly web interface that summarizes signal quality metrics through interactive visualizations and a summary table. A free and well-documented user interface can increase accessibility to signal quality assessment procedures historically only available to researchers with computer science and electrical engineering backgrounds. As a result, our signal quality assessment dashboard can contribute to more methodological transparency, reproducibility, and rigor that empowers researchers from diverse methodological backgrounds to make more informed decisions about the reliability and validity of their data when ambulatory biosensor data collection systems are used.

References

1. Goodwin, M.S.: Passive telemetric monitoring: novel methods for real-world behavioral assessment. In: Handbook of Research Methods for Studying Daily Life. pp. 251–266. Guilford Press, New York (2012)
2. Mishra, V., et al.: Continuous detection of physiological stress with commodity hardware. ACM Trans. Comput. Healthcare **1**, 1–30 (2020)
3. Weenk, M., et al.: Continuous monitoring of vital signs using wearable devices on the general ward: Pilot study. JMIR Mhealth Uhealth **5**, 1–12 (2017)
4. Pantelopoulos, A., Bourbakis, N.G.: A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**, 1–12 (2010)
5. Orphanidou, C.: Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations. Springer, Cham (2018)
6. Fine, J., et al.: Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. Biosensors **11**, 1–36 (2021)
7. Madhavan, G.: Plethysmography. Biomed. Instrum. Technol. **39**, 367–371 (2005)
8. Kusumoto, F.: ECG Interpretation. Springer Nature, Cham (2020)
9. Elgendi, M.: Optimal signal quality index for photoplethysmogram signals. Bioengineering (Basel) **3**, 21 (2016)
10. Quintana, D.S., Alvares, G.A., Heathers, J.A.: Guidelines for reporting articles on psychiatry and heart rate variability (GRAPH): recommendations to advance research communication. Transl. Psychiatry **6**, 1–10 (2016)
11. Stupple, A., Singerman, D., Celi, L.A.: The reproducibility crisis in the age of digital medicine. NPJ Digital Med. **2**, 3 (2019)
12. Yamane, N., Mishra, V., Goodwin, M.S.: Developing an open-source web-based data quality assessment pipeline for analysis of ambulatory cardiovascular data in individuals with autism. In: Paper presented at the International Society for Autism Research Annual Meeting (INSAR 2023), Stockholm, Sweden, 3 May 2023 (2023)

with ASD, as well as applying specific inclusionary and exclusionary criteria (e.g., tolerating certain fidgeting movements in the ASD group, analyzing the same number of trials and time points in both groups) to pre-processing procedures [55]. Motion and behavioral data can be used to formulate these criteria if available. Analyses of HeartView's signal quality metrics with motion data also support decision-making about excluding specific data points or applying additional signal-cleaning procedures on ambulatory cardiovascular data. Further, HeartView's signal quality metrics can support researchers during feasibility testing of experimental protocols involving young children or children with behavioral challenges and wearable cardiovascular devices in pilot studies.

Within the WESAD dataset, our pipeline captured differences in data missingness and the number of invalid segments between PPG and ECG recordings from two different devices. Compared to ECG data from the RespiBAN, PPG data from the Empatica E4 contained more invalid segments and missing peaks per segment. This finding is consistent with those of previous studies comparing the data quality of cardiovascular signals derived from ECG and PPG devices [56, 57]. However, such discrepancies in signal quality could be due to a couple of confounding reasons. First, because PPGs from the Empatica E4 were recorded using fewer inputs than ECGs from the RespiBAN (i.e., a single input versus three leads, respectively), the Empatica E4 had an inherently greater likelihood of collecting data with missing peaks or invalid segments. Second, while the RespiBAN is meant to be worn on the chest, the Empatica E4 is worn on the wrist. Differences in sensor and electrode placement sites result in different levels and types of artifacts. For example, Zhang, Zhou, and Zeng [61] found that ECG data collected from sites on the upper arm were more susceptible to motion artifacts and muscle noise than ECG data collected from the chest. While we did observe differences between wrist-derived PPG and chest-derived ECG data, a limitation of the WESAD dataset is that recordings are limited to only one placement site per signal type. Future work should evaluate HeartView based on differences in signal quality using at least two placement sites per device.

The current functions and feature set of the HeartView dashboard limit its data processing and quality assessment to cardiovascular data collected with the Actiwave Cardio, Empatica E4, and RespiBAN. To increase HeartView's functional generalizability across devices, future iterations of the pipeline and dashboard will incorporate a data transformer to streamline data pre-processing by translating data from various devices by brands commonly used in research (e.g., Polar, Bittium, Shimmer, etc.) into a single format. Further, we plan to add more data processing algorithms and SQA functions specific to different devices and sensor types. For example, researchers may benefit from various algorithmic options at different pipeline steps to evaluate their relative performance when assessing signal quality. Such algorithms may incorporate acceleration measures to corroborate cardiovascular data quality. Given differences in device capabilities to output higher-level values such as HR and IBI (or lack thereof) [62], we also plan to add an IBI detection algorithm for raw PPG data. This need for additional functionality specific to devices and sensor types highlights the importance of an extensible, community-driven approach to software development, whereby researchers and developers use their own datasets to contribute to open-source code and provide valuable feedback on usability, utility, and reproducibility.

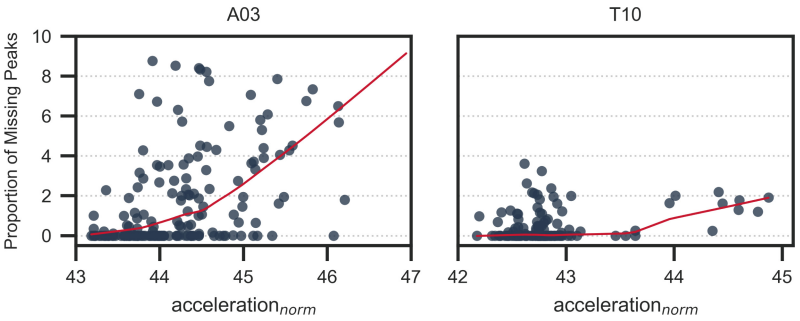


Fig. 5. Scatterplots of proportions of missing peaks per segment and normalized area-under-the-curve acceleration magnitude values in 'A03' and 'T10.' The red line represents locally weighted/estimated scatterplot smoothing curves.

6 Discussion and Future Work

HeartView is an extensible, open-source signal quality assessment pipeline with a web-based dashboard. With the HeartView dashboard, we provide researchers a tool to visually inspect ambulatory cardiovascular data's signal quality (i.e., numbers of missing beats and invalid segments) on a web browser using open-source Python libraries and frameworks. Assessments of basic quality and physiological feasibility of cardiovascular signals from wearable systems can improve the reliability of their measurements—an issue that continually precludes more widespread adoption in clinical contexts [58–60]. Our evaluation of the HeartView pipeline revealed group differences in signal quality within two datasets that support its incipient internal and external validity. We also present and discuss results from two randomly selected cases to demonstrate HeartView's ability to inform decisions about subsequent data processing procedures.

Within the enTRAIN dataset, the HeartView pipeline captured signal quality differences between ECG recordings of TD children and children with ASD. Specifically, ECG recordings of children with ASD contained significantly more invalid segments and missing peaks per segment than those of TD children. Our results are congruent with previous work by Kylliäinen and colleagues [55], who report more artifactual data collected from children with ASD or other developmental delays than from TD children, due to stereotypic behaviors and sensory differences in children with ASD. Further visual inspection of raw ECG signals with high percentages of missing and invalid data also revealed many segments containing motion artifacts and relatively weak or flat line signals. Indeed, our correlation analyses for two enTRAIN cases suggest that data missingness can be attributable to various noise sources. As observed in the case of 'A03,' data missingness can result from increased physical motion. However, it can also result from sources unrelated to physical motion, as observed in the case of 'T10.' In such instances, researchers may benefit from corroborating physiological data quality with ground truth labels, such as video-based annotations of gross motor movements (e.g., running around the observation room, fiddling with the ambulatory recording device). Additionally, researchers may decide to devote more time to cleaning data from children

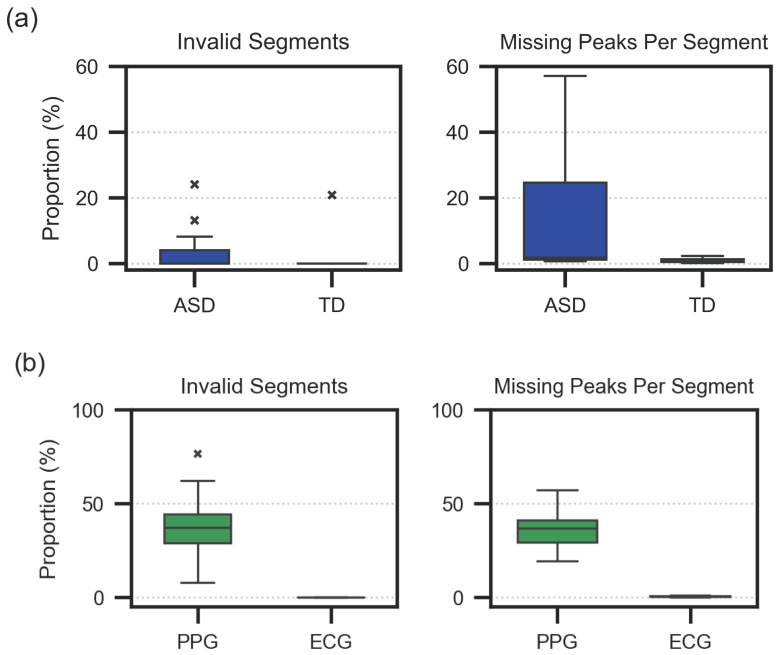


Fig. 4. Comparison of signal quality metrics between data of: (a) children with autism spectrum disorder (ASD) and typically developing (TD) children in the enTRAIN dataset; (b) photoplethysmograph (PPG) and electrocardiograph (ECG) devices in the WESAD dataset.

5.2 Cases: 'A03' and 'T10'

We randomly selected cases 'A03' and 'T10' in the enTRAIN dataset to demonstrate a use case for HeartView's signal quality metrics informing data processing procedures. Overall, across $n = 173$ sliding segments, 'A03's ECG recording contained an average of 3.47% ($SD = 6.88\%$) missing peaks per segment. Spearman's rank correlation coefficients for 'A03's proportions of missing peaks and normalized AUC values of acceleration magnitude also reveal a moderate, positive relationship between physical motion and data missingness ($\rho = .52$, $p < .001$). Across $n = 149$ sliding segments in 'T10's ECG recording, we found an average of 0.94% ($SD = 4.64\%$) missing peaks per segment and a non-significant correlation between physical motion and data missingness ($\rho = .12$, $p = .14$). Figure 5 illustrates the relationships between ECG data missingness and physical motion for 'A03' and 'T10.'

and muscle noise, followed by powerline interference filters. All enTRAIN and WESAD recordings were segmented into 60-s windows and then trimmed to the start of the first experimental condition and the end of the last experimental condition for each participant before testing for differences.

We also present a potential use case of HeartView's signal quality metrics with data from one randomly selected participant with ASD and one randomly selected TD participant from the enTRAIN dataset. Considering previous findings [52–55] suggesting that ambulatory physiological data from children with ASD may be noisier than that of TD children due to increased physical activity, we sought to demonstrate that HeartView can reveal the relationship between ECG signal quality and physical motion in children. For each enTRAIN participant, the Actiwave Cardio device also contained a 3-axis accelerometer that collected acceleration data at 32 Hz. First, we smooth the acceleration data using a quarter-second moving average filter and then calculate acceleration magnitude values. We then compute second-by-second AUC values of acceleration magnitude using Riemann sums and normalize each value using min-max normalization so that each normalized AUC value is scaled to the range of [0, 1]. Next, we aggregate all AUC values into 60-s sliding windows with 15-s steps. Here, we use a sliding window approach to account for any time lags between physical motion and signal artifact onsets. Therefore, in each 60-s window, the AUC value of acceleration falls in the range of [0, 60]. Finally, we compute Spearman's rank correlation coefficients to evaluate correlations between AUC values and proportions of missing R peaks across all sliding windows.

5 Results

Below, we present the results of our assessment of the HeartView pipeline and demonstrate a use case of HeartView's signal quality assessment metrics with two randomly selected subjects from the enTRAIN dataset.

5.1 Group Differences

Overall, the results of our analyses substantiate our hypotheses. Among the ECG recordings of children with ASD in the enTRAIN study, we found greater proportions of invalid segments with a moderate effect size ($U = 156.0$, $p = .03$, $r = 0.32$) and average proportions of missing peaks per segment with a large effect size ($U = 210.0$, $p = .001$, $r = 0.53$) compared to the ECG recordings of TD children (see Fig. 4a).

Across all WESAD participants, Wilcoxon signed-rank tests revealed that PPG data recorded contained significantly more proportions of invalid segments ($z = -4.17$, $p < .001$, $r = 0.76$) and average proportions of missing peaks per segment ($z = -4.17$, $p < .001$, $r = 0.76$) than ECG data with large effect sizes (see Fig. 4b).

Stress and Affect Detection (WESAD). All adult participants and children's caregivers provided written informed consent to participate in the research before data collection.

The enTRAIN study [12] was carried out to investigate socio-affective behavior in 23 typically developing (TD) children and 11 children with autism spectrum disorder (ASD) while collecting interpersonal physiological data [50] during a series of standardized social-emotional regulation tasks. Children's cardiovascular data in the enTRAIN dataset comprise a total of 82.6 h ($M = 2.4$, $SD = 0.6$) of raw ambulatory ECG recorded at 1024 Hz with the chest-worn Actiwave Cardio by CamNtech, which has demonstrated good reliability and validity with gold-standard cardiovascular measures [51]. Children in the study had a mean age of 4.0 years ($SD = 1.1$).

WESAD [13] is a publicly available dataset featuring physiological and motion data recorded from 15 healthy adults during 'neutral,' 'stress,' and 'amusement' affective states. We used 24.1 h ($M = 1.6$, $SD = 0.2$) of ECG recorded at 700 Hz with the chest-worn RespiBAN and 29.7 h ($M = 2.0$, $SD = 0.2$) of PPG simultaneously recorded at 64 Hz from the Empatica E4. Participants in the WESAD study had a mean age of 27.5 years ($SD = 2.4$).

4.2 HeartView Assessment

HeartView outputs metrics regarding the basic quality (i.e., whether beats are identifiable) and physiological feasibility (i.e., whether IBI values are valid) of cardiovascular data and visualizes them on a web-based dashboard. We evaluate the HeartView pipeline by assessing whether group differences in signal quality can be captured in each dataset. Specifically, we tested differences in the numbers of missing peaks per segment and invalid segments between TD children and children with ASD in enTRAIN and between PPG and ECG recordings in WESAD. We hypothesized the following:

H_1: Our pipeline can capture group differences in the number of missing peaks and invalid segments between (a) ECG recordings of TD children and children with ASD and (b) PPG recordings from the Empatica E4 and ECG recordings from the RespiBAN devices.

Our rationale for H_{1a} is based on the observation that children with ASD display increased motor stereotypies [52], wandering behaviors [53], and symptoms of attention-deficit/hyperactivity disorder [54] compared to typically developing peers. As a result, these increased movement behaviors are likely to introduce more frequent signal artifacts into data collected from this population in psychophysiological experiments [55].

Our rationale for H_{1b} is based on previous work demonstrating that ECG devices tend to provide higher-quality data than PPG devices, which are subject to motion artifact [56, 57]. This discrepancy is likened to differences in sampling rate and mechanical configuration (e.g., optical sensors versus electrodes). Thus, we expect the data quality of PPG recordings to be impacted by signal artifacts more than that of ECG recordings.

Because data distributions were found to be non-normal for all groups with quantile-quantile plots and the Shapiro-Wilk test, we tested group differences with the Mann-Whitney test in the enTRAIN dataset and with the Wilcoxon signed-rank test for paired PPG and ECG recordings in the WESAD dataset. PPG and ECG data from the WESAD dataset is time-synchronized using the timestamps recorded by the RespiBAN upon initialization. In all ECG recordings, we first apply filters to eliminate baseline wander

missing numbers of beats per segment; and (3) a line chart of the raw and filtered cardiovascular data with an overlaying scatterplot of detected peak locations. Two separate buttons are provided to access additional line charts of the corresponding IBI series and raw accelerometer data. We also include a range slider tied to a callback function that takes user-selected segment values and outputs a filtered view of the line charts. For example, in the top-right view in Fig. 3, the bottom panels contain line charts displaying raw and filtered ECG data from the Actiwave Cardio with points denoting locations of detected peaks within the first 20 s of segment 5.

Fig. 3. Multiple panels of the HeartView dashboard. *Top left*: Launch view of the off-canvas containing user input elements; *Top right:* Dashboard view of electrocardiograph (ECG) signal quality assessment and visualization of raw and filtered ECG with peaks; *Bottom left*: Dashboard view with a visualization of inter-beat interval (IBI) series; *Bottom right:* Dashboard view with a visualization of raw acceleration data.

4 Methods

In the following subsections, we discuss the datasets and data pre-processing and analysis procedures used in our assessment of the HeartView pipeline.

4.1 Datasets

Two datasets were used to assess the utility and incipient internal and external validity of HeartView. We leverage cardiovascular data from two studies, enTRAIN and Wearable

Signal Quality Metrics. HeartView performs SQA on the basic quality (i.e., whether peaks are identifiable for reliable HR and IBI extraction) and physiological feasibility of a signal (i.e., whether the number of extracted peaks is valid) across segments of a user-customizable length. Thus, the pipeline measures signal quality based on the number of missing peaks per segment and invalid signal segments.

Missing Peaks. HeartView determines the number of missing peaks against an expected number of peaks. The pipeline derives this expected number of peaks by computing the median of all second-by-second HR values observed within each segment. We chose to use the median of all second-by-second HR values given its robustness to outliers. Second-by-second HR values are derived with the following steps. First, for each calculated IBI value, the pipeline computes a HR value by dividing the IBI value from 60,0000. Next, each second-by-second HR value is calculated using the harmonic mean of HRs (i.e., the reciprocal of the mean of the reciprocals of HRs) observed in a 2-s window based on Graham's approach [48].

$$\bar{x} = \frac{n}{\sum {}^{1}/_{x_i}} \tag{1}$$

In (1), n represents the number of HR values, and x_i represents a HR value at a timepoint i within the 2-s window. Thus, $\bar{x}$ represents the harmonic mean HR at one second, and the expected number is set to the median of all observed $\bar{x}$ in one segment. If the number of detected peaks is greater than the expected number of peaks in a segment, the missing number of peaks in that segment is set to zero; otherwise, the pipeline derives the number of missing peaks by calculating the difference between the number of detected peaks and the expected number of peaks.

Invalid Peaks. After peak extraction, HeartView counts invalid segments based on whether the number of detected peaks in a segment falls outside the range of [30:220] bpm. Based on prior work [2, 47, 49], the upper bound is set to the maximum human HR value, and the lower bound is set conservatively to a value close to the lower bound of the human physiological range.

3.2 Dashboard

We developed the HeartView dashboard using the open-source Dash framework, which consists of a Flask server that communicates with front-end React components [32]. Multiple callback functions with user input and state arguments are mapped to Dash core components, including a file upload component, data segmentation field, and checklist buttons corresponding to filter types (baseline wander, muscle activity, and powerline interference). These functions are then called separately to output interactive charts and a summary table. All user input components are contained within an off-canvas menu that can be toggled to appear or disappear from the left side of the screen.

As illustrated in Fig. 3, the main dashboard contains three separate panels: (1) a data summary panel, including information about the uploaded data file, computed signal quality metrics, and an export button to save all pre-processed data and signal quality information; (2) a bar chart with overlaying bars corresponding to the expected and

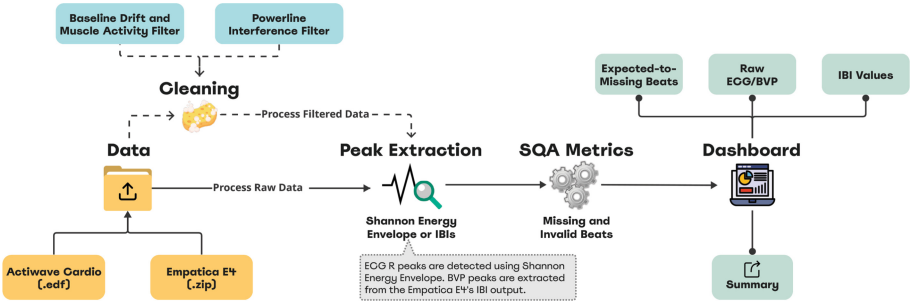


Fig. 2. HeartView pipeline architecture.

eliminate noise from the ECG data. For example, baseline wander and muscle noise can be eliminated using a bandpass filter between 0.5 and 45 Hz, and powerline interference at 60 Hz can be removed using a notch filter with a quality factor of 20.

PPG. Each archive file of the Empatica E4 comprises a set of CSV files containing raw and pre-processed data for HRV analysis. HeartView reads pre-processed inter-beat interval (IBI) values for later peak extraction and raw BVP values with timestamps for plotting and visual inspection purposes.

Acceleration. HeartView can also extract and process acceleration data from EDF, CSV, and archive files of the Actiwave Cardio, Empatica E4, and RespiBAN. The pipeline contains functions for smoothing raw data, converting from g-force to meters per second squared, and computing area under the curve (AUC) of acceleration magnitude. AUC is a commonly used proxy for movement over a given time window [33, 34], particularly when complex time or frequency-domain features are not under consideration at the SQA stage. In the present context, a higher AUC value indicates greater motion.

Peak Extraction. HeartView identifies peak locations from filtered data using the algorithm by Manikandan and Soman [35] for ECG data and from pre-processed IBI output from the Empatica E4 for PPG data[2]. R peaks from the ECG waveform are automatically detected using a Shannon Energy Envelope (SEE) estimator, peak-finding logic based on the Hilbert-transform [40], and zero-crossing point detection [35, 41]. This R peak detection algorithm has been validated using ambulatory ECG recordings from the MIT-BIH arrhythmia database [42, 43], demonstrating 99.8% average detection accuracy, 99.9% sensitivity, and 99.8% positive prediction. Peak locations in the PPG waveform are identified using timestamps from the pre-processed IBI file output provided by Empatica. Although multiple PPG peak detection algorithms exist [24, 44–46], HeartView uses the IBI time series provided by Empatica, given the device's widespread use [47] and the fact that existing algorithms have not been validated on datasets that are standardized to assess the performance of PPG peak detection algorithms.

[2] As our primary intent is to introduce a tool for researchers to perform initial assessment checks in their SQA, HeartView currently includes only one of several possible algorithms for ECG beat detection [36–39]. Future users could select and implement additional or alternative state-of-the-art algorithms.

three implement a GUI. ECGAssess is Python-based software that performs automated SQA and binary classification of the acceptability of multi-lead ECG data collected in clinical contexts for medical diagnosis [27]. ANSLab [31] provides both open-source and licensed, closed-source MATLAB-based software options through OpenANSLab and ANSLab Professional, respectively. The software suite contains modules that allow physiological data pre-processing on text files, artifact editing, and analysis. Additional functionalities in ANSLab Professional include batch processing, HRV analysis, and reading of multiple file types. To our knowledge, ANSLab is the only other GUI-based software suite with functions to process PPG data. Additionally, it is the only software suite capable of generating and exporting configuration files, a functionality that we plan to add to a future iteration of HeartView. Although open-source, OpenANSLab does not run as a standalone executable application and thus requires an installation of MATLAB, which is not free and therefore inaccessible to many outside of academia. In contrast, HRVTool is a standalone MATLAB application that performs ECG data processing and HRV feature extraction [30].

While the abovementioned open-source tools are valuable, uses are generally limited to those with programming skills. Further, some possess functionalities restricted by paywalls. In contrast, we propose a data pre-processing pipeline with an accompanying free, GUI-based solution for researchers without programming skills to perform necessary preliminary checks for basic quality and physiological feasibility of both ECG and PPG data. Our approach delivers an open-source, well-documented, and extensible web interface intended to increase efficiency and accessibility to a broader range of researchers who may not otherwise be able to conduct rigorous SQA on their ambulatory cardiovascular data.

3 HeartView Overview

We developed HeartView in Python 3.9 and its accompanying web-based dashboard using Plotly's Dash framework (version 2.8.1). Dash is an open-source Python framework built on Flask, Plotly.js, and React [32].

3.1 Data Processing Pipeline

The HeartView pipeline performs three main procedures (see Fig. 2) before outputting summary information on the dashboard: (1) data pre-processing (i.e., transformation and cleaning); (2) peak extraction; and (3) SQA metric computation.

Data Pre-processing. HeartView begins by reading and transforming raw ECG, PPG, and accelerometer data into Pandas data frames using device-specific file reading functions. Acceptable file types include European Data Format (EDF) files from the Actiwave Cardio and archive files from the Empatica E4. In addition, HeartView uses Pandas to read and pre-process comma-separated value (CSV) files generated from these devices, as well as the RespiBAN.

ECG. HeartView extracts timestamps and raw ECG values in units of millivolts from each EDF or CSV file of the Actiwave Cardio. Next, optional filters are applied to

indices (SQIs) to detect signal artifacts. SQIs are statistical or machine learning-based measures that heuristically describe characteristics and acceptability of signal waveforms [14–17] and are thus binary in most cases (i.e., "acceptable" or "unacceptable") [5]. Statistical SQIs may include kurtosis [18], skewness [16], signal-to-noise ratio (SNR) [19], and signal power [20]. Present machine learning-based SQIs commonly include measurements derived with support vector machine classifiers [21, 22] and neural networks [23, 24].

Several open-source data processing software and libraries [25–31] are available and can be applied to SQA of cardiovascular data. Some of the most popular data processing Python packages, such as NeuroKit2 and pyphysio, also support the computation of common statistical SQIs, including kurtosis and SNR [25, 28]. Other physiological data processing libraries are available for assessing basic quality and physiological feasibility checks, including filtering and visualizing signals and deriving peaks. BioSPPy, for example, provides a library of standard biosignal processing functions and feature extraction algorithms, including filtering, QRS complex detection, and visualization [26]. NeuroKit2, a community-driven Python package, contains functions to derive different types of peaks, filter signals, and compute HR [28]. Table 1 presents several software packages and libraries and their available features and functions relevant to the SQA of ambulatory PPG and ECG data.

Table 1. Overview of popular cardiovascular data processing software and libraries.

Package	FR	CE	SF	BD	V	CA	PPG	ECG	GUI
ANSLab	◐	●	●	●	●	◐	●	●	●
BioSPPy	○	○	●	●	●	●	●	●	○
ECGAssess	●	○	●	●	●	●	○	●	●
HeartPy	○	○	●	●	●	●	●	●	○
HRVTool	●	○	○	●	●	●	○	●	●
NeuroKit2	○	○	●	●	●	●	●	●	○
pyphysio	○	○	●	●	○	●	●	●	○
HeartView	●	⊕	●	●	●	●	●	●	●

○ Not existing ◐ Partially existing ● Completely existing ⊕ In development

Acronyms: *FR* = File reader; *CE* = Configuration exporter; *SF* = Signal filtering; *BD* = Beat detection; *V* = Visualization; *CA* = Code available; *PPG* = Photoplethysmography; *ECG* = Electrocardiography; *GUI* = Graphical user interface

In an informal survey of user needs distributed by one of our co-authors to 421 researchers and engineers[1] who process and analyze physiological data, 78% of respondents favored using well-documented, open-source software with user-friendly graphical user interfaces (GUIs). Based on our audit of existing popular open-source tools, only

[1] The survey sample comprised 31% researchers and 69% engineers from the Society of Psychophysiological Research (SPR), the IEEE International Machine Learning for Signal Processing (MLSP) workshop, and snowball sampling using personal contacts and social media.

Wearable devices use multiple biosensors to capture peripheral physiological signals. In the case of cardiovascular activity, photoplethysmography (PPG), to measure blood volume pulse (BVP), and electrodes, to record electrocardiograph (ECG) data, are the most common. PPG is an optical method of measuring volumetric changes in blood profusion [7]. Pulse rate, a proxy for heart rate (HR), is a function of the changes in light absorbed or reflected by blood flowing through a particular measurement site. ECG is a technique for observing and recording changes in heart electrical activity.

Before deriving summary statistics and making inferences based on PPG and ECG data, artifacts should be inspected using the expected morphological characteristics and dynamics of PPG and ECG signals. In a PPG waveform, each dominant peak represents a change in the absorption or reflection of light due to increased blood flow during systole (i.e., heart contraction). Similarly, the ECG signal contains three waveforms—the P, QRS, and T—corresponding to independent events in the cardiac cycle [8]. The most dominant wave is the QRS complex, which represents the depolarization of the heart's ventricles as a contraction begins. Correct detection of R peaks ensures that QRS complexes are captured, thus confirming valid heartbeats [9].

The primary goal of physiological SQA is to identify outliers, signal artifacts, and missingness to increase the reliability and validity of physiological measurements; however, the process varies across research teams. There is no unified approach to assessing signal quality in biosensor data [5], as standardized and transparent reporting of custom data preprocessing procedures is lacking [10]. In addition, most SQA procedures exist in closed-source pipelines, limiting reproducibility and uniformity across studies [11]. We developed HeartView to increase the reproducibility of and accessibility to SQA procedures typically performed only by trained researchers with computational backgrounds.

HeartView is a Python-based, open-source, extensible SQA pipeline and dashboard that visualizes and summarizes segment-by-segment quantification of missing and invalid beats in ambulatory cardiovascular data obtained in research contexts. SQA of a signal's basic quality and physiological feasibility is essential for making informed decisions about further data cleaning and processing procedures [5]. SQA of basic quality addresses whether beats are identifiable for reliable HR and heart rate variability (HRV) calculation. At the same time, physiological feasibility describes whether HR and inter-beat interval (IBI) values are valid. We demonstrate the utility of our pipeline in assessing the quality of physiological signals on two datasets covering different use cases: ECG data collected from children with and without autism spectrum disorder (ASD) [12], and a publicly available dataset containing PPG and ECG data collected from healthy adults [13].

2 Related Work

The level of SQA one performs depends on research integrity and clinical purpose. For instance, additional algorithmic development may be necessary in clinical contexts to assess specific waveform characteristics—e.g., whether the P, QRS, or T waves are identifiable—to diagnose conditions like myocardial ischemia [14] and heart disease [15]. Indeed, most work on SQA uses clinical datasets to derive and evaluate signal quality

commercially available and becoming progressively smaller and lighter. However, due to their size and ambulatory nature compared to traditional stationary systems, modern wearable system signals are more susceptible to artifacts, increasing missing or distorted data. Common sources of ambulatory signal artifacts include powerline interference, baseline wander, muscle activity, physical movement, and pressure disturbance [5, 6]. Figure 1 illustrates examples of ambulatory signal corruption by different artifacts.

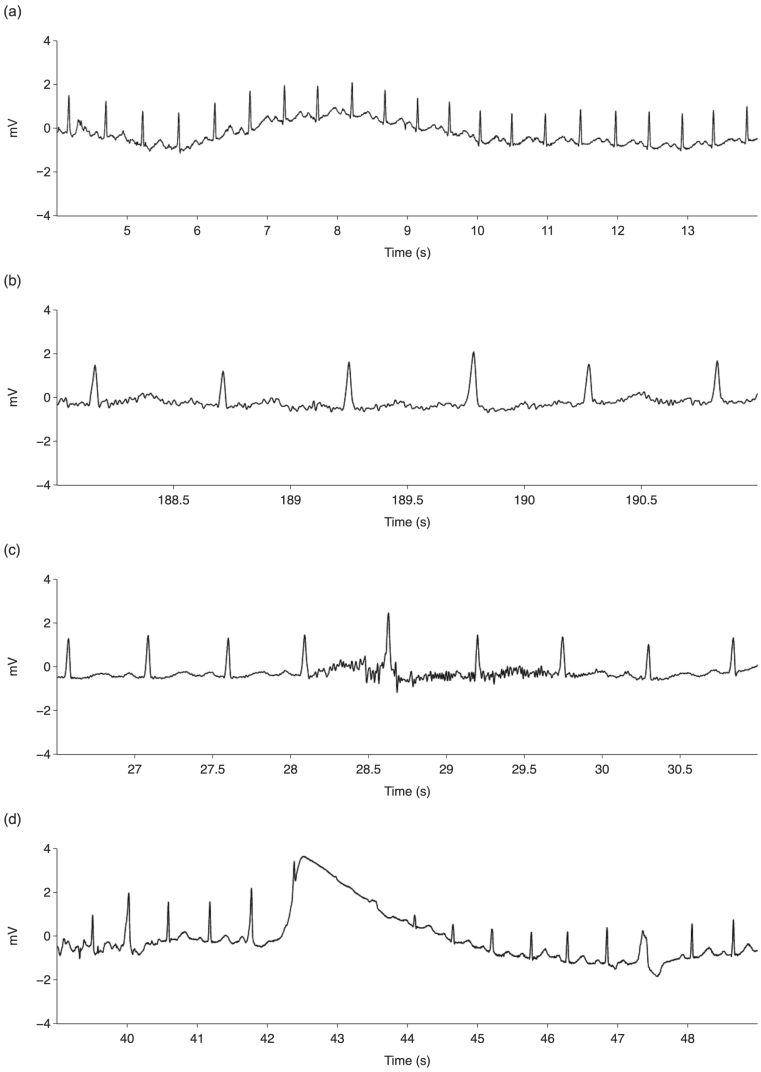


Fig. 1. Ambulatory electrocardiograph signals corrupted with (a) baseline wander, (b) powerline interference, (c) muscle activity, and (d) pressure disturbance.

HeartView: An Extensible, Open-Source, Web-Based Signal Quality Assessment Pipeline for Ambulatory Cardiovascular Data

Natasha Yamane(✉), Varun Mishra, and Matthew S. Goodwin

Khoury College of Computer Sciences and Bouvé College of Health Sciences, Northeastern University, Boston, MA 02115, USA
{yamane.n,v.mishra,m.goodwin}@northeastern.edu

Abstract. Wearable sensing systems enable peripheral physiological data to be collected repeatedly in naturalistic settings. However, the ambulatory nature of wearable biosensors predisposes them to common signal artifacts that researchers must address before analysis. Signal quality assessment procedures are time-consuming and non-standardized across research teams, and transparent reporting of custom, closed-source pipelines needs improvement. This paper presents HeartView, an extensible, open-source, web-based signal quality assessment pipeline that visualizes and quantifies missing beats and invalid segments in heart rate variability (HRV) data obtained from ambulatory electrocardiograph (ECG) and photoplethysmograph (PPG) signals. We demonstrate the utility of our pipeline on two datasets: (1) 34 ECGs recorded with the Actiwave Cardio from children with and without autism, and (2) 15 sets of ECGs and PPGs recorded with the RespiBAN and Empatica E4, respectively, from healthy adults in the publicly available WESAD dataset. Our pipeline demonstrates interpretable group differences in physiological signal quality. ECGs of children with autism contain more missing beats and invalid segments than those without autism. Similarly, PPG data contains more missing beats and invalid segments than ECG data. HeartView has a graphical user interface in the form of a web-based dashboard at https://github.com/cbslneu/heartview.

Keywords: Signal Quality Assessment · Data Pipelines · Ambulatory Cardiovascular Data · Electrocardiography · Photoplethysmography

1 Introduction

Signal quality assessment (SQA) involves detecting and evaluating outliers, artifacts, and missingness in signal-based data using expected signal morphology and dynamics. This procedure is an increasingly important step during and after data collection, as wireless ambulatory technologies are gaining popularity for their ability to monitor physiological states continuously and unobtrusively in both research and clinical settings [1–4]. Many wearable devices that capture peripheral physiological signals in free-living contexts are

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 107–123, 2024.
https://doi.org/10.1007/978-3-031-59717-6_8

Datasets and Big Data Processing

20. Pranvera Beqiraj, M.: The right to be heard in the European Union – case law of the court of justice of the European Union. Europ. J. Multidisciplinary Stud. **1**(1), 264–269 (2016)
21. European Court of Justice.: *Estel NV v Commission of the European Communities* App no 270/82, [13–16] (1984)
22. General Court.: *Eyckeler & Malt AG v Commission of the European Communities* App no T-42/96, [79–80] (1998)
23. European Court of Justice.: *Aalborg Portland A/S, Irish Cement Ltd, Climents français SA, Italcementi – Fabbriche Riunite Cemento SpA, Buzzi Unicem SpA and Cementir – Cementerie del Tirreno SpA v Commission of the European Communities* App no Cases C-204/00 P, C-205/00 P, C-211/00 P, C-213/00 P, C-217/00 P and C-219/00 P, [126] (2004)
24. General Court.: *Solvay SA v Commission of the European Communities* App no T-30/91, [88–92] (1995)
25. European Court of Justice.: *Club Hotel Loutraki AE, Vivere Entertainment AE, Theros International Gaming, Inc., Elliniko Casino Kerkyras, Casino Rodos, Porto Carras AE and Kazino Aigaiou AE v European Commission*, App no C-131/15 P, [46] (2016)

References

1. König, I.R., et al.: What is precision medicine? Europ. Respiratory J. **50**(4), 1–12 (2017)
2. The European Parliament and the Council of the European Union.: Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU *OJ* [2017] L 117/176
3. Finck, M.: Automated decision-making and administrative law. In: Cane, P. et al. (eds.), The Oxford Handbook of Comparative Administrative Law, vol. 1, 657–675 (2021)
4. The European Parliament, the Council and the Commission.: Charter of Fundamental Rights of the European Union *OJ* [2012] C 326/391
5. Kańska, K.: Towards Administrative Human Rights in the EU. Impact of the Charter of Fundamental Rights. Europ. Law J. **10**(3), 296–326 (2004)
6. Borh, A., Memarzadeh, K.: The rise of artificial intelligence in healthcare applications. In: Bohr, A., Memarzadeh, K. (eds.) Artificial Intelligence in Healthcare, vol. 1, pp. 25–60. Academic Press, London (2020)
7. Zhao, Y.: CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine **61**(103030), 1–14 (2020)
8. 2cureX, Choose the Right Treatment for Each Patient. https://usercontent.one/wp/www.2curex.com/wp-content/uploads/2023/01/IndiTreat-brochure-1.pdf?media=1676030860. Accessed 12 July 2023
9. Chen, R., Chen, C.: Artificial Intelligence. An introduction for the Inquisitive Reader. Vol 1, CRC Press, Boca Raton and Oxon (2022)
10. The European Parliament and the Council of the European Union.: Regulation (EU) of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC *OJ* [2017] L 117/1
11. European Commission.: Factsheet for Manufacturers of in vitro diagnostic medical devices. European Union (2020)
12. The European Parliament and the Council of the European Union.: Directive 98/79/EC of the European parliament and of the Council of 27 October 1998 on *in vitro* diagnostics medical devices *OJ* [1998] L 331/1
13. Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., Zatloukal, K.: Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. New Biotechnol. **70**, 67–72 (2022)
14. Kristjánsdóttir, M.V.: Good Administration as Fundamental Right. Icelandic Review of Politics & Administration **9**(1), 237–255 (2013)
15. Teo, T.W., Choy, B.H.: in. In: Tan, O.S., Low, E.L., Tay, E.G., Yan, Y.K. (eds.) Singapore Math and Science Education Innovation. ETLPPSIP, vol. 1, pp. 43–59. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1357-9_3
16. See for example European Court of Justice.: *Teresa Cicala v Regione Siciliana* App no C-482/10 (ECJ) (2011)
17. Lock, T.: Article 41 CFR Right to good administration. In: Kellerbauer, M., Klamert, M., Tomkin, J. (eds.) vol. 1, pp. 2204–2207. Oxford University Press, Oxford (2019)
18. European Court of Justice.: *Fiskano AB v Commission of the European Communities* App no C-135/92, [40] (1994)
19. European Court of Justice.: *Transocean Marine Paint Association v Commission of the European Communities* App no 17/74, [15] (1974)

the competitor is seeking may – in fact – not be included in the clinical evidence, as the algorithm does not demonstrate a causal link. Moreover, the duty to state reasons imposed on the Notified Body may become cumbersome. The Notified Body may be confronted with the potentially impossible exercise of verifying the clinical evidence, which may result in the Notified Body making their decision on incorrect information. Thus, this may adversely affect a sound regulatory decision-making procedure.

The use of *intellectual property rights* may worsen the hurdles posed to both the right to be heard and the right to access one's file, upon which the competitors may rely, since they may form an additional obstacle to comprehend how the algorithm underlying the ADM system in precision medicine works. Since the algorithm comprises the company's competitive advantage over their competitors, the algorithm underling 'CUP-AI-Dx' and 'IndiTreat' may be protected by the regime of intellectual property rights – most likely the regulatory framework of trade secrets. While these trade secrets are submitted to the Notified Body – which will thus not negatively impact the duty to state reasons –, they are not shared with the general public – which includes the competitors – under the right to access one's file. As a result, the competitors may also face difficulties effectively exercising their right to be heard. Therefore, a sound regulatory decision-making process may be hampered.

5 Conclusion – A Research Agenda

This contribution has demonstrated that the general characteristics of ADM systems in precision medicine – i) the self-learning ability, and ii) the lack of a causal link, and iii) the use of intellectual property rights – may very well imperil the right to good administration. In particular, this piece has illustrated the dangers to the three subrights expressly mentioned in Article 41(2) EU Charter, namely i) the right to be heard, ii) the right to access one's file, and iii) the duty to state reasons – and thereby to sound regulatory decision-making procedures.

This piece does not provide solutions to overcome the three established hurdles, rather it suggests three research lines to diminish these perils. The *first* line of recommendation is conducting more research aimed at achieving explainable AI. Specifically local explainability appears a promising field for ADM systems in precision medicine, since this field studies how AI may explain how it has reached the specific outcome based on the input data. In the meantime, however, this paper suggests focusing on creating more accurate and fairer algorithms, which is the *second* line of suggested research. Such algorithms may be achieved by ensuring that complete, representative, and accurate data are inserted both during the creation of the ADM system in precision medicine and during the input phase when the ADM system is in operation. The *third* proposal is an approach as opposed to a line of research, as this contribution calls for conducting interdisciplinary research. Consequently, not only researchers in the field of data science should explore how to overcome the established challenges to the right to good administration, but also experts in the field of law and healthcare professionals should be involved due to the interdisciplinary nature of the topic at hand, which covers all these fields of study. More importantly, these three specialisations should collaborate to achieve explainable ADM systems in precision medicine that conform to the right to good administration.

4 Obstacles to the Right to Good Administration

The three general features of ADM systems in precision medicine – namely i) the self-learning ability, ii) the proof of correlation instead of causation, and iii) the use of intellectual property rights – is a breeding ground for perils to the subrights of the right to good administration under Article 41(2) EU Charter, namely i) the right to be heard, ii) the right to access one's file, and iii) the duty to state reasons.

While grasping how the algorithm works may already be cumbersome – since this exercise may require expert knowledge and skills – *the self-learning ability* only further hinders deciphering how the algorithm reached its outcome – or even creates an inconceivable activity. This burden to the comprehension of how the algorithm works is created due to the algorithm's ability to evolve independently based on the input of its environment. Concretely, this means that the outcomes of 'CUP-AI-Dx' predicting the location of the primary cancer and of 'IndiTreat' forecasting the suitability of medicinal products for colorectal cancer may be based on different rules than those initially programmed. Consequently, the competitors and the Notified Body may not be able to decipher how the algorithm reached its prediction. Since the competitors of the company drawing up the EU Declaration of Conformity – and thus submitting the clinical evidence for review by the Notified Body – may rely on the right to be heard, it is questionable to which extent the competitors can effectively make their points of view known before the adoption of the administrative decision due to this self-learning ability. Thus, this means that the right to be heard may be at risk. The same holds true as regards the right to access one's file. The Notified Body may provide unrestricted access to the clinical evidence to the competitors, but they will most likely not grasp what the information entails and its significance, which renders the right to access one's file meaningless. The self-learning ability also jeopardises the duty to state reasons, as the Notified Body – even though provided with the clinical evidence – may not fully understand how the ADM system has reached its outcome. Particularly, the question arises whether the Notified Body can adequately review the accuracy of the clinical evidence and verify the conclusions drawn by the manufacturer. Thus, the Notified Body may not be able to substantiate their decision as regards the review of the clinical evidence in a clear and an intelligible manner, which leaves the sound regulatory decision-making process at risk. Furthermore, the Notified Body responsible for the conformity assessment may face difficulties to pinpoint how the algorithm may evolve in the future and how this may affect the clinical evidence. In short, a sound regulatory decision-making procedure may be at risk.

The same holds true as regards the *lack of a causal link* between the input data, which consists of the patient's blood or tissue sample, and the output data, which forms the likelihood of a diagnosis or of the suitability of a pharmaceutical product. Zooming in on the ADM systems 'CUP-AI-Dx' and 'IndiTreat', this means that their outcomes may be based on correlation – as opposed to causation. As a result, the competitors and the Notified Body may not gather the right picture as to how the algorithm underlying 'CUP-AI-Dx' and 'IndiTreat' work, and whether the algorithm functions accurately. Thus, the *lack of causation* may hinder the right to be heard of the competitors. Since any alleged link may – in fact – evidence correlation, the question arises whether competitors can effectively make use of their right to be heard as they may not be able to get to the heart of the matter. This also remains true as regards the right to access one's file. The data

this right only materialises if the foreseen administrative decision may have adverse consequences for the individual. The right to be heard, thus, enables the administrative authority to consider the individual's point of view during the decision-making process. [17] This right consists of two components: first, public administration is obliged to notify the individual about the existence of the pending administrative decision, and second administrative authorities are to ensure that the individual is given the opportunity to effectively make their point of view known before adopting the administrative decision [20]. The above is also applicable to any other individual – not being the addressee of the administrative decision –, who is adversely affected by the adoption of the decision [6].

2. the right to access one's files (subparagraph b) [21] should be given to the individual both before and after the administrative authority adopts its decision. When providing access to their file before deciding on a case, the individual can give full effect to their right to be heard, as the individual can acquaint themselves with the information related to them held by public administration. Put differently, the right to access one's file is a vital precondition to effectively enjoy the right to be heard [22]. Given the status of the right to access one's file as an essential prerequisite of the right to be heard upon which individuals can rely who are negatively affected by the administrative decision – but are not the addressee –, the right to access one's file is, thus, also applicable to such individuals. When given access after the adoption of the administrative decision, the individual has the opportunity to understand the reasoning underlying the administrative decision and can thereby decide to seek – and if needed prepare for – judicial review [15]. Public administration are to provide the individual relying on their right to access their file all relevant information in their possession, except information covered by professional secrecy or business secrets [23, 24]
3. the duty to state reasons (subparagraph c)[7] requires public administration to state the reasons for their decision in a sufficiently precise manner that would allow the individual to understand the underlying reasons of the administrative decision. This would enable the individual to decide whether to appeal the decision in front of the court, which then can adjudicate based on the stated reasons [25]. The duty to state reasons serves a threefold purpose. First, the administrative decision-making procedure becomes more transparent, as it allows the individual to comprehend why the decision is taken and to decide whether to seek judicial redress (*individual perspective*). Second, administrative authorities are now to ponder upon which reasoning their decision is based, which counters arbitrary decision-making (*public administration perspective*). Third, the duty to state reasons is a prerequisite to perform effective judicial review (*judiciary perspective*) [15].

[7] The duty to state reasons is not based on general principles of EU law, but rather on existing Treaty provisions, which is elaborated in the case law of the Court of Justice of the European Union. See Craig, P.: Article 41. In Peers, S., Hervey, T., Kenner, J., Ward, A. (eds), vol. 1, pp. 1125–1152. Hart Publishing, Oxford (2021).

authorities.[4] Article 41 EU Charter is an umbrella concept[5] that contains a diverse set of rights and principles aimed at protecting the individual against the arbitrary use of power by administrative authorities. To this end, this myriad of rights and principles dictates how public administration ought to behave, especially in relation to individuals. Concentrating on Article 41's wording, the first paragraph encompasses an overall provision that entitles the individual '*[…] to have their affairs handled impartially, fairly and within a reasonable time […]*' [16]. This general right is further clarified in the second paragraph, which provides a non-exhaustive list of subrights. In particular, Article 41(2) EU Charter lists three rights that – undoubtedly – fall within the ambit of the right to good administration, namely i) the right to be heard, ii) the right to access one's file, and iii) the duty to state reasons. However, apart from these three rights that are explicitly mentioned, the umbrella right to good administration may encompass other rights and principles [15].

Before delving into the three procedural subrights under Article 41(2) EU Charter, the author holds that these three subrights are vital to the contextual principle of transparency in the light of the *In Vitro* Diagnostic Medical Devices Regulation, as both pursue sound regulatory decision-making procedures. Article 41(2) EU Charter mentions the following three subrights of the right to good administration, namely:

1. the right to be heard (subparagraph a)[6] is a context-specific right, which means that its content hinges on the circumstance under which it is invoked [15]. However, in general this right requires public administration to provide the individual an opportunity to make their stance effectively known before the adoption of the administrative decision that may adversely affect the individual concerned [18, 19]. This means that

[4] The author is aware that Notified Bodies are not necessarily part of public administration, which means that Article 41 EU Charter is not applicable. Nevertheless, the author argues that Notified Bodies may be regarded to fall under public administration based on a case handed down by the European Court of Justice, namely *A. Foster, G.A.H.M. Fulford-Brown, J. Morgan, M. Roby, E.M. Salloway and P. Sullivan and British Gas plc*, App no C-188/89. In this case, the European Court of Justice held that any body – irrespective of its legal form – may be on an equal footing with public administration, if that body is responsible for providing a public service under the control of the State per a measure adopted by the State. The author maintains that the same holds true as regards Notified Bodies. First, Notified Bodies are responsible for providing a public service, namely they are responsible for the conformity assessment, which is a prerequisite for the placement on the market of an ADM system in precision medicine. Second, Notified Bodies are both placed under the supervision of the State (see Articles 39 and 41 *In Vitro* Diagnostic Medical Devices Regulation) and appointed by the State (see Articles 35, 36, 38 *In Vitro* Diagnostic Medical Devices Regulation). Consequently, Notified Bodies are to comply with Article 41 EU Charter.

[5] The author is aware of the debate as regards the precise content of the right to good administration, as penned in Article 41 EU Charter. However, it is not this paper's aim to exhaustively discuss its elements. For an analysis of the content of the right to good administration, see for example Kanska, K.: Towards Administrative Human Rights in the EU. Impact of the Charter of Fundamental Rights. European Law Journal 10(3), 296–326 (2004).

[6] The right to be heard is also encapsulated in the case law of the European Court of Justice, see for example European Court of Justice.: *Transocean Marine Paint Association v Commission of the European Communities* App no 17/74, [15] (1974).

3 The Right to Good Administration

3.1 General Remarks

Before becoming a fully fledged human right within the context of the EU,[3] various elements of the – contemporary – right to good administration was already recognised by the Court of Justice of the European Union (**CJEU**) in its case law [5]. The right to good administration – in whatever form – is a pivotal human right that can hardly be overestimated. First, the right to good administration is an enabling right that facilitates individuals to effectively enjoy their fundamental rights, such as the right to an effective remedy. More concretely, the individual cannot be expected to enjoy their right to an effective remedy in case the administrative authority does not provide the underlying reasons for its decision in a clear and an intelligible manner. Put differently, the right to good administration is a vital precondition to exercise other fundamental rights [14]. Second, the right to good administration prescribes that the behaviour of public administration should be in accordance with written and unwritten law, which also includes their conduct towards individuals. Consequently, this right provides individuals with enforceable rights when interacting with public administration [14]. This demonstrates that the right to good administration is not merely a right that facilitates other human rights, but also – and perhaps more importantly – a 'stand-alone' human right [14].

3.2 EU Charter

Even though the EU Charter devotes an article to the right to good administration, much ink has been spilled about its precise status and scope. Based on its phrasing, only the Institutions, Bodies, Offices, and Agencies of the EU fall within the remit of the Charter right to good administration [15]. This reading has also been confirmed by the CJEU [16]. However, the CJEU has refined this black and white approach, and holds that general principles underlying the right to good administration are applicable to Member States when implementing EU law.

The right to good administration, as embedded in Article 41 EU Charter, is a procedural fundamental right [17] and plays a crucial role in procedures before administrative

[3] In some legal orders, good administration is still regarded a principle.

medical device, the Notified Body is to review the clinical evidence on its accuracy and verify the conclusions drawn by the manufacturer.[2]

Aiming our attention at the clinical evidence that is to be submitted by the manufacturer to the Notified Body, the clinical evidence aims to ensure that the *in vitro* diagnostic medical device is safe and produces the expected clinical benefits. Article 2(36) states that clinical evidence consists of [2].

- clinical data, and;
- results of the performance evaluation

Looking further into the results of the performance evaluation, the manufacturer is required – under Article 56(3) – to demonstrate: [2].

- scientific validity (Article 2(38)): this requires the *in vitro* diagnostic medical device to illustrate an association between the analyte and a clinical condition or physiological state [2]. Specifically in the context of an ADM system in precision medicine, this entails that the outcome produced by the underlying algorithm must indicate a link with a clinical condition or a physiological state [13];
- analytical performance (Article 2(40)): the *in vitro* diagnostic medical device is to show that the device can accurately discover and measure an analyte [2]. Looking at ADM systems in precision medicine, the output data should be accurate – as opposed to the detection and measurement of an analyte being accurate [13].
- clinical performance (Article 2(41)): here the *in vitro* diagnostic medical device must demonstrate a (medical) correlation between the results and the clinical condition or the physiological state [2]. Against the background of ADM systems in precision medicine, the output of the underlying algorithm ought to have a (medical) correlation with the clinical condition or the physiological state [13].

The above requirements linked to the EU Declaration of Conformity show that the principle of transparency, which aims to facilitate sound regulatory decision-making, is well-embedded in the *In Vitro* Diagnostic Medical Devices Regulation. During the conformity assessment procedure, the manufacturer is to submit a bulk of evidence demonstrating that the *in vitro* diagnostic medical device is safe to use and produce the anticipated clinical benefits, which requires the manufacturer to be transparent about how their *in vitro* diagnostic medical device works and their effects.

[2] For the review of clinical evidence during the conformity assessment of Class C *in vitro* diagnostic medical devices, which includes ADM systems in precision medicine that predict the likelihood the diagnosis fits the patient's profile, see Article 48(7), para. 1 and Annex IX, Sect. 4.4 *In Vitro* Diagnostic Medical Devices Regulation or Article 48(8), para.1 and Annex X, Sect. 3(c) *In Vitro* Diagnostic Medical Devices Regulation. For the review of clinical evidence during the conformity assessment of companion diagnostics, which encompasses ADM systems in precision medicine that predict the suitability of pharmaceuticals based on the patient's profile, see Article 48(7), para. 3 and Annex IX, Sect. 4.4 *In Vitro* Diagnostic Medical Devices Regulation or Article 48(8), para. 1 and Annex X, Sect. 3(c) *In Vitro* Diagnostic Medical Devices Regulation.

Table 1. Risks of the different classes of *in vitro* diagnostic medical devices

	Risk To Individual	**Risk To Public Health**
Class A	Low	Low
Class B	Moderate	Low
Class C	High	Moderate
Class D	High	High

in 2017. [2] Against this backdrop, it is, thus, no surprise that the *In Vitro* Diagnostic Medical Devices Regulation places the principle of transparency in the limelight. This principle is further echoed throughout the Recitals. To this end, the European legislator acknowledges in Recital 4 that transparency was not a concern in the previous legislative act, which resulted in the introduction of this type of obligations in the current *In Vitro* Diagnostic Medical Devices Regulation [2]. Unfortunately, determining the content of the principle of transparency is no clear-cut task, as it is a multifaceted concept, whose meaning is dependent on the context in which it is used. The question then arises what entails this principle of transparency in the *In Vitro* Diagnostic Medical Devices Regulation. In this legislative act, transparency is – simultaneously with adequate access to information – crucial for, amongst others, sound regulatory decision-making procedures, as mentioned in Recital 40 [2]. Consequently, this aim is set in the transparency obligations that are permeated in the main body of the *In Vitro* Diagnostic Medical Devices Regulation. However, these transparency obligations imposed on *in vitro* diagnostic medical devices vary and are determined by their classification, and thus by the risks posed to the individual using the *in vitro* diagnostic medical device and to the general public health. Depending on its classification, which in the case of ADM systems in precision medicine for the purpose of diagnosing patients and proposing a treatment plan including pharmaceuticals is 'Class C', additional requirements are applicable to safeguard transparency.

Transparency Obligations. Focussing on the requirements to place a 'Class C' *in vitro* diagnostic medical device on the market of the EU, Article 17 stands out seeing its all-encompassing scope, as it demands to draw up the EU Declaration of Conformity that certifies that all requirements of the *In Vitro* Diagnostic Medical Devices Regulation are observed [2]. As stipulated in Articles 10(5) and 15(3)(b), this comprehensive obligation is imposed on the manufacturers of all *in vitro* diagnostic medical devices [2]. Further, by signing the EU Declaration of Conformity, the manufacturer takes full responsibility that the ADM system in precision medicine complies with the *In Vitro* Diagnostic Medical Devices Regulation (Article 17(3)) [2]. As stated in Article 48, one of the obligations to obtain an EU Declaration of Conformity is to perform a conformity assessment, [2] which – quite literally – is an assessment to affirm that the requirements of the *In Vitro* Diagnostic Medical Devices Regulation have been fulfilled (Article 2(32) [2]. This evaluation can be done with or without the involvement of a third party, the Notified Body, which is designated by the Member State to perform the conformity assessment (Article 2(33) and (34)) [2]. Specifically in the case of a 'Class C' *in vitro* diagnostic

on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [10] (**Medical Devices Regulation**) or by the *In Vitro* Diagnostic Medical Devices Regulation. The Medical Devices Regulation governs medical devices in the broad sense – the main restricting factors consisting of their use having a medical purpose and the manufacturer's intended purpose, as mentioned in Article 2(1) [10]. The *In Vitro* Diagnostic Medical Devices Regulation does not merely require this medical purpose and the manufacturer's intended purpose but also – in accordance with Articles 2(1) and (2) – demands that these medical devices provide specific information in the medical context acquired by the examination of samples of the human body – which includes those originating from blood or human tissue – in a controlled environment, such as a test tube or petri dish [2]. As ADM systems in precision medicine mostly base a suggested diagnosis or a medical treatment plan on genomic data that is derived from bodily samples – be it from blood or from human tissue – and is examined in a controlled environment, these ADM systems are governed by the *In Vitro* Diagnostic Medical Devices Regulation – which will thus be the focus of this contribution. To this end, the author notes that ADM systems may solely be regulated by the Medical Devices Regulation provided they only use samples that do not come from the human body but rather from, for instance, medical imaging or data given by the patient.

Zooming in on the legal context, a distinction is warranted as regards the assessment of an ADM system in precision medicine that – on the one hand – suggests a diagnosis and – on the other hand – suggests a pharmaceutical regimen. While both these ADM systems are '*in vitro* diagnostic medical devices' under Article 2(1) [2], ADM systems that propose a diagnosis fall under Article 2(2)(a), and those that determine drug sensitivity are a specific *'in vitro* diagnostic medical device' under Article 2(2)(e), namely a 'companion diagnostic', as they i) identify which patients are likely to respond to a treatment plan, and ii) identify which patients are anticipated to suffer serious negative side-effects due to the regimen (see Article 2(7)(a) and (b)) [2].

However, to determine the applicable regime in the *In Vitro* Diagnostic Medical Devices Regulation, it does not suffice to define the ADM systems in precision medicine in these relatively general terms, they also need to be further classified in accordance with the classification rules (Article 47(1) and Annex VIII). [2] *In vitro* diagnostic medical devices are grouped in Class A to Class D, which is determined by the risks posed to the individual and to public health in general [11]. Table 1 illustrates these risks posed by *in vitro* diagnostic medical devices in order to be categorised as 'Class A', 'Class B', 'Class C' or 'Class D'.

In sum, the *In Vitro* Diagnostic Medical Devices Regulation imposes the least stringent obligations upon 'Class A' *in vitro* diagnostic medical devices, and establishes the most demanding requirements on those grouped in 'Class D'. Both ADM systems recommending a diagnosis and ADM systems proposing a treatment plan involving medicinal products are placed in Class 'C', see Rule 3(f) Annex VIII and Rule 3(h) Annex VIII, respectively [2].

The Principle of Transparency. In accordance with Recital 1, the principal purpose of the *In Vitro* Diagnostic Medical Devices Regulation is to ensure transparency, which was amongst the main aims behind the revision of the preceding legislative act [12]

2 Setting the Scene

2.1 The Factual Context - Automated Decision-Making Systems in Precision Medicine

ADM systems in precision medicine use data retrieved from samples stemming from, for instance, the patient's blood or tissue to predict the likelihood that the diagnosis or the treatment plan fits the patient's unique profile [6]. As a result, ADM systems may be a useful tool for physicians in diagnosing their patients or determining a suitable treatment plan for their patients.

An example of an ADM system for the purpose of medical diagnoses in the field of precision medicine is 'CUP-AI-Dx', which identifies – by using RNA – the location of the primary cancer of patients diagnosed with the rare disease 'carcinoma of unknown primary' [7]. This type of cancer is difficult to treat since the primary cancer is unknown. As such, the ADM system may facilitate physicians to diagnose their patients by localising the primary cancer, and thereby help them to treat their patients. Another example in the field of precision medicine but focussing on treatment plans entails the identification of a suitable drug treatment for colorectal cancer by the ADM system 'IndiTreat'. This test helps decide which medicinal product is likely the most suitable for patients suffering from colorectal cancer by examining the patient's profile against a particular set of pharmaceutical products. Thus, this ADM system may help physicians to set up an effective treatment plan [8].

These ADM systems in precision medicine are characterised by three generic features, namely i) the self-learning ability, ii) the lack of a causal link, and iii) the allocation of intellectual property rights. First, *the self-learning ability* is rooted in the use of machine learning and deep learning techniques, which provides ADM systems with the ability to independently learn from its environment. Put differently, the algorithm underlying the ADM system has acquired the ability to self-evolve – that is to say without human intervention. As a result, grasping how the ADM systems reach its outcome based on the input data may be a cumbersome – if not an impossible – task. Second, these ADM systems may *lack a causal link*, since they may establish correlation between the input data and the acquired outcome. Thus, any alleged link between the input and the output may be solely a coincidence or caused by noise [9]. Third, these ADM systems largely enjoy protection rooted in *intellectual property rights*. Seeing the involvement of machine learning and deep learning techniques in the creation of these ADM systems – which requires specialised skills and competences and substantial resources –, ADM systems are mainly developed by the private sector [3]. Since the algorithm resulting from machine learning and deep learning may be the enterprise's competitive advantage, companies may opt to protect this advantage by using intellectual property rights, and more specifically the legislative framework governing trade secrets. Consequently, the protection mechanism provided by intellectual property rights may constitute an additional hurdle – and thus exacerbate – unravelling how the ADM system works.

2.2 The Legal Context - *In Vitro* Diagnostic Medical Device Regulation

General. ADM systems used to diagnose disease or to set up medical regimes are governed by Regulation (EU) of the European Parliament and of the Council of 5 April 2017

Union (**EU**), these ADM systems are governed by Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on *in vitro* diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU [2] (***In Vitro* Diagnostic Medical Devices Regulation**), as a bodily sample – for instance samples from blood or human tissue – serves as input data.[1] Concisely, as mentioned in its Article 1(1), the *In Vitro* Diagnostic Medical Devices Regulation governs the procedure of the ADM system from the moment they are placed on the market of the EU [2]. More concretely, as the development of these ADM systems requires expertise and ample of resources – as demonstrated by the use of machine learning and deep learning techniques –, the private sector predominantly creates them [3] and are thus to comply with the legal obligations stemming from the *In Vitro Diagnostic* Medical Devices Regulation. When interacting with the private sector, the public administration is to comply with the right to good administration, as embedded in Article 41 of the Charter of Fundamental Rights of the European Union [4] (**EU Charter**). This right to good administration includes a diverse subset of rights and principles, which all aim to safeguard the individual's right to defence during administrative proceedings [5]. However, this umbrella right to good administration may be at risk due to the generic features of the algorithm underlying ADM systems, which includes i) the self-learning ability, ii) the evidence of correlation – as opposed to causation –, and iii) the allocation of intellectual property rights.

The aim of this paper is to explore how these three general aspects of the algorithm underlying the ADM systems used in precision medicine affect the right to good administration. To this end, this contribution paints the background, which comprises the context in which these ADM systems are used (Sect. 2.1) and the legal framework that consists of the *In Vitro Diagnostic* Medical Devices Regulation with a specific focus on the principle of transparency it pursues and its transparency obligations (Sect. 2.2). After, this piece scrutinises the legal framework that comprises the right to good administration, which simultaneously consists of the legal benchmark against which the effects of the three general features of ADM systems in precision medicine are evaluated. After introducing the right to good administration (Sect. 3.1), this paper focusses on Article 41 EU Charter in which the right to good administration is embedded. In particular, this contribution dissects the three subrights that are expressly mentioned in Article 41(2) EU Charter (Sect. 3.2). Subsequently, this piece specifies the obstacles posed, caused by the characteristics of ADM systems to the right to good administration as outlined in Article 41(2) EU Charter (Sect. 4). Lastly, this paper concludes and proposes a research agenda in which the author suggests research recommendations to overcome the hurdles to the right to good administration (Sect. 5).

[1] The author points out that ADM systems that base their prediction on other data than those retrieved from human blood or tissue – for instance medical records or medical imagery – are not considered by the *In Vitro Diagnostic* Medical Devices Regulation but rather fall within the scope of the more generic Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. For more information, see Sect. 2.2.

Automated Decision-Making Systems in Precision Medicine – The Right to Good Administration at Risk

Sarah de Heer(✉)

Department of Law, Lilla Gråbrödersgatan 4, 222 22 Lund, Sweden
sarah.de_heer@jur.lu.se

Abstract. Automated decision-making (**ADM**) systems – whose algorithms are based on Artificial Intelligence and more specifically on machine learning and deep learning – predict the likelihood of an outcome based on profiling the input data. ADM systems, which are predominantly developed by the private sector, are a promising device for the field of precision medicine, where medical intervention is based on the patient's unique profile that consists of their genomic data, medical records data, environmental data, and lifestyle data. Such ADM systems are used when diagnosing or creating a treatment plan for patients. As these ADM systems take a bodily sample, for example from blood or human tissue, to predict which diagnosis or drug regime is most suitable, they are governed by the *In Vitro* Diagnostic Medical Devices Regulation in the European Union. However, the general features inherent to coding algorithms based on machine learning and deep learning – amongst others i) the self-learning ability, ii) the lack of a causal link, and iii) the use of intellectual property rights –, may form perils to the right to good administration that prescribes the legal norms of administrative conduct, including towards individuals. Particularly, the right to be heard, the right to access one's file, and the duty to state reasons may face considerable hurdles. Thus, this contribution aims to scrutinise these risks to the right to good administration and proposes a research agenda to overcome them.

Keywords: Automated Decision-Making System · Precision Medicine · Right to Good Administration

1 Introduction

Automated decision-making (**ADM**) systems are tools based on Artificial Intelligence (**AI**) that predict the likelihood of an outcome by means of profiling. Their developers particularly rely on machine learning and deep learning – two subsets of the more general AI. The use of ADM systems is especially alluring in precision medicine, which is a subfield of medicine that adapts medical interventions to the patient's profile. In this context, the algorithm underlying the ADM system considers, amongst others, genomic data, medical records data, environmental data and/or lifestyle data [1]. In the European

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 92–104, 2024.
https://doi.org/10.1007/978-3-031-59717-6_6

Iwaya, L.H., Babar, M.A., Rashid, A.: Privacy Engineering in the Wild: Understanding the Practitioners' Mindset, Organisational Culture, and Current Practices (2022). arXiv preprint arXiv: 2211.08916

Joo, E., Kononova, A., Kanthawala, S., Peng, W., Cotton, S: Smartphone Users' Persuasion Knowledge in the Context of Consumer mHealth Apps: Qualitative Study. JMIR Mhealth Uhealth **9**(4), e16518 (2021). https://mhealth.jmir.org/2021/4/e16518, https://doi.org/10.2196/16518

Kolovson, S., et al.: Understanding participant needs for engagement and attitudes towards passive sensing in remote digital health studies. In: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare, Association for Computing Machinery, pp. 347–362 (2020)

Kordzadeh, N., Warren, J.: Communicating personal health information in virtual health communities: an integration of privacy calculus model and affective commitment. J. Assoc. Inf. Syst. **18**, 45–81 (2017)

Nurgalieva, L., O'Callaghan, D., Doherty, G.: Security and privacy of mHealth applications: a scoping review. IEEE Access **8**, 104247–104268 (2020). https://doi.org/10.1109/ACCESS.2020.2999934

Lau, J., et al.: Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. In: Proc. ACM Hum.-Comput. Interact. 2(CSCW): Article 102 (2018)

Leon, P., et al.: Privacy and behavioral advertising: towards meeting users' preferences. In: PPS '15: Second SOUPS Workshop on Privacy Personas (2015)

Janic, M., Wijbenga, J.P., Veugen, T.: Transparency enhancing tools (TETs): an overview. In: 2013 Third Workshop on Socio-Technical Aspects in Security and Trust, pp. 18–25. IEEE, June 2013

Cunha, J.A.O.G.d., Aguiar, Y.P.C.: Reflections on the role of nudges in human-computer interaction for behavior change: software designers as choice architects. In: Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems. Diamantina, Brazil, Association for Computing Machinery: Article 56 (2020)

Danaher, B.G., et al.: From black box to toolbox: outlining device functionality, engagement activities, and the pervasive information architecture of mHealth interventions. Internet Interv. **2**(1), 91–101 (2015)

Deci, E.L., Ryan, R.M.: Self-determination theory. In: Van Lange, P.A.M., Kruglanski, A.W., Higgins, E.T. (eds.) Handbook of Theories of Social Psychology, pp. 416–436. Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n21

Degeling, M., et al.: We value your privacy … now take some cookies: measuring the GDPR's impact on web privacy. Informatik Spektrum **42**(5), 345–346 (2018)

Detweiler, C.A., Hindriks, K.V.: A survey of values, technologies and contexts in pervasive healthcare. Pervasive Mob. Comput. **27**, 1–13 (2016)

Peters, D., Calvo, R.A., Ryan, R.M.: Designing for motivation, engagement and wellbeing in digital experience. Front. Psychol. **9** (2018). https://doi.org/10.3389/fpsyg.2018.00797

Schomakers, E., Lidynia, C., Ziefle, M.: Listen to my heart? how privacy concerns shape users' acceptance of e-health technologies. In: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 306–311 (2019). https://doi.org/10.1109/WiMOB.2019.8923448

Ferreira, A., et al.: Perceptions of Security and Privacy in mHealth. In: HCI for Cybersecurity, Privacy and Trust, Cham, Springer International Publishing (2021)

Fishbein, M.: A theory of reasoned action: Some applications and implications. Nebr. Symp. Motiv. **27**, 65–116 (1979)

Floridi, L., Cowls, J., Beltrametti, M., et al.: AI4People—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5

Fogg, B.J.: A behavior model for persuasive design. In: Proceedings of the 4th international Conference on Persuasive Technology, pp. 1–7, April 2009

Guo, X., et al.: The privacy–personalization paradox in mHealth services acceptance of different age groups. Electron. Commer. Res. Appl. **16**, 55–65 (2016)

Gupta, B., Chennamaneni, A.: Understanding online privacy protection behavior of the older adults: an empirical investigation. J. Inf. Technol. Manag. **29**, 1–13 (2018)

Hutchinson, H., et al.: Technology probes: inspiring design for and with families. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Ft. Lauderdale, Florida, USA, Association for Computing Machinery, pp. 17–24 (2003)

Poyner, I.K., Sherratt,R.S. : Privacy and security of consumer IoT devices for the pervasive monitoring of vulnerable people. Living in the Internet of Things: Cybersecurity of the IoT - 2018, 2018, pp. 1–5 (2018). Doi: https://doi.org/10.1049/cp.2018.0043

Icek, A.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**(2), 179–211 (1991)

Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary. Washington (DC): National Academies Press (US); 2010. 5, Healthcare Data as a Public Good: Privacy and Security. https://www.ncbi.nlm.nih.gov/books/NBK54293/

Institute of Medicine (US); Grossmann C, Powers B, McGinnis JM, editors. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington (DC): National Academies Press (US); 2011. 8, Fostering the Global Dimension of the Health Data Trust. https://www.ncbi.nlm.nih.gov/books/NBK83578/

Table 5. (*continued*)

Emerging Themes	Privacy Challenges (Codes)
	Unclear time-value tradeoff
	Enable automatic prompts for reduced manual/mental comparisons
	App launch and preset schedule for data change notification
	Interaction and illustration for privacy change engagement
	Social risk-reward engagement
	Uncertainty in AI

References

Acquisti, A., et al.: Nudges for privacy and security: understanding and assisting users' choices online. ACM Comput. Surv. 50(3): Article 44 (2017)

Al Ameen, M., Liu, J., Kwak, K.: Security and privacy issues in wireless sensor networks for healthcare applications. J. Med. Syst. **36**, 93–101 (2012). https://doi.org/10.1007/s10916-010-9449-4

Andrews, V.: Analyzing awareness on data privacy. In: Proceedings of the 2019 ACM Southeast Conference, pp. 198–201. Association for Computing Machinery, Kennesaw (2019)

Arora, S., Yttri, J., Nilse, W.: Privacy and Security in Mobile Health (mHealth) Research. Alcohol Res. Current Rev. **36**(1), 143–151 (2014)

Atienza, A.A., et al.: Consumer attitudes and perceptions on mHealth privacy and security: findings from a mixed-methods study. J. Health Commun. **20**(6), 673–679 (2015). https://doi.org/10.1080/10810730.2015.1018560

Bartoletti, I.: AI in Healthcare: Ethical and Privacy Challenges. Springer International Publishing, Cham (2019)

Bertino, E., et al.: Internet of Things (IoT): Smart and Secure Service Delivery. ACM Trans. Internet Technol. **16**(4): Article 22 (2016)

Caine, K.: Local standards for sample size at CHI. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 981–992, May 2016

Calvo, R.A., Peters, D., Vold, K., Ryan, R.M.: Supporting human autonomy in AI systems: a framework for ethical enquiry. In: Burr, C., Floridi, L. (eds) Ethics of Digital Well-Being. Philosophical Studies Series, vol. 140. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50585-1_2

Cameron, J.D., Ramaprasad, A., Syn, T.: An ontology of and roadmap for mHealth research. Int. J. Med. Informatics **100**, 16–25 (2017). https://doi.org/10.1016/j.ijmedinf.2017.01.007

Chen, Y., et al.: Privacy games. ACM Trans. Econ. Comput. **8**(2), Article 9 (2020)

Christman, J.: Autonomy in Moral and Political Philosophy. The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Chun Tie, Y., Birks, M., Francis, K.: Grounded theory research: a design framework for novice researchers. SAGE Open Med. **7**, 2050312118822927 (2019). https://doi.org/10.1177/2050312118822927

Table 5. (*continued*)

Emerging Themes	Privacy Challenges (Codes)
	Awareness about where data exists in the wild and corrective steps to reduce its footprint
External influencers	Avoiding situations where privacy beliefs are challenged
	Knowing data footprint in connected social networks
	Balance user privacy advocacy and improving services
	Community for those with stigmatized health conditions
Unique Motivators	Mapping app's utility to types of required privacy interactions
	UX/UI not designed for diverse intended users' needs
	Robust and quality-driven app vetting to produce trust
	Privacy information designed for simple, personalized risks/controls
	Data types with high motivation
Supporting contextual autonomy through accessibility	Visual limitations persist and inhibit ability pursuant modality
	Ability dependent on environment
	Discern problem solving (self-diagnosis and resolution) VS. Seeking professional consultation
	Attention limitation
	Efficient, simple, and gratifying enable ability
Triggering autonomy through automation	Early declaration of an app intended use and relationship to your data
	Undesirable early interactions impacting attention and experience
	High frequency notices inducing questions and fatigue
	Overcoming negative historical connotations

(*continued*)

Table 4. Methodological process

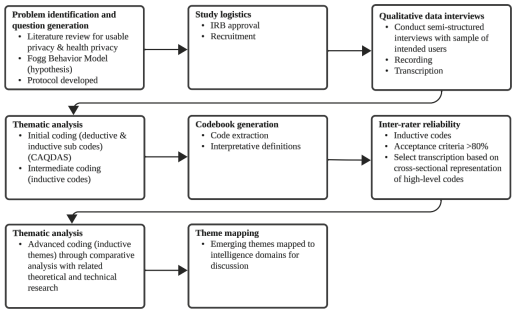


Table 5. Initial & Intermediate Thematic Codes

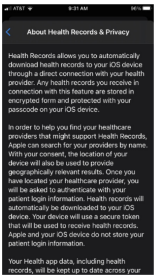
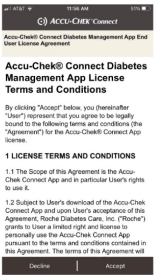
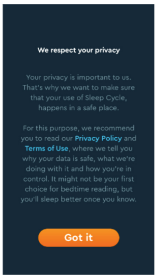
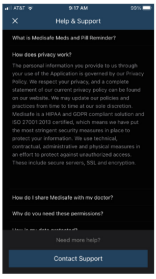
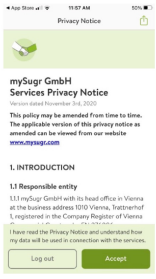
Emerging Themes	Privacy Challenges (Codes)
Disconnected cognition for assessing value	Modern data privacy interpretations
	Knowledge, emotional, and identity benefits
	Disconnected knowledge to propose value in pervasive environments
Challenged agency - limited freedom of choice	Incongruent consent
	Insufficient options
	Information retrieval
	Inflexible and disabled solutions
	Multivariate burdens and utilities
	Unclear privacy requirements for players
	Privacy requirements for preventing harm to organization and user
Customizable, trustworthy, and engaging solutions to build sensational experiences	UX/UI for improving privacy engagement
	Privacy volume and comprehension affecting app authenticity and trust
	Custom feature development that enable privacy interactions
Privacy awareness anticipations for the future	Balance between social and personal privacy paradigm
	Privacy competence to ensure end user interests are core

(*continued*)

Table 2. Preliminary Themes

Question	Answer	Theme
Meaning of "data"	Collected	Informational inputs and outputs
	Interpreted	
	Analyzed	
	Processed	
	Decision derivation	
Meaning of "privacy"	Confidentiality	Limiting exposures
	Access	
	Hopelessly implausible	Socio-technical influences
	Personal protection	Personal interventions
	Control	

Table 3. Study interview mobile health privacy policy probes (stimuli)

Apple (General)	Accu-Chek Connect (Diabetes)	Sleep Cycle (Sleep)	Medisafe (Medications)	mySugr (Diabetes)
				

6 Conclusion

In this paper, we used existing behavior models as a lens to understand users' privacy experiences, behaviors, and perspectives toward mHealth data privacy policies. From 15 semi-structured interviews with adult users of mHealth applications, we extend knowledge of users' experiences and unmet needs for privacy policy design that influence users' behaviors toward mHealth applications. Through the lens of the Theory of Planned Behavior and SDT, we characterize factors beyond personal data control and consent that influence users' sense of autonomy when engaging with mHealth privacy policies. Finally, we provide unique considerations for privacy policy design that focus on improving consent preferences, transparency of privacy control statuses, and building trust on a multiple levels.

Acknowledgments. Special thanks to the participants that shared their experiences and Davide Bolchini, Ph.D., for assisting with editing.

Appendix

Table 1. Participant Demographics

P#	Gender	Age	Highest education level	OS	mHealth app usage	Privacy notice frequency
1	Male	43	Graduate degree	iOS	Daily	Weekly
2	Female	31	Graduate degree	AOS	Daily	Daily
3	Male	47	Graduate degree	iOS	Monthly	Monthly
4	Male	31	Graduate degree	iOS	Daily	Weekly
5	Female	39	Undergraduate degree	iOS	Daily	Annually
6	Male	49	Graduate degree	iOS	Monthly	Monthly
7	Female	39	Some college	iOS	Daily	Weekly
8	Female	67	Undergraduate degree	iOS	Daily	Monthly
9	Female	56	Doctorate degree	iOS	Daily	Daily
10	Male	28	Undergraduate degree	iOS	Weekly	Monthly
11	Female	57	Graduate degree	iOS	Daily	Weekly
12	Male	37	Some college	iOS	Weekly	Weekly
13	Female	32	Undergraduate degree	AOS	Daily	Weekly
14	Male	28	Undergraduate degree	iOS	Weekly	Monthly
15	Male	59	Some college	iOS	Daily	Weekly

in not overwhelming the intended users with options that seemingly appear outside of the perceived guardrails. This means that while some configurable parameters may be needed, having too many options can lead to confusion or make users feel like they are completely on their own in their privacy decisions. This nuanced position is contrast to the notion of having complete awareness and control as premised in previous research (Andrews, (2019), Schaub et al., (2017)), but it more importantly must match users expectations about broadly applied privacy preference modeling across mHealth apps. Moreover, consent modeling is inherently dynamic over time and between people who share information in mHealth app environments. For example, collaborative privacy sharing models (Petronio, 2010) take into account multiple parties, but how these collaborative agreements change with participant preferences over time present uncertainty, thus designing for clear user-roles and control are critically important. Lastly, fostering trust and transparency requires a transparency about black box environments where data are collected, processed, and stored. Such transparency-enhancing tools are acknowledged as helping promote privacy and trust (Janic, et al., (2013)), but an importantThis mechanism of for this visibility must ensure traceability of data usages and clear verifiable levels of control by the user over data in those specific environments.

5.3 Limitations

One limitation of our work is that our study is retrospective of behaviors and does not actually observe behaviors with privacy policies. While the self-reported accounts provided by participants provide insights into their experiences with mHealth privacy policies, additional studies of direct user behavior may uncover additional challenges and design implications. Our work is also qualitative which is useful for providing an in-depth understanding of users' attitudes and perceptions. However, one tradeoff of qualitative work is that findings are not generalizable (Leung, 2015). We provide a rich, thick description to aid transferability, but our work like other qualitative work (Joo et al., (2021), Martin-Hammond et al., (2019), Zhang et al., (2021)) is likely limited based on the context in which it was carried out. Further, as we collected data we began to see recurring ideas, and continued until we stopped seeing new data, which. This is consistent with the processes for analyzing qualitative data., however our While our sample is small and may be limited by certain participant demographics, such as some participants that were familiar with familiarity with privacy policies, we believe this to still be a valuable step in the direction to explore this research further and look to consider age-based and other demographic perspectives in future work and therefore some users' privacy concerns may not be represented in our findings. Our $N = 15$ is slightly higher than other qualitative interviews (Caine, 2016) conducted by HCI researchers, however, data reached saturation at 12 participants where we noticed consistent responses with fewer new points emerging. We completed an initial review of all transcripts excluding those without response variation to determine those to include in agreement calculations. We used (McDonald et al., 2019) for determining our agreement approach. Our analysis approach was consistent with their arguments against solely using IRR for agreement. These are considerations for future research.

5.2 Balancing Perceptions of mHealth Privacy Autonomy and Automation

Privacy autonomy, through the lens of independent engagement and trust, is inherently challenged because of limited support and choices (Cunha et al., (2020), Schaub et al., (2017)). However, our examination of users' self-reported behaviors indicates that understanding mHealth privacy language is also a barrier to fully engaging with a technology. We extend the need to decouple the domains of being informed and intentional consent in order to focus research towards understanding language over simply improving consent interactions. Making information comprehensible for users reach educational and governance systems because both are needed to scale health and consumer applications, especially when data collection is essential in the user's health journey. In some mHealth technologies, such as the wearable Apple Watch, users are able to select the data that may be collected about them (i.e., biometric identifiers) conditional to the practices that are employed in a given app's functionality. Using gamification techniques may be another considerable way of building user knowledge about the personal data collected about them, which is supported in the research by (Simon et al., 2021) that describes cognitive absorption for privacy decision-making because of engaging gamification. Other research (Mavroeidi et al., 2020) also considers using gamification for engaging users about privacy. However, we also suggest that in the future, it could be useful to investigate the role of gamification in building value constructs aimed to incentivize (or motivate) learning and understanding complex mHealth data privacy.

Collective privacy influence is a unique area that emerged from our interviews because it highlights both the benefit and responsibility of understanding social contexts when developing policies about mHealth privacy. Although this finding is similar to users engaging in health-based communities for health support (Kordzadeh et al., (2017), Danaher et al., (2015)), it is unique in the sense of mHealth privacy because it exposes a gap where individuals and their social influences are not currently aligned, which affects these users' perceived privacy control, thus autonomy and trust. This is similarly discussed in research by (Gupta, 2018) that identifies external influences (e.g., such as prior experience) on older adults' general privacy behaviors, but the emerging theme from our interviews recognizes the social norms, highlighted in behavioral theories such as TPB (Icek, 1991), play a unique role in dynamic mHealth privacy behaviors. Various socially oriented topics arose from our data ranging from meta views on balancing social and personal privacy initiatives, situation avoidance for privacy discourse, data privacy footprint in social networks, balancing privacy advocacy and improving services, and providing community for individuals with stigmatized health conditions. While these topics range in variety and abstraction, they construct a basic model for the relationship between social influences and perceived individual privacy attitudes, thus extending work (Zou et al., (2020), Guo et al., (2016)) by detailing unique behavioral intentions used as a vehicle for trust in mHealth systems. It is for this reason that we suggest that future usable mHealth privacy research must continue to investigate these topics and explore opportunities to leverage and enhance these outside constructs for the development of truly autonomy-supporting privacy interactions.

Furthermore, potential design directions that strike a balance between autonomy and automation may need to focus primarily on consent, transparency, and trust. Specifically, designing preferences at the time when the user provides consent must be careful

data, consent, and nudging interactions (Cunha et al., (2020), Degeling et al., (2018), Schaub et al., (2017), Utz et al. (2019)). Yet, due to some of the historically untrustworthy actions that occurred that sit at the intersection of health and privacy (Grossmann et al., 2011), some users still are wary, influencing their perceptions of mHealth technologies. To address questionable trust and adoption in mHealth systems that collect and process health data, researchers and industry practitioners have seemingly held the position that providing more ways for users to access and manage personal data is a sufficient baseline for control (Schaub et al., (2017), (Atienza et al., 2015)). However, we uncovered that users' perceptions and expectations of privacy control often do not equate to the independence that is needed for autonomy. Therefore, based on our data we conclude that there is a conflated belief that control is the same as autonomy. In the design of health and well-being technologies, often autonomy extends beyond the binary concepts of control and is defined as a users' feeling of agency or their ability to act based on their goals and values (Peters et al., 2018). In the context of mHealth privacy, while it is reasonable to assume that personal responsibility is essential for consenting and using mHealth technologies, non-privacy-neutral perceptions inherently exist when users are tasked with deciding to use a service or not (binary opt-in vs. opt-out). This is further compounded by the fact that notice-choice structures present content that are likely not to be read in the first place (Meier et al., 2020). When agreements are in place with conditions that are non-negotiable to the user's existing motives or beliefs, this creates questions of perceived control over one's privacy and whether the application actually supports users' autonomy, and is deserving of trust.

Our findings highlight users' beliefs that there are not enough alternative ways of getting people to engage with their mHealth data privacy practices, specifically informed consent interactions. Further, offloading all the privacy decisions at the launch of a new app is not only contradicting the benefit of the mHealth app, but it also ignores the dynamic ways that people choose to be informed and interact with their sensitive information. In the future, it would be beneficial for usable privacy researchers and industry professionals to explore alternative strategies that focus on personalized and emotional engagements with mHealth data privacy in order to support autonomy, while also distinguishing this work from traditional views about privacy control that often focuses on actions. In similar discussions, (Christman, 2020) distinguishes basic and ideal autonomy where basic autonomy implies that users are free from influence and imply they are not under constricting conditions. We also posit that patients with health conditions who seek support from mHealth technologies are inherently constricted in autonomy and thus are forced to weigh utility-tradeoffs unfairly. While this context of autonomy relates to other work regarding "contextual integrity" (Wijesekera et al., (2015), Zimmer, (2018)), we find that meeting user expectations is not simply about control (e.g., permissions, etc.), but also recognizing the role of changing awareness and mental processing, particularly on the side of social and historical influences. Our research extends prior work that examines contextual factors related to general privacy policy design (Micinski et al., (2017), Votipka et al., (2018), Squicciarini et al., (2014)), but we extrapolate factors unique to supporting autonomy with mHealth's privacy interactions.

In summary, we learned that subjective norms around interoperable environments (e.g., wearables and remote patient monitoring) in the health context, social influences, and changing motivations each influence participants' privacy interactions, which ultimately affects their perception of autonomy and trust in those interactions. By uncovering these subjective norms, we identified unique relationships with trust that may not have been clearly articulated previously in the mHealth design space.

5 Discussion

Through the lens of TPB, our research finds that mHealth users are unengaged with privacy policies and feel there is a chasm between their individual needs and the controls provided by the privacy community. Overall, our research suggests that mHealth users generally agree that privacy policies are beneficial and crucial for mHealth applications; however, they encounter persistent challenges when engaging with those policies. Specifically, we found that not sufficiently characterizing user' perceptions of internal and external motivations may obfuscate real opportunities for making privacy language more engaging and bridging the gap to help users understand essential information. As such, one result may be that end-users do not understand how the design of these privacy solutions are intended to protect them. Thus, having superficial awareness without knowledgeable engagement does not actually support autonomy and trust in mHealth applications. Our research builds on existing literature [Cunha et al., (2020), Leon et al., (2015), Vilaza et al., (2019), Gupta, (2018)) by advocating for the development of privacy solutions to be behaviorally and contextually orientated in order to uncover real user facing problems when interacting with mHealth privacy policies. Improving users' ability to understand language and recognize dynamic mHealth privacy environments relies on systematically assessing motivational intentions that engage users beyond basic privacy awareness. Further, we found that perceived control over one's mHealth privacy is unrealized partly in fact due to the inability to tangibly see, interact, and understand what privacy means when engaging with a mHealth technology. Thus, many users feel they lack autonomy when engaging with mHealth privacy policies, but due to the criticality of the context - managing health, users feel compelled to comply or completely disengage despite their concerns. Our results show that designing for motivationally charged engagement and understanding by leveraging social factors may be one effective way to optimize autonomy-support and trust in mHealth solutions. We discuss these implications in the following sections.

5.1 More Control Does Not Equal More Autonomy

Our work considers perceived autonomy through the lens of Self Determination Theory (Deci et al., 2012), where autonomy is having the choice and will to act according to personal goals and values. For health-related technology design, (Calvo et al., 2020) distinguishes autonomy from independence and control, noting that perceived autonomy can also be influenced by individual behaviors, lifestyle or society, which in-turn impacts adoption. Significant work has been done to simplify the experience that users have with privacy policies (Acquisti et al., 2017) and provide them with more control over their

fear of societal stigma, or alternative paths to build confidence in their decisions through social networks affecting their autonomy when engaging with mHealth applications.

4.3 Contextual Nascence: Navigating Black Box Interoperability and Historical Preconceptions

We learned that participants' interactions with privacy notices vary significantly but may originate from unique historical preconceptions, such as black box interoperability and the evolution of health and technology. For example, interoperable mHealth environments collect and process various forms of sensor and self-reported personal health data. This data is often embedded in artificial intelligence (AI) or other personalized systems whose architecture enables health management solutions similar to those described in other work (Danaher et al., 2015). While these architectures are innovative, participants expressed privacy concerns about technologies such as proactive AI systems that continually collect their data and push untraceable targeted-marketing material. This raises broader questions about the role of emerging technologies such as IoT or voice technologies in shaping users perceptions of mHealth technologies and users interactions with health applications provided by those devices. For example, P09 stated, "…Siri and the Amazon Echo are listening all the time [and] can get information and they're going to hear private health information. If, you know, somebody's listening or they're going back and reviewing vital recordings, as they're supposedly trying to make Siri better and more interactive with better programming. There are people who hear that private information. So because it's recorded near your house, private information could also be out there if that's what happens, what was recorded at that point in time." Participants were generally uncertain about the AI black box (Lau et al., 2018), but were tangibly concerned about the inability to trace data effectively across its lifecycle, and especially when it is shared or sold for other purposes. For example, P10 stated, "…this kind of goes along the lines of sharing or understanding how my data is being shared with other companies or the service provider I am doing business with…If I start to get targeted or oddly specific targeted ads that seem to be coming from my interactions with one system in particular, that could prompt me to take a look and maybe try to get a better understanding of just how much data is being collected and how it's being used. And just kind of trying to connect the dots if I get very targeted marketing on different devices and I can try to trace it back to a certain application…".

Participants also perceived healthcare's historical evolution as a motivational factor towards privacy. Health technologies such as mHealth are burdened with negative historical connotations for various reasons such as public cases of individuals' health data rights being violated. Participants explained that the rapid prevalence of notices for various technologies is one reason why some pre-mHealth generations have negative views about mHealth technologies. For example, P07 stated, "I think at this point, I'm young enough to expect them [privacy policies] to be there and old enough to remember when they weren't." Another negative historical connotation was explained by P02, who stated, "the older cases of like Henrietta Lacks, they used her [information] and she never knew." We found that these historical references and events influenced how participants perceived privacy in certain social groups.

participants sometimes see imbalances between technological advances and personal privacy on a macro socio-technical level. Some participants feared that society values the speed and convenience of mHealth technologies more than understanding their privacy implications. For example, P01 stated, "[the] balance of convenience and technology is one that's a very difficult one for me that I kind of struggle with just because I know how much you are giving up in the sense of privacy… I like to be more informed where that balance is with every individual device or piece of software, or whatever it is that I am interacting with… I like to be informed on how much of my life I am giving away or my information or my private data, or we will see how much of my soul I am selling to save eight minutes or to gain some form of convenience… I don't think that there is enough concern in the general populace for the level of information that is being collected about every individual…" Some participants therefore held a belief true privacy is hopelessly implausible due to the advancing tech market, which is consistent with attitudes of fatalism (Joo et al., (2021)). For example, P01 stated, "Unfortunately, privacy is a pipe dream that most people have given up." suggesting that they feel most users do not have autonomy in privacy decisions whether they like it or not. However, there were other users that were hopeful that future mHealth privacy research will consider these social concerns and influences and their implications.

Participants explained their privacy decisions are sometimes negatively affected by the type of health condition their mHealth device supports. Participants mentioned that mental health and addiction heighten their privacy decisions because these conditions have potential social stigmas and insurance implications. One participant was concerned about billing insurance for mental health issues. For example, P05 stated, "…when I worked in a health setting, you see a lot of patients coming in for mental health reasons or addiction reasons, and they didn't want their insurance billed, or they didn't want it to go through certain channels because they wanted to keep it highly private." Another participant was concerned about their employer receiving sensitive information about their addiction. For example, P08 stated, "…so say I had a drug problem or something, and there was an app for what I'm trying to handle that or something, I wouldn't want the fact that I was a drug addict being shared with an employer or anyone really. So, those kinds of things that could be looked on negatively…" As stated by these participants, social stigmas around sensitive health conditions have a role to play in their privacy decisions and also could impact their sense of autonomy leading to the decision not to engage with a mHealth application.

Participants also described a need for alternative modes of privacy discourse outside of traditional manufacturer notices that are typically provided. To address this, participants shared that they sometimes leverage social networks to communicate about critical mHealth privacy issues. One participant, P14 stated, "I think most people would learn through media or social media quicker than probably that a business would be notifying you that your data was compromised". Thus, participants expressed that their participation in social communication channels are needed for timely information that may affect their privacy decisions. So, participants expressed sometimes experiencing collective privacy influence. This collective influence could lead to developing apathy about privacy decisions due to perceived societal norms, more stringent views due to

P10 stated, "It [consent] is definitely a gray area because technically by the book I am checking the box that I read and understand the notice. But if that is the only way that I'm going to be able to use this system [a mHealth system] that I want to use, there's not really much of an option for me to get further clarification or additional resources to fully understand my data privacy rights, as far as how that company or that service is handling things." P10 also later stated, "It's sort of the ultimatum, you either use the [mHealth] system or you don't. That's the only decision that you're allowed as a user…" Another participant, P09, shared a similar sentiment, "Do I really want to access this information, or do I want to have to go through a manual way or not do anything at all? Most of the time, I just accept it because I want to be able to pay my [medical] bill online or I want to be able to access MyChart [a personal health record] information online, and in order to do that, I have to accept it. So since you don't really get a choice and you need to get to it, you pretty much have to accept it anyway." These findings are consistent with research by (Utz et al., 2019) that notes the tensions users face when weighing tradeoffs between being informed and giving consent when interacting with privacy policies more broadly. Yet, participants discuss that when in these situations they feel they have limited autonomy especially if it is necessary or critical for them to use a mHealth system.

When manufacturers require consent without ensuring that users are sufficiently informed, or when not having access to a device or service is the only alternative option to consent, users shared that they begin to experience feelings that perpetuate transactional compliance as more important than costumer feelings or expectations. For example, P01 stated, "I think it's [providing privacy notices] strictly for compliance's sake. I don't know how many people actually read the privacy notices. So I hesitate to say it protects the consumer. I mean, it should protect consumers. It should protect both parties, quite frankly, but I just don't see that actually happening. I mean, I can't imagine, or I have to imagine the percentage of people that actually read terms of service or privacy notices or anything along those lines is remarkably small." As a result, participants believed that their interests are secondary to a manufacturer's compliance requirement, and they therefore experienced mistrust. We discovered that this mistrust is tied to the institutional systems and processes that govern mHealth applications. Although participant's degree of trust was inconclusive, some participants shared that it was attributed to unclear pre-market processes and the manufacturer or governing institution's history as it relates to quality. For example, P06 stated, "if I had a suspicion…my default is that makers have been vetted through the app store and they are trustworthy. But if I felt like there was something about their quality or trustworthiness that [there were] some sort of red flag, I might go into the privacy agreements. To be honest, if it was made… [by someone] that usually doesn't have our best interests in mind…that would motivate me to look [at] trust and quality." The limited choices related to consent and transparency of institutional practices, led participants to feel they had less autonomy in their privacy decisions related to mHealth applications.

4.2 Social Influences on Privacy Perceptions Influence Decisions

We learned that the perceptions of society or others also sometimes influenced participants' privacy decisions. Akin to the influence of social norms in TPB, we found that

Using the codebook, two researchers independently coded two of the transcripts to further refine the codebook and establish inter-rater reliability (McDonald et al., 2019) between the researchers. Inter-rater reliability was assessed to determine the likelihood that two independent reviews of the same participant transcript yield similar outcomes. Thus, we wanted to determine if the generated codes were interpreted similarly between independent reviewers. Pre-defined acceptance criteria for the reliability score was set at 80% or greater on an individual quotation level aligned with existing practice. The researchers defined rules prior to analysis, which established which transcripts would be coded; rationale supporting this decision was based on the participant's code distribution and nominal representation of educational background compared to other subjects (undergraduate degree). Of 41 codes and definitions, the inter-rater reliability score of 80% was exceeded after initial comparisons and a round of discussion to address agreements and disagreements. Once consensus was established, one researcher coded the entire subset of transcripts using the codebook.

4 Findings

We found a need for additional focus on autonomy in mHealth privacy interactions. Participants had mixed-attitudes about the value and usefulness of mhealth privacy policies. For instance, we found that subjective norms and perceived control (beyond actual data controls provided) uniquely contributed to users' sense of autonomy in interactions with mHealth privacy policies. Participants believed that these additional factors should be considered in privacy policy design to facilitate personalized, engaging, and meaningful interactions in highly dynamic mHealth privacy situations. Our results suggest that beyond the ability to control personal data, users' sense of autonomy in privacy interactions may also rely on the ability of designers to truly engage users to understand how the design of solutions are intended to protect them. In the following subsections, we present results of what participants told us about their unmet needs for autonomy with mHealth privacy interactions.

4.1 Incongruent Informed Consent is a Barrier to Engagement and Trust

We learned that participants felt they often had to consent to mHealth privacy policies without being fully informed about them. Participants did not attribute this problem to a lack of information, but rather the question of what it means to be "informed" and how the information presented (with the goal of informing) engages the user. For example, P14 explained consent is often a binary choice but emphasized the distinction between consenting and being informed. They stated, "Theoretically, yes…if it comes down to that binary choice and if the consumer is being informed, then that's consent. If you're signing up for something and you're not being informed, that's not informed consent. So that's a different argument and that's a different situation…" So, while participants mentioned a lack of engagement with policy information, they also challenged the notion that listing information in a policy is sufficient for engaging users and helping them understand its meaning.

Participants also encountered situations where they lacked understanding of the information presented but felt obligated to consent in order to access the services. For example,

voluntary consent to its practices. While we used FBM to design our study, we framed our analysis through the TPB to understand participant's interactions with mHealth privacy and how it relates to perceived autonomy (see Appendix Table 4 for an overview of this methodological process).

3.2 Participants

Participants were recruited from a local community in the surrounding areas of a midwest city in the United States. They were required to be age 18 years or older, and be current users of mhealth applications and consent to privacy policies. We chose a broad age range and were not intending to compare differences based on demographics at this phase of research. No participants were excluded. Participants' ages ranged from 28–67 years old (Appendix Table 1). All participants had a smartphone or mobile phone with internet access. Participants ranged in mHealth usage and frequency of privacy notice engagements. Participants encountered the policies through a variety of devices including Apple Watch, IoMT (Internet of Medical Things), iPhone, Alexa, smart appliances, Electronic Medical Records (EMR), and Fitbits among others. To further understand participants' existing views on data and privacy, we also asked them to share what they felt data and privacy mean (Appendix Table 2).

3.3 Study Procedures

During each 60-min semi-structured interview, we asked participants about their experiences with privacy notices when using mobile applications including health related apps. Additionally, we shared with participants various mHealth privacy notice examples as probes (Appendix Table 3) to help participants reflect on their own encounters with mHealth privacy policies, their attitudes and behaviors toward them, and factors they felt influenced their attitudes and behaviors (Hutchinson et al., 2003). Finally, we asked participants to reflect on barriers and challenges they faced, if any, and to brainstorm ideas of how they feel one might improve interactions with privacy policies to improve their sense of autonomy when engaging. Each participant was asked to complete a demographic and background survey at the end of the study. These questions were gathered to understand participant characteristics and technology experiences. We conducted interviews until we stopped hearing and seeing new data (i.e., saturation) (Chun Tie et al., 2019). After completing all interviews, we began analysis of the data.

3.4 Data Analysis

We audio recorded all interviews and transcribed them prior to data analysis. We used thematic analysis situated in Grounded Theory (GT) to analyze our data. The GT research process consists of collecting qualitative data, inductively assigning codes to data to develop themes, comparing themes with external research, and building theory from these themes (Chun Tie et al., 2019). This inductive process considers data saturation and external research comparisons to iteratively refine codes and themes to support the theory (Chun Tie et al., 2019). Once we confirmed a level of support from existing literature, we generated a codebook to guide our deductive coding process.

et al., 2021). However, similar to research on mobile crowdsourcing and trust authentication, the collection and processing methods in this environment are inadequately expounded on, thus consumers are left with gaps in knowledge about their data journey and have seemingly limited trust in the wearables they use (Feng et al., 2018). While researchers explore trust and control in various privacy models, once mHealth devices begin to collect data, ultimately, users are therefore left with the belief that their data is shared in a black-box environment intended to capitalize on their use of the technology without providing sufficient awareness of their data integrity. In this research, we explore users' perceptions of privacy policies to uncover factors they perceive to influence their autonomy in privacy policy interactions beyond the existence of privacy control(s). By doing so, we aim to understand how user self-reported behaviors are influenced, or not, by their sense of perceived autonomy in those interactions and identify design considerations for future autonomy-preserving privacy interactions.

3 Methods

Our interviews aimed to answer the following research questions:

- RQ1: What are users' current experiences with mHealth privacy policies?
- RQ2: What are users' attitudes and behaviors toward privacy policies for mobile applications that collect and use personal health data and why?

3.1 Theoretical Framing

Because we wanted to understand users' behaviors when engaging with the design of privacy policies in mhealth applications, we initially started with the Fogg Behavior Model (FBM) to help frame questions in our study protocol because of its focus on user behavior and technology design (Fogg, 2009). Fogg's model describes that user behaviors can be influenced by recognizing user motivation and ability, and potentially designing triggers that characterize those relationships (Fogg, 2009). However, we later expanded our theoretical framing during the analysis phase after exploring the data and realizing that broader concepts were emerging related to the Theory of Reasoned Action (TRA). The TRA focuses on motivations such as intents and a person's ability to act or adapt to behaviors according to them (Fishbein, 1979). Within the TRA, humans are viewed as rational decision-makers that when faced with a decision of pros and cons, adequately weigh them consistently and predictably in accordance with the most optimal economic benefit (Fishbein, 1979) – this decision-making is similar to privacy calculus in our research context. We quickly realized through iterative thematic analysis that TRA would succumb to limitations about perceptions of user control, which is why we explored a similar model that allowed us to focus on that component of behavioral intent. To characterize the relationship between intention and behavior, the Theory of Planned Behavior (TPB) describes intentions as multi-faceted, which rely on perceived levels of individual behavioral control, subjective norms, and attitudes (Icek, 1991). These dimensions of intent are dynamic and inherently conflict with behavioral economics where decisions are considered rational and reliable. In our data, we began to see concerns emerging that were related to understanding health privacy language and subsequent

confirms that privacy is both dynamic and subjective, and is susceptible to change over time along with different contexts, which is the basis for Privacy Regulation Theory. Irwin theorizes privacy as the control and feedback over information flow, which our work expands; however, the framework produced focuses heavily on contextualizing health environment monitoring solutions rather than mobile health, which we posit has proximal differences in interpretation (Moncrieff et al., 2009).

To address black-box perceptions in pervasive health technology, designing for transparency and choice are important in passive data sharing to reduce privacy concerns. Current design is still only accounting for upfront choice and transparency, and little with how choice and engagement are actively managed after data is shared (Kolovson et al., 2020). Some researchers posit "…the crux of modern machine learning: the reliance on powerful but intrinsically opaque models. When applied to the healthcare domain, these models fail to meet the needs for transparency that their clinician and patient end-users require. We review the implications of this failure, and argue that opaque models (1) lack quality assurance, (2) fail to elicit trust, and (3) restrict physician-patient dialogue. We then discuss how upholding transparency in all aspects of model design and model validation can help ensure the reliability and success of medical AI…" this forms the basis for not just opaque AI models in healthcare but also opaque data journeys in mHealth (Quinn et al., 2022). While regulation may be the de facto standard for ensuring privacy between interoperable devices like fitness trackers and smartwatches, device requirements subject to FDA and HIPAA are not widely acknowledged due to lack of awareness and misidentifying medical device classifications (Motti, 2019).

2.3 Privacy Control Versus Autonomy

Often, privacy behaviors appear to be dictated by technology that simply aims to provide control(s); through this lens, we see a challenge in autonomy due to a lack of self-direction, identity, and intrinsic factors (Deci et al., 2012). However, based on Self-Determination Theory (SDT), the premise of true autonomy in this context is the feeling that one is both being in control and willing to engage in good privacy-preserving behaviors – simply, we must transcend from designing controls to designing autonomy. We posit that privacy-by-design is being challenged in unique ways due to the complexity of systems that collect, process, and maintain data. While regulations such as GDPR and HIPAA exist to govern data practices and have an important role (Premarathne et al., 2015), their principles are collectively reduced to compliance-centric models, which leaves little room to improve usable mechanisms beyond 'cookies' (Degeling et al., 2018) or other usable privacy mechanisms. It is for this reason that existing privacy-preserving infrastructures are not fully capable to keep up with the needs of consumer mHealth innovations. An example of this resides in the health IoT environment where consumers value privacy over novel utilities and feel the two are somehow negotiated against each other (Zou et al., 2020). Researchers have aimed to address problems that exist between humans and ubiquitous computing, but mobile health wearables and applications in particular, have unique challenges related to secure interoperability between devices, databases, and governing infrastructures, which have created negative privacy perceptions (Ometov et al., 2021). These perceptions are perpetuated by the advancing need to continuously collect sensing and individual data to generate insights (Ometov

in a proactive tradition. It is no longer ethical to misconstrue preference and control as a sufficient end towards proactive privacy. Researchers enable preference selection and other usable privacy interactions as a means to promote control over one's health data, but control is only one variable that presupposes another integral and widely overlooked virtue of autonomy. It is this belief that motivates our work and distinguishes this research from others that focus on elderly populations (Detweiler et al., 2016). More specifically, the composition that makes up autonomy is not well understood and established within advancing interconnected health systems. We will further explore the topic in later sections.

2.2 Pervasive mHealth and Black Box Use Cases

Pervasive health through the use of sensor technology has generated broad and deep insights (Wang et al., 2022). Some mHealth sensing information architectures and functionality enable health interventions by leveraging behavioral change through different engagement techniques. Some features that drive engagement such as forums present unique privacy challenges as well (Danaher et al., 2015). Other applications such as virtual health communities' research has also explored the topic of privacy (Kordzadeh et al., 2017), yet much of this research focuses on supporting human-human communication rather than human-machine communication, which makes the domain unique. What makes this area of HCI unique is the Mobile health (mHealth) component, which is defined as, "the use of mobile devices to monitor or detect biological changes in the human body, while device management entities, such as hospitals, clinics, or service providers, collect data and use them for healthcare and health status improvement" (Park, 2016), and is similar to others' (Ruotsalainen et al., 2012) definition in the context of pervasive health. mHealth can also include self-reported health data provided by users through consumer-focused personal tracking and reporting applications (Radbron et al., 2019). Although the growing ubiquity of mHealth applications has seemingly large potential upside to improve health through innovative and connected solutions such as IoT (Bertino et al., 2016), researchers are faced with navigating the need for large amounts of data with the complex domain of opaque health privacy (Quinn et al., 2022). To this end, many mHealth technologies have the large upside potential to transform healthcare through integrated machine learning capabilities and artificial intelligence. Although many contributions have been made in this arena, health-related stigmas can influence privacy perceptions and perpetuate concerns of the technology's utility (Arora et al., 2014). Design for sharing behavioral data in social constructs as leverage of peer support for health monitoring; also establishes engagement with data privacy across a lifecycle as an interesting research avenue (Vilaza et al., 2019).

Even though policies such as HIPAA provide protections for personal health data, users often still have concerns about what data is collected about them and how it is used (Al Ameen, 2012). As such, researchers are exploring ways to reduce negative impacts and perceptions through contextualizing privacy concerns in this space (Ferreira et al., 2021). For example, some researchers note that some mHealth privacy concerns are associated with age and can be used to tailor mobile applications to these users (Ferreira et al., 2021). Significant work has been done to define regulatory frameworks and user constraints in IoT environments (Poyner et al., 2018). Other work by Irwin Altman

data privacy decisions. In addition, we noticed that each probe, while having some common elements, had unique user interfaces from a visualization perspective. We found that certain themes were consistent with existing usable privacy research. We also found that users' sense of autonomy, perceived control and willingness (Deci et al., 2012) in mHealth privacy policy interactions were often influenced by factors other than available data control and consent. Our work builds on usable privacy and mHealth behavioral research by extending knowledge of factors that impact users' interactions with mHealth data policies. Our work extends prior research (Atienza et al., 2015) exploring users' attitudes and behaviors toward existing privacy policies. However, we focus in the context of mHealth data exploring users' experiences engaging with applications that collect their personal health information to support them in managing their health. Our work contributes to the broader research community by merging concepts of behavioral design and usable privacy to improve language understanding, and promote trust in this environment. Specifically, we extend prior research (Audie et al., 2015) that explores users' attitudes and behaviors toward existing privacy policies.

2 Related Work

We acknowledge a few foundational domains which foreground our work. We see an evolving data collection surge where emerging questions of privacy and trust ensue. We believe health data privacy is particularly relevant at this intersection of technology and ethics, and describe these domains in detail below.

2.1 Ethics on Privacy, Trust, and Technology Acceptance

Researchers in the field of ethics have considerably investigated privacy, trust, and acceptance. In a world where pervasive automation advances significantly, AI researchers have developed frameworks to optimize personal autonomy (Calvo et al., 2020) and foreground risks (Floridi et. al., 2018). Frameworks in this space consider privacy a pillar of ethical design and essential for technology acceptance, especially in the mHealth domain (Mantovani et.al., 2017). As such, privacy as a construct has a paramount position that does not only facilitate ethical AI, but also affects utility, control, trust, and acceptance goals. Researchers recognize the broad application of AI, but ethical design must establish privacy as a basic individual right that withstands the deliverance of evolving pervasive systems (Bartoletti, 2019). The reasons consumers have a strong affinity for privacy is due to several complex factors. Researchers know that variation in demographics such as age and the type of data collected (i.e., health) can either help or hurt the trust they have in a health technology, and ultimately its acceptance or adoption (Poyner et al., (2018), Schomakers et al., (2019), Wang et al., (2019), Martin-Hammond et al., (2019), Guo et al., (2016)). While significant work has been done to improve consumers' willingness to accept health technology, some experience a sense of fatalism that is perpetuated by the evolving health system they interact with (Joo et al., (2021)). This fatalism is a sign of migrating chasms between the nature of perceived privacy and the growth at which consumers are exposed to new health promotions. Researchers explore this intersection, but many do not entirely approach grounding health privacy

promising results for reducing users' cognitive load through limiting excessive reading and decreasing users' burden through design considerations (Schaub et al., 2017). mHealth privacy research has also examined consumer's abilities to consent to data practices, such as how their data is used and stored, which is a core component of Health Insurance Portability and Accountability Act (HIPAA), General Data Protection Regulation (GDPR), and the Common Rule, three regulations that regulate the processing of personal data, outline provisions of human subject research, and safeguards privacy of medical data and other personal data (Nurgalieva et al., (2020), Arora et al., (2014)). Practitioners have also explored technical solutions that improve users' privacy awareness and ability to comply with regulatory requirements (i.e., GDPR and HIPAA) (Iwaya et al., 2022). Yet, oftentimes existing practices and approaches emphasize obtaining consent, sometimes neglecting that people ignore, or fail to understand the risks and implications of using an application or their participation in its data usage (Degeling et al., (2018), Schaub et al., (2017)). Therefore, users are often faced with the classic tradeoff between application (i.e., app) utility and privacy which ultimately leads to a black box where users are not fully informed about their data privacy rights. This presents a chasm for users and privacy policy designers positioned at the intersection of legal compliance and usable privacy design.

The rapid emergence of connected mHealth solutions has enabled more personalized and informed care (Steinhubl et al., 2015) but the ability to understand user attitudes and behaviors towards mHealth data privacy is a known trust-related barrier to user adoption (Lynch et al., (2017), "Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care", (2010), (Zou et al., 2020)) and remains a challenge. One open challenge is that these solutions often ignore other relevant factors such as dynamic intent or perceived control that might impact users' behaviors in the context of healthcare (Ruotsalainen et al., 2012). We must therefore further understand the factors beyond data control and consent that influence user behaviors in the mHealth context to identify appropriate opportunities and solutions to address users' needs when interacting with privacy policies. It is imperative to better understand the intricacies of individual data privacy behaviors when interacting with mHealth applications to derive further design considerations that can inform this ubiquitous and evolving data-driven environment of mHealth. We posit that user attitudes and behaviors have a deterministic contribution in helping to identify the strengths and limitations of current privacy policies that are designed to help motivate individual data privacy behaviors, engagement, and understanding.

In this paper, we investigate users' attitudes and self-reported behaviors when engaging with mHealth data privacy policies to understand context-specific factors and opportunities to improve mHealth privacy policy design. We conducted interviews with 15 adults that use mHealth applications to understand their attitudes and behaviors toward existing mHealth privacy policies, challenges, and opportunities for improvement of these policies. During interviews, we used a focused set of probes (Appendix Table 3) to support reflection when sharing their prior experiences with mHealth data privacy policies. We selected these applications because they covered broad reaching health domains ranging from wellness (i.e., sleep) to mission-critical healthcare management (i.e., diabetes care and medication adherence), which were believed to have unique elements in

Exploring Users' Perspectives of Mobile Health Privacy and Autonomy

Thomas Starks(✉), Kshitij Patil, and Aqueasha Martin-Hammond

School of Informatics, Computing, and Engineering, Indiana University – Indianapolis, Indianapolis, IN 46202, USA
{tmstarks,kshpatil,aqumarti}@iu.edu

Abstract. The increased use of mobile health (mHealth) applications and the corresponding exchange of sensitive data has underscored privacy concerns. Privacy notices are often unengaging or incomprehensible, leading to questions of informed consent and trust. While studies have focused on providing solutions aimed to simplify privacy language and reduce cognitive burden, often overlooked are the behavioral aspects of individual attitudes, norms, and perceived control that lead to dynamic intentions for engagement. In this paper, we use existing behavior models as a lens to understand users' privacy experiences, behaviors, and perspectives toward mHealth data privacy policies. In 15 semi-structured interviews with adult users of mHealth applications, participants encountered persistent challenges when engaging and articulating the value of privacy. Participants do not understand how privacy notices are designed, which leads to superficial awareness and control that does not actually support their perceptions of autonomy and trust in mHealth. As a result, users felt sub-optimal autonomy when engaging in privacy interactions. We discuss design considerations for autonomy-supporting privacy notices that may help users feel a greater sense of agency when interacting with mHealth applications.

Keywords: Human-centered computing · Human computer interaction (HCI) · Empirical studies in HCI · Security and privacy · Human and societal aspects of security and privacy · Usability in security and privacy First Section

1 Introduction

Technology-driven health solutions such as mHealth applications are both prolific and challenging for privacy policy researchers, designers, and practitioners. mHealth applications are categorized as any mobile device that captures and obtains health-related data to improve quality-of-care (Cameron et al. 2017), which span diabetes management, sleep, medication, and general health and wellness, among others. The data accompanying mHealth applications require privacy policy designers to consider both regulatory compliance and individual privacy behaviors when crafting user policies, frameworks and solutions that meet privacy goals. For example, privacy design research has shown

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 70–91, 2024.
https://doi.org/10.1007/978-3-031-59717-6_5

32. Powell, A.: AI Revolution in Medicine. Harvard Gazette, November 2020. https://news.harvard.edu/gazette/story/2020/11/risks-and-benefits-of-an-ai-revolution-in-medicine/
33. Qiu, S., Liu, Q., Zhou, S., Wu, C.: Review of artificial intelligence adversarial attack and defense technologies. Appl. Sci. **9**(5), 909 (2019)
34. Quinn, T.P., Jacobs, S., Senadeera, M., Le, V., Coghlan, S.: The three ghosts of medical AI: can the black-box present deliver? Artif. Intell. Med. **124**, 102158 (2022)
35. Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., Qadir, J.: Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput. Biol. Med. 106043 (2022)
36. Ross, P., Spates, K.: Considering the safety and quality of artificial intelligence in health care. Jt. Comm. J. Qual. Patient Saf. **46**(10), 596 (2020)
37. Rubinger, L., Gazendam, A., Ekhtiari, S., Bhandari, M.: Machine learning and artificial intelligence in research and healthcare. Injury (2022)
38. Scott, I., Carter, S., Coiera, E.: Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform. **28**(1) (2021)
39. Seppänen, M., Hyrynsalmi, S., Manikas, K., Suominen, A.: Yet another ecosystem literature review: 10+1 research communities. In: 2017 IEEE European Technology and Engineering Management Summit (E-TEMS), pp. 1–8 (2017). https://doi.org/10.1109/E-TEMS.2017.8244229
40. Sujan, M.A., White, S., Habli, I., Reynolds, N.: Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare. Saf. Sci. **155**, 105870 (2022)
41. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J.: How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat. Med. **27**(4), 582–584 (2021)
42. Xing, L., Giger, M.L., Min, J.K.: Artificial Intelligence in Medicine: Technical Basis and Clinical Applications. Academic Press, Cambridge (2020)
43. Yang, L., Ene, I.C., Arabi Belaghi, R., Koff, D., Stein, N., Santaguida, P.L.: Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur. Radiol. **32**(3), 1477–1495 (2022)

14. Jia, Y., McDermid, J.A., Lawton, T., Habli, I.: The role of explainability in assuring safety of machine learning in healthcare. IEEE Trans. Emerg. Top. Comput. (2022)
15. Jiang, L., et al.: Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. J. Int. Med. Res. **49**(3), 03000605211000157 (2021)
16. Kallus, N., Puli, A.M., Shalit, U.: Removing hidden confounding by experimental grounding. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
17. Lavrakas, P.J.: Encyclopedia of Survey Research Methods. Sage Publications, Thousand Oaks (2008)
18. van Leeuwen, K.G., Schalekamp, S., Rutten, M.J., van Ginneken, B., de Rooij, M.: Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur. Radiol. **31**(6), 3797–3804 (2021)
19. Lekadir, K., Quaglio, G., Garmendia, A.T., Gallin, C.: Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts. EPRS (European Parliamentary Research Service) (2022)
20. Macrae, C.: Governing the safety of artificial intelligence in healthcare. BMJ Qual. Saf. **28**(6), 495–498 (2019)
21. Magrabi, F., et al.: Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb. Med. Inform. **28**(01), 128–134 (2019)
22. Manikas, K.: Revisiting software ecosystems research: a longitudinal literature study. J. Syst. Softw. **117**, 84–103 (2016). https://doi.org/10.1016/j.jss.2016.02.003, https://www.sciencedirect.com/science/article/pii/S0164121216000406
23. Manikas, K.: Supporting the evolution of research in software ecosystems: reviewing the empirical literature. In: Maglyas, A., Lamprecht, A.-L. (eds.) Software Business. LNBIP, vol. 240, pp. 63–78. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40515-5_5
24. Manikas, K., Hansen, K.M.: Software ecosystems – a systematic literature review. J. Syst. Softw. **86**(5), 1294–1306 (2013). https://doi.org/10.1016/j.jss.2012.12.026, https://www.sciencedirect.com/science/article/pii/S016412121200338X
25. Martin, C., et al.: The ethical considerations including inclusion and biases, data protection, and proper implementation among AI in radiology and potential implications. Intell.-Based Med. 100073 (2022)
26. McCradden, M.D., Joshi, S., Anderson, J.A., Mazwi, M., Goldenberg, A., Zlotnik Shaul, R.: Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. J. Am. Med. Inform. Assoc. **27**(12), 2024–2027 (2020)
27. Moore, C.M.: The challenges of health inequities and AI. Intell.-Based Med. 100067 (2022)
28. Muehlematter, U.J., Daniore, P., Vokinger, K.N.: Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit. Health **3**(3), e195–e203 (2021)
29. Newaz, A.I., Sikder, A.K., Rahman, M.A., Uluagac, A.S.: A survey on security and privacy issues in modern healthcare systems: attacks and defenses. ACM Trans. Comput. Healthc. **2**(3), 1–44 (2021)
30. Page, M.J., et al.: The Prisma 2020 statement: an updated guideline for reporting systematic reviews. Syst. Control Found. Appl. **10**(1), 1–11 (2021)
31. Paton, C., Kobayashi, S.: An open science approach to artificial intelligence in healthcare. Yearb. Med. Inform. **28**(01), 047–051 (2019)

[25] Martin, C., DeStefano, K., Haran, H., Zink, S., Dai, J., Ahmed, D., Razzak, A., Lin, K., Kogler, A., Waller, J., et al.: The ethical considerations including inclusion and biases, data protection, and proper implementation among ai in radiology and potential implications. Intelligence-Based Medicine p. 100073 (2022)
[12] Gupta, S., Gupta, A.: Dealing with noise problem in machine learning datasets: A systematic review. Procedia Computer Science **161**, 466–474 (2019)
[1] Barh, D.: Artificial Intelligence in Precision Health: From Concept to Applications. Academic Press (2020)
[13] Hamid, S.: The opportunities and risks of artificial intelligence in medicine and healthcare. Apollo - University of Cambridge Repository (2016)
[16] Kallus, N., Puli, A.M., Shalit, U.: Removing hidden confounding by experimental grounding. Advances in neural information processing systems **31** (2018)

References

1. Barh, D.: Artificial Intelligence in Precision Health: From Concept to Applications. Academic Press, Cambridge (2020)
2. Bohr, A., Memarzadeh, K.: Artificial Intelligence in Healthcare. Academic Press, Cambridge (2020)
3. Borycki, E., Kushniruk, A.: Artificial intelligence and safety in healthcare. In: AI and Society, pp. 17–32. Chapman and Hall/CRC, Boca Raton (2022)
4. Briganti, G., Le Moine, O.: Artificial intelligence in medicine: today and tomorrow. Front. Med. **7**, 27 (2020)
5. Crossnohere, N.L., Elsaid, M., Paskett, J., Bose-Brill, S., Bridges, J.F.: Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J. Med. Internet Res. **24**(8), e36823 (2022)
6. Center for Devices and Radiological Health: Artificial intelligence and machine learning (AI/ML)-enabled medical d, October 2022. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
7. Galitsky, B., Goldberg, S.: Artificial Intelligence for Healthcare Applications and Management. Academic Press, Cambridge (2022)
8. Geiping, J., Fowl, L., Somepalli, G., Goldblum, M., Moeller, M., Goldstein, T.: What doesn't kill you makes you robust (ER): adversarial training against poisons and backdoors. arXiv preprint arXiv:2102.13624 **1**(7) (2021)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
10. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
11. Group, I.S.W., et al.: "Software as a medical device": possible framework for risk categorization and corresponding considerations. In: International Medical Device Regulators Forum (2014)
12. Gupta, S., Gupta, A.: Dealing with noise problem in machine learning data-sets: a systematic review. Procedia Comput. Sci. **161**, 466–474 (2019)
13. Hamid, S.: The Opportunities and Risks of Artificial Intelligence in Medicine and Healthcare. Apollo - University of Cambridge Repository (2016)

[35] Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., Qadir, J.: Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Computers in Biology and Medicine p. 106043 (2022)
[9] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
[8] Geiping, J., Fowl, L., Somepalli, G., Goldblum, M., Moeller, M., Goldstein, T.: What doesn't kill you makes you robust (er): Adversarial training against poisons and backdoors. arXiv preprint arXiv:2102.13624 **1**(7) (2021)
[33] Qiu, S., Liu, Q., Zhou, S., Wu, C.: Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences **9**(5), 909 (2019)
[10] Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
[42] Xing, L., Giger, M.L., Min, J.K.: Artificial intelligence in medicine: technical basis and clinical applications. Academic Press (2020)
[3] Borycki, E., Kushniruk, A.: Artificial intelligence and safety in healthcare. In: AI and Society, pp. 17–32. Chapman and Hall/CRC (2022)
[19] Lekadir, K., Quaglio, G., Garmendia, A.T., Gallin, C.: Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts. EPRS (European Parliamentary Research Service) (2022)
[26] McCradden, M.D., Joshi, S., Anderson, J.A., Mazwi, M., Goldenberg, A., Zlotnik Shaul, R.: Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. Journal of the American Medical Informatics Association **27**(12), 2024–2027 (2020)
[20] Macrae, C.: Governing the safety of artificial intelligence in healthcare. BMJ quality & safety **28**(6), 495–498 (2019)
[27] Moore, C.M.: The challenges of health inequities and ai. Intelligence-Based Medicine p. 100067 (2022)
[7] Galitsky, B., Goldberg, S.: Artificial Intelligence for Healthcare Applications and Management. Academic Press (2022)
[4] Briganti, G., Le Moine, O.: Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine **7**, 27 (2020)
[31] Paton, C., Kobayashi, S.: An open science approach to artificial intelligence in healthcare. Yearbook of medical informatics **28**(01), 047–051 (2019)
[42] Xing, L., Giger, M.L., Min, J.K.: Artificial intelligence in medicine: technical basis and clinical applications. Academic Press (2020)
[37] Rubinger, L., Gazendam, A., Ekhtiari, S., Bhandari, M.: Machine learning and artificial intelligence in research and healthcare. Injury (2022)
[34] Quinn, T.P., Jacobs, S., Senadeera, M., Le, V., Coghlan, S.: The three ghosts of medical ai: Can the black-box present deliver? Artificial intelligence in medicine **124**, 102158 (2022)
[14] Jia, Y., McDermid, J.A., Lawton, T., Habli, I.: The role of explainability in assuring safety of machine learning in healthcare. IEEE Transactions on Emerging Topics in Computing (2022)

can be attributed in part to the immaturity of the field, however, the potential impact of risks in the medical domain are severe. The dynamic nature of AI models compared to traditional medical devices requires a stronger focus on the post-market phase from the regulators. Therefore, the study underscores the need for a more comprehensive understanding, and clear and robust regulatory guidelines to navigate through these potential hazards.

7 Conclusion

In this paper we investigate the adoption of AI in medical devices. Currently, concerns regarding the safety risks surrounding AI-based medical devices currently stand in the way of their wider adoption. In this study we conduct: (1) a survey of the safety risks of AI-enabled Medical Devices published between 2012 and 2023, (2) an analysis of AI-based medical devices in the EUDAMED database. and (3) a survey on the perceptions of Medical AI ecosystem stakeholders. Our analysis body includes 29 reviewed papers, 71 AI-based medical devices and seven responded questionnaires out of an original 130 participants. Our findings show that the presence of unique risks, such as bias or lack of transparency, in AI-enabled Medical Devices is undeniable. Looking at data available at EUDAMED we can see that it is currently hard to even pinpoint which devices in EU use AI and we have to look at company websites, press statements or published papers to discover that. We also uncovered that many AI enabled devices in EU deal with severe conditions such as arrhythmia or stroke, which further underlines the severity of potential risks manifesting. Experts and companies in the Medical AI ecosystem feel a need for guidance and regulation that covers the whole life-cycle of AI products, with more emphasis on the post-market phase, and incorporates aspects related to data-centric risks of the products. This demonstrates an openness to more structured guidelines from the industry. However, our research suggests that regulators feel that do not have expertise in AI, indicating that a gap exists between the complexities of AI technology and the understanding of those responsible for its oversight. Based on the findings we propose, that more clear and encompassing regulatory guidelines would be needed to mitigate the risks of AI-enabled Medical Devices in EU.

Appendix A - Literature Body of the Literature Survey

[21] Magrabi, F., Ammenwerth, E., McNair, J.B., De Keizer, N.F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P.J., Vehko, T., Wong, Z.S.Y., et al.: Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications. Yearbook of medical informatics **28**(01), 128–134 (2019)
[38] Scott, I., Carter, S., Coiera, E.: Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health & Care Informatics **28**(1) (2021)
[2] Bohr, A., Memarzadeh, K.: Artificial intelligence in healthcare. Academic Press (2020)

to fill out the survey represents a broader trend among regulators. This could point to a serious issue in the EU legislation of AI-enabled medical devices.

The concerns and pain-points pointed out by the participants show that many risks in AI -enabled medical devices are data-centric, such better guidelines on sufficient amounts of data. This feedback from actors currently active in the ecosystem indicates that EU regulators have not sufficiently addressed several data-centric risks of AI-enabled medical devices.

Furthermore the survey participants also underlined the need for better post-market guidance. This highlights another unique feature of AI that the EU might not have tackled sufficiently. Namely the changing nature of AI models and algorithms and the large amounts of risks stemming of it that can manifest in the post-market phase. Unlike many traditional medical devices such as contact lenses or pumps, AI performance can vary widely in different locations, for example in different hospitals and can also dangerously degrade when coming in contact with new data while in productions. These unique aspects would need to be clearly addressed by the EU regulators.

6 Discussion

The majority of AI devices are identified in the survey on EUDAMED were within the field of radiology or cardiology and most commonly dealt with critical conditions or illnesses, such as cancer or stroke. The fact that devices on the EU market commonly deal with critical illnesses highlights the potential severe outcomes of various risks not being properly mitigated. It was challenging for this study to identify AI in devices as the current information in EUDAMED is inadequate. EUDAMED does not contain information on whether a device is utilising AI and lacks information that would be needed to assess the safety of AI-based devices, e.g. information on the data used for the device.

In the literature survey of the risks and mitigation strategies of AI-enabled Medical Devices most papers discussed data-centric risks in various detail. Other risk categories identified were cybersecurity risks, transparency related risks and lastly user and system interaction related risks. The amount and nature of risks identified in combination with the domain mission criticality of the devices underline the importance of good praxis in the adaptation of AI and the high risks in improper adaptation.

The analysis of stakeholder perceptions found that several post-market data-related risks presented in the academic literature, such as bias, were also a concern for the stakeholders, who emphasised the need for better testing and regulatory guidance to address such risks. Some stakeholders felt that the current EU regulation on Medical-AI is inadequate, citing a lack of post-implementation guidance and guidelines on data sufficiency as examples. This points to a need for regulatory guidelines that in a larger degree take into account the dynamic and data-centric properties of AI enabled medical devices. However, there was indication that regulators feel a lack of expertise about AI.

The findings of this paper highlight lack of regulation and establishment of common understanding of safety risks of AI-enabled Medical Devices. This

5.2 AI-Based Medical Devices in EU

The current data fields available in EUDAMED point to a possible regulatory issue, as they do not contain a lot of information that would be needed to assess the safety of AI-based devices, such as information about the data - for example potential biases in the training data or amount of data used for training. As the review of the safety risks showed, having clear documentation about the design of the system, including the data used, helps mitigate risks associated with AI-enabled devices, such as the black-box nature of such devices.

Analysis of the distribution of risk classes revealed that the medical AI devices in the EUDAMED dataset had a higher proportion of "class IIa" and "class IIb" risk classes and a lower proportion of "IVD general" risk classes compared to the non-AI devices.

"Class IIa" risk classes are considered to be of low to moderate risk, while "class IIb" risk classes are considered to be of moderate risk. "IVD general" risk classes are not included in class IIa or IIb, and their risk level is not specified.

This suggests that the medical AI devices in EUDAMED are often lower to moderate risk compared to the non-AI devices. It is worth noting that this comparison is based on the proportion of devices in each risk class, and it is not necessarily indicative of the overall risk level of the medical and non-AI devices.

The second finding of the analysis of EUDAEMD is that most AI-enabled devices are dealing with critical or serious illnesses and conditions, such as stroke or cancer. Analysis of the correlation between the severity of the condition or illness targeted and the risk class of the device showed that devices targeting serious illnesses and conditions doe not necessarily get a higher risk class. This is not surprising as, the severity of the targeted condition or illness is just one factor among many considered when assigning a risk class to a medical device, with the intervention of the device carrying the most significant weight.

5.3 Survey of Stakeholder Perceptions

The low number of survey participants means that the results are not suitable for representing the medical device ecosystem as a whole, since such a small sample size can lead to sampling bias. However, the results are still useful for supplementing the literature review and for providing insights into potential safety concerns form the stakeholders perspective. Furthermore, the finding points out potential pain-points in the EU regulation of medical devices from the stakeholders perspective. Additionally, the results could potentially be used to inform future research or to identify areas for improvement in the distribution process. The fact that none of the regulators filled out the survey coupled with the fact that one of the regulators reported that he feels that they do not have sufficient information to fill out the survey, points to possible gap in regulators knowledge of AI- enabled devices. While it must be noted that since we only have one datapoint we currently have weak evidence. None the less, this is an interesting finding that could point to a future research direction. It is possible that the regulator who wrote back indicating that they did not have sufficient knowledge

When examining the current regulation of AI enabled devices three participants did not feel that the current system in EU was sufficient in terms of ensuring the safety of AI/ML based Medical Devices, while two participants were unfamiliar with the system and two felt that the current system was sufficient. Reasons noted for feeling that the system was insufficient were: (i) No guidance had been published by the European authorities on surveillance following implementation; (ii) Notified Body scrutinises for safety, clinical experts review Clinical Evaluation Report; (iii) Does not sufficiently account for potential biases or oversights in the training and validation data; (iv) Re-certification of data-centric AI and learning algorithms are not fully incorporated; (v) Lack of life-cycle understanding;

Lastly, the changes the participants would like to see in the medical device regulation were: (a) more focus on post market surveillance; (b) more streamlined process, that are less dependant on the availability of notified bodies or their specific interpretations; (c) Better guidelines for post-market surveillance; (d) Better guidelines on ensuring sufficiency of data.

5 Analysis

5.1 Literature Survey

It is evident from the literature survey that the main risks stem from the AI-enabled devices reliance and interaction with data - not only during the pre-launch phase, but also during production.

This is due to the fact, that unlike traditional medical devices that function in a rather predictable, deterministic way, AI-enabled devices can evolve and change their behavior based on the data they interact with.

This means that for AI-enabled devices, post-market surveillance and real-world performance monitoring are as, if not more, important. This requires a change in regulatory frameworks, which have traditionally focused heavily on the pre-market phase where devices are tested extensively in controlled lab settings.

Second aspect, unique to AI-enabled medical devices, is risks related to the interpretability of AI-enabled Devices. Dangerous or unhelpful patterns learned by the model can be difficult to detect, as for many AI- models it is difficult and at times impossible to understand why they have reached a certain conclusion. Furthermore, the lack of transparency can exacerbate risks related to user-system interaction. To address this, regulatory frameworks need to include requirements on the level of transparency and interpretability of AI systems. A very promising direction is the use of explainable AI (XAI) techniques, that aim to make the decision-making process of AI systems more interpretable to humans.

In conclusion, the dynamic nature of AI-enabled medical devices, as well as their complexity, calls for a significant shift in thinking when designing regulatory frameworks.

Table 1. Risks and mitigation strategies identified in the literature

Risks and Mitigation Strategies	
Risk	Mitigation Strategy
Bias [3,21]	1. Pooling of data from various countries and organisations to create large and diverse data sets,across various areas, such as race and ethnicity [3,42] 2. Verify AI technology product claims on local data set [3] 3. Comprehensive multi-location evaluation studies to identify instabilities [19] 4. Reporting performance of models across relevant subgroups [26]
Hidden Confounders	No Mitigation Strategy Suggested
Noise and artefacts in model inputs [12]	Polishing, such as relabeling of data, or filtering out the noise [12]
Adversarial Attacks	1. Adversarial training 2. Generative Adversarial Network (GAN) 3. Statistical approaches [2]
Data Privacy Attacks [2]	1. De-identification algorithms [2] 2. Federated approaches for decentralised AI [19] 3. Full disk encryption [2] 4. Masking measures [2]
Lack of Tranparency (Blackbox nature of AI) [20,21]	1. An 'AI passport' for standardised description and traceability of medical AI tools [19] 2. Auditing [21] 3. For some models, visualization software, i.e. as SHAP and LIME [27,38]
Input errors [7]	Providing the user with background information and a glossary of clinical terms used in the model [7]
Cognitive biases	Training healthcare specialists to not lose vigilance [21]
Automation Complacency [21]	1. Improving the interpretability of AI systems [21] 2. Curriculum combining medicine and engineering to allow for better understanding of the workings of models [4] 3. Training healthcare specialists to not lose vigilance [21]

(in a scale from 1-5). The aspect receiving the lowest score in terms of influence was *the medical specialisation the device is deployed in* with an average of 3. Whether it is used for critical, serious or non critical, illness/condition received an average rating of 4.1 The remaining scores were: *Testing the AI algorithm on data from the hospital where it is deployed in the launch phase* - 3.28. *How much data has been used to train the device* - 4.14. *Interaction between the user and the device* - 4.14. *The users understanding of AI/ML* - 3.28.

GAN-Generative Adversarial Network (GAN) is a type of machine learning model, that can be used to generate adversarial data for a model to classify to bolster a model's robustness against attacks [33].

Statistical Approaches - using statistical tests, adversarial inputs from the operational data can be detected. Statistical tests rely on the fact that adversarial examples are statistically different from other inputs [10].

Federated Learning. This technique allows the training of an algorithm on sensitive data, present at multiple decentralized sites, without the exchange of data. For example, a number of hospitals can contribute toward the training of a model, without the data itself ever leaving each hospital's data center [42].

4.2 EU AI-Based Medical Device Survey

At the time of the gathering the data of this paper[4], the EUDAMED database lists 955 medical software items, which are reduced to 765 unique devices after eliminating duplicates. Excluding lower risk devices left 327, with 5 listed in the AI Radiology database, 13 in the FDAs database for AI based medical devices The other devices are manually labeled following the protocol outlined in Chap. 3. This results in 71 AI devices. These are labeled as serious (10), non-serious (20), or critical (41). More AI devices are classified as class IIa (low to moderate risk) than non-AI devices (72% versus 68%). The most common target body parts are the heart (13 devices) and lungs (10 devices), with 13 devices targeting multiple parts. Some devices belong to two specialities. The most common speciality the device are aimed at was radiology with 28 devices, followed by cardiology with 14 devices.

4.3 Survey of Stakeholder Perceptions

The survey is send out to 130 potential respondents. Seven provided a valid response. Four of the respondents are experts and three are working in a SaMD company. The responses do not include any regulators. However, one of the regulators reports that they feel that they do not have sufficient information to fill out the survey.

The respondents report that *pathology* and *emergency* medicine are areas that AI can be used while *radiology* and *nuclear medicine* are areas where AI is underused.

Participants are requested to evaluate various aspects of AI-enabled medical devices for their potential impact on device safety. The selected characteristics are based on prominent elements from the literature survey and guidelines from the International Medical Device Regulators Forum [11].

In this question, the participant assess that whether the devices are: (a) informing of options for treatment/diagnosing, or (b) for aiding in treatment or in diagnoses. (b) had the most influence. This element received a score of 4,6

[4] Extraction date: 2022.09.29.

Cybersecurity Risks [2,38]. AI-enabled medical devices largely share common cybersecurity risks with non-AI healthcare systems, but the use of AI in healthcare increases exposure to data privacy and integrity risks, due to creating an increased need between the interconnectivity between systems and dataset [2]. Resulting attacks can compromise model accuracy, lead to harmful predictions, re-identify de-identified data, or result in data loss. Various cybersecurity risks are discussed below.

AI increases *reidentification* opportunities in anonymized patient datasets, exemplified by Liangyuan's research [2], which demonstrated that over 90% of adults' physical activity data could be reidentified using ML models.

Adversarial Attacks on AI, categorized into white-box attacks that employ subversion, such as gradient-based techniques, and black-box attacks that poison datasets [2] leading to harmful or incorrect predictions or undetectable software corruption, are not easily detectable [38].

The risks discussed can also interplay and mutually amplify each other, such as the interaction between sampling and diagnostic bias or automation complacency and lack of transparency, with the latter making it more difficult to identify bias in the training data.

Mitigation Strategies. Most papers included in the survey presented potential mitigation strategies for the risks. In this chapter mitigation strategies from the summary table warranting additional clarification are described.

Transparency. *Visualization tools for increased transparency.*

Visualization tools for ML predictions, like Local Interpretable Algorithm-Agnostic Explanations (LIME) and Shapley Values (SHAP), help visualize the key features influencing the algorithm's predictions However, a challenge remains in clinicians understanding the language of these explanations. To address this, a platform connecting medical experts with ML researchers could help establish standardized representations of explanations [35].

Cybersecurity. *Encryption* various encryption measures are usually employed for data in transfer [2].

Adversarial Training - machine learning technique, which improves models robustness and generalization ability by training it to learn data samples that are designed to be have small and often human-imperceptible differences from the original data, but, which a model misclassifies [9]. For example images, with added pixels. This technique helps the model becomes more resistant to errors and to better handle real-world inputs that may be similarly ambiguous [8].

Masking Measures. Masking techniques, such as adding random statistical noise, collapsing variables, creating synthetic data, or using ML models to generate statistically similar datasets, are commonly employed when sharing data with external stakeholders to protect sensitive information [2].

setting are physicians prescribing medication based on indicators not present in the health record [16]. They can reduce both model generalizability and interpretation [42].

Transparency-Related Risks [14,19–21,34,37,42]. Risks related to the explainablility or interpretability also referred to as transparency of AI based devices were also a common topic in the literature.

Lack of Transparency: i.e. the 'black box' nature of many AI systems, like deep learning models, makes their decision-making process unclear [20,21,42]. This lack of transparency can make it difficult to determine the accuracy or reliability of the AI's output, may erode trust of patients and healthcare specialists and may make it more difficult to identify and correct errors [14,34,37]. A well-known example of a difficultly identifiable error is a case of AI models predicting pneumonia mortality risk mistakenly labeled asthma patients as lower risk of mortality, since they were treated more aggressively and quickly according to hospital protocol, which reduced their risk of death [42]. Transparency can be split into [19]: *Traceability* (clarity of AI development and usage) and *Explainability* (clarity of AI decisions).

User and System Interaction Related Risks [7,19–21,26,31,38]. The least discussed aspect influencing the safety of the devices was User and System Interaction related risks. Examples of such risks are input errors, automation complacency, and cognitive bias of the user. Input errors can occur as a result of the user misspelling, confusing clinical terms, users employing local definitions or misrepresenting findings. This issue is further exacerbated by quickly changing medical definitions [7]. Automation complacency refers to when specialists rely too heavily on the models predictions. Research has shown that specialists tend to over-rely and delegate full responsibility to systems and lose vigilance or become deskilled [21,31]. Evidence suggests that when a clinician is uncertain, they may defer to models predictions [26]. The black box nature of many modern AI-systems will likely contribute to the worsening of this phenomenon [21,38]. Over time, automation complacency might lead to misdiagnoses and inappropriate interventions, as algorithms may lean towards overdiagnosis by detecting subclinical findings [38].

Cognitive bias of the user includes errors that are closely related to automation complacency of the users. Cognitive biases have many forms and can include to misunderstandings of statistics and mathematical rationality or be one of many forms of human cognitive biases, such as Search satisfying: Ceasing to look for further information or alternative answers when the first plausible solution is found [7].

Various User and System Interaction risks are exacerbated by healthcare specialists limited knowledge of AI. Varied studies have showed that, that healthcare specialists have received little education regarding AI and do not rate their knowledge of AI highly hidden [19].

risks stemming from user-machine interaction. Summary of various risks and their mitigation strategies can be found in Table 1 at the end of the chapter.

Results of the survey show that data-related risks are most prevalent in the literature. These include aspects such as data drift, distributional shift or calibration drift.

Data-Centric Risks [1,3,12,13,19,21,25,34,38]. Many risks mentioned in the literature stem from the data used for the Artificial Intelligence.

Bias is a prominent topic in literature and mainly stems from disparities between training and operational data. Bias errors occur, since machine learning models do not generalize well beyond the data they were trained on [21]. Bias has different forms such as distributional shift, which arises when the underlying data distribution used for model development differs from the data where the model is deployed [34]. For instance, a skin cancer prediction model may perform poorly on dark-skinned patients [3], revealing a distributional shift due to selection bias, which can occur when marginalized populations are not adequately represented in the training data.

If not effectively implemented, evaluated, and regulated, AI solutions in the future may perpetuate and amplify systemic disparities and human biases, contributing to healthcare inequities [19].

Distributional shift can also occur due to minor difference in the radiology equipment in different hospitals [21] resulting in medical images, such as X-rays with slightly different characteristics. Another sub-type of bias is calibration drift, which can occur due to unanticipated changes in clinical practices or patient behaviour [38].

However, bias can also stem from factors beyond the shortcomings of inadequate training data. Such as measurement bias - omitting critical data-fields during model training. For example an algorithm predicting survival of post-menopausal women, that did not perform well, partly because it lacked relevant blood test results [38]. Another source for bias can be incorrect data. This is especially true in cases where data from consumer-facing health apps are merged with clinical data to create predictions. An instance is the Fitbit PurePulse Trackers' unreliable heart rate measurements [13].

Bias is further worsened by the characteristics of data-sets available. Healthcare data is often sparse and imbalanced, for example contain more samples of patients with a mild condition, due to naturally occurring distribution. This is especially prominent in fields such as pathology and mental health [1].

Further data related risks include noise and artefacts in model inputs and hidden confounders. *Noise and artefacts in model inputs* - Noise in data refers to meaningless or irrelevant data, that the model can pick up on [12]. Noisy data is often caused by the differences in or issues with medical equipment used. For example, scanning errors or differences in hospital imagining protocols [25]. Noise can also come from Imaging artifacts and poor imaging quality [25].

Hidden Confounders. are factors unmeasured in the observational data affecting both treatment and outcome. An example of hidden confounders in a clinical

of companies found on the EUDAMED database and manually labeled as AI and secondly using various online databases for companies, such as Danish startup ecosystem database or EIT health database. Experts were found by looking at speakers at relevant conferences, authors of relevant papers or looking at interest groups representing Medical-Device manufacturers. The survey was sent to 130 individuals, including 31 experts, 50 regulators, and 49 individuals from SaMD companies.

4 Findings

4.1 Literature Survey

When applying the literature survey protocol we retrieve a total of 27.073 results. 24 of the originally resulted papers are included in the final literature body. An additional five papers are added by snowballing. The literature survey process can be summarized in Fig. 1. The complete reference list can be found in the Appendix 7. Details of the included papers are also summarised in Fig. 1. The papers covered the time period 2016–2022; they came from the European Union (EU), United Kingdom (UK), United States of America (USA), and Canada; and they were primarily journal articles. Journals published in 2022 were most common.

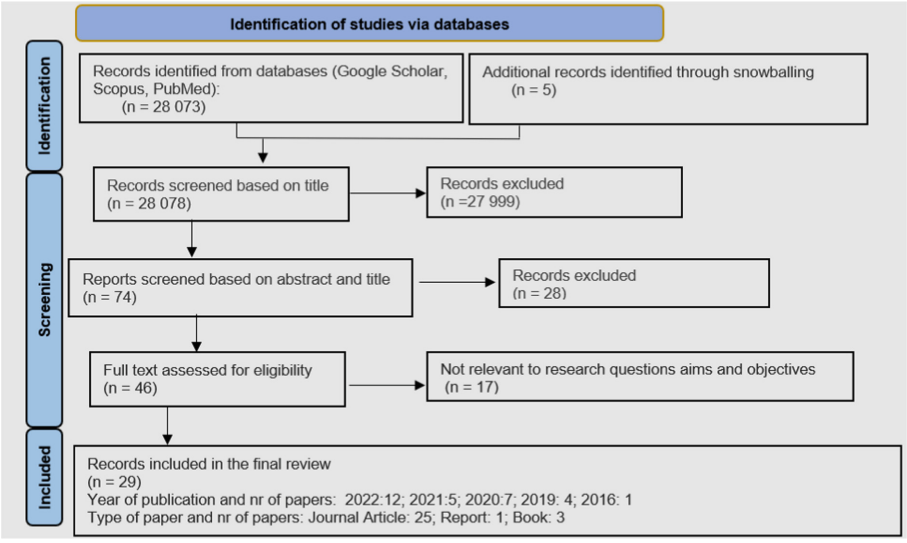


Fig. 1. Process for defining the literature body.

The risks identified in the literature survey can be classified into four categories: data-centric risks, transparency-related risks, cyber-security risks, and

specialties while avoiding excessive granularity for the paper's purpose. Furthermore, the devices are categorized based on the risk categorization principles of the International Medical Device Regulators Forum (IMDRF) set out in Possible Framework for Risk Categorization and Corresponding Considerations [11].

Therefore the devices are further manually labeled across following dimension: the point of healthcare situation or condition the software is intended to be used in; the body part targeted and the medical speciality the device belongs to.

Furthermore, from the point of healthcare situation or condition the software is intended to be used in, the devices were classified in the following categories: critical situation or condition; serious situation or condition; Non-serious situation or condition [11]. Classifying the devices from the point of healthcare situation or condition is done by two labellers - one working in the healthcare sector and another in IT.

3.3 Medical Device Stakeholder Survey

In order to get a more complete view on the area, we conduct a survey of the EU stakeholders in AI-based medical device area. The main focus areas are the current use of AI-based medical devices in EU, potential risk factors and EU legislation regarding Medical Devices. In this survey we intend to validate the findings from the literature survey and rate the risks of using AI-based medical devices. The survey is conducted online with the survey link being sent out to the potential participants. Before launching, the survey is piloted it using a pre-screening [17]. The survey is pre-screened by four researchers with knowledge of both medical and IT field. All lists of options for multiple-choice questions, with the exception of the question about participants role, are randomized to decrease potential measurement errors [17].

The survey has four sections: *Background information* aimed at defining the role of the participant (Expert, Working in SAMD company, Regulator or Other); *AI in the EU market* aimed at defining areas with AI-enabled devices and possible overuse of them; *Safety risks of AI-enabled medical devices* focusing on rating and prioritizing the risks that are noted in the literature; and *Regulation of AI-enabled devices in EU* focusing on the largest technical challenge for ensuring the safety of AI/ML based Medical Devices in EU.

Furthermore the survey collects input in whether the participants felt that the current medical device regulation in EU is sufficient in terms of ensuring the safety of AI/ML based Medical Devices and what changes they would like to see in the medical device regulation in the EU. In this analysis we categorized the participant roles as following: *regulator*, a person working in a AI- enabled SaMD company or an expert in the area. *Expert*, a person who had published research in AI enabled Medical Devices in EU or was or had been part of an expert group or think tank such as the EU expert group on AI. *Regulator*, included in Notified Bodies in the European Union found by looking through the NANDO database for Notified Bodies. SAMD using AI companies were companies which in publicly declared using AI in their devices. Such companies were found firstly by the list

The papers are screened by title, abstract, and full text against the inclusion and exclusions criteria defined. After the papers were selected for inclusion, backwards and forward snowballing is used to find further relevant papers and gather a comprehensive and diverse set of studies that are relevant to the research question being addressed.

For all of the papers reviewed, the following information are extracted: (1)Type of article: journal, conference article or book; (2) Bibliographic data such as publication year; (3) Safety risks listed/discussed; (4) Ways of mitigating the safety risks if they were listed/discussed; (5) Reviewer notes, comments, and recommendations from surveying the article.

Although the current study focuses on medical devices limited in the EU region, geographical limitations were intentionally excluded from the literature survey to ensure an adequate literature body and variability in results.

3.2 EU AI-Based Medical Device Survey Protocol

We survey the approved software medical devices in EU and identify the devices with AI-supported functions. To do so, we survey the European database on medical devices (EUDAMED). We extract[3] all software medical devices and collect (a) Trade name, (b) Manufacturer and (c) Classification (risk class).

Having collected the initial device body, validated and cleaned the data, we process it as following. In this study we focus on medium to high risk devices, thus we exclude the low risk devices (class I and Class A) from the dataset. This group is chosen for exclusion, since the devices are subjected to a different, less rigorous approval process.

Furthermore, EUDAMED does not currently provide information on the description of the device, including whether a device is using AI or not. Thus, the resulting data are manually annotated as either AI and non-AI for filtering out non-AI devices. To validate the device data, we follow a three-source approach: FDA list of AI/ML devices, AI for Radiology database, and device publicly available data. As the first step of annotation the devices are cross-referenced with our first two sources: the FDA list of Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices and the AI for Radiology database. AI for Radiology database is a database of CE-marked AI software products for clinical radiology based on vendor-supplied product specifications [18]. The FDA list of AI-enabled Medical Devices is a non-exhaustive list periodically updated by FDA based on publicly available information [6]. The rest of the devices are then manually labeled based on the publicly available data on Google. In this step, we use the device manufactures websites and press releases as primary data sources.

The resulting dataset is categorized by medical specialties using a modified version of the European Union of Medical Specialists' list, created in collaboration with two medical experts. The modified list aims to include all relevant

[3] We apply the Python library BeautifulSoup with extraction date 2022.09.29.

overview of medical AI devices approved by the US Food and Drug Administration, that indicated that evaluation process can mask vulnerabilities of devices when they are deployed on patients. Muehlematter et al. [28] report on a comparative study of Medical Devices approved by FDA and CE-marked in EU between years 2015–2020.

While many papers have investigated the safety risks of AI-based medical devices [15,29,36], we were not able to find a dedicated literature survey of the safety risks of AI-based medical devices.

3 Methodological Approach

In this study we apply a combination of quantitative and qualitative approaches to present multiple findings about AI-based Medical Devices. This mixed approach is chosen to enable triangulation in order to examine the current use and potential safety risks of AI-enabled Software as a Medical Device and from research literature, devices on the EU market and practitioner's viewpoint. The approach applied is: (a) we review the literature of safety risks associated with AI-enabled medical devices; (b) we analyze the current AI-enabled Medical Devices on the EU market achieved by the collection and manual labelling of data from the European Database of Medical Devices EUDAMED; and (c) we survey the stakeholders of the ecosystem of the European medical devices.

3.1 Literature Survey of Risks of AI-Enabled SaMD

We conduct a literature survey on the risk of AI-enabled software as a medical device (SaMD)[1]. We define a protocol based on the PRISMA methodology [30] and by leveraging our previous experience on literature surveys and systematic literature reviews [22–24,39]. Our survey protocol includes:

Sources. The defined literature sources are: (i) Google Scholar, (ii) PubMed, and (iii) Scopus.

Search string. *((safety) OR (risks)) AND (healthcare) AND (((machine learning)) OR (deep learning)) OR (artificial intelligence)*[2].

Inclusion/exclusion criteria. In order for the paper to be included in the Literature Review the following criteria has to be met: 1) Paper discusses the safety risks of AI enabled devices in medicine; 2) The paper is from the time period 2012–2023; 3) The paper is in English;

[1] SaMD is defined by the International Medical Device Regulators Forum (IMDRF) as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device.".

[2] Further variation of search keywords were tested that included, among other, "AI ML & Safety & Medical Device & Medicine"; "ML & Safety Risks & Medical Device & Healthcare".

device area [32]. AI-based medical device software holds great promise in addressing the challenges faced by healthcare systems in the European Union, such as the aging population, inefficient medical systems, and lack of healthcare workers. However, these AI-enabled solutions also come with risks, from inaccurate predictions to incorporating various biases. These issues raise concerns about safety risks, which can consequently lead to a lack of trust and pose a barrier to the wide-scale adoption of AI into clinical practices. Lack of information about these devices and mitigation of various risks further decreases trust. In the EU context various aspects of AI- enabled medical devices, such as their characteristics, are unexplored. This paper aims to provide an overview of risks associated with AI-based medical devices and describing the AI-based medical software devices currently on the EU market, with focus on factors affecting their safety. The core questions explored in this paper are:

1. What are the safety risks of AI-enabled medical Devices, and what strategies exist to mitigate them?
   Extensive focus has been put into creating frameworks for evaluating AI-based Medical Devices [5]. However, to the best of the author's knowledge, no survey of the safety risk of such devices has so far been conducted.
2. What kind of AI-based Medical Devices can be currently found on the EU market?
   There has been a lot of discussions and work put into regulating AI in the European Union. However, in comparison to the USA regulatory body (FDA), EU is lagging behind in terms of providing information about AI-enabled Medical Devices. Indeed, until the launch of EUDAMED there lacked a central database of Medical Devices on the EU market.
3. How do the stakeholders of the AI-enabled Medical Device ecosystem perceive the use, risks and regulation of AI-enabled Medical Devices?
   Analysis on stakeholders' perception of Medical AI is so far largely focused on healthcare specialists and their views on AI [40,43]. However, little is known how companies, researchers and regulators perceive the current use of AI various safety risks of AI-enabled Medical Devices and the regulation on Medical-AI in EU.

The rest of the paper is structured as follows: in Sect. 2, we analyze the background and related works; in Sect. 3, we provide a brief overview of the regulation of medical devices in EU; in Sect. 4, we explain the methodologies used in various research steps; in Sect. 5, we present the findings of the research and in Sect. 6, we provide an analysis of the research findings. Following that, we provide a discussion section, where we delve into the implications of our findings.

This paper aims to contribute to the current discussions of legislative and regulatory reforms intended to regulate AI/ML-based medical devices.

2 Background and Related Work

Most studies on AI-based medical devices focus on the US market and on devices approved by the FDA. For example, Wu et al. [41] published a comprehensive

Investigating AI in Medical Devices: The Need for Better Establishment of Risk-Assessment and Regulatory Foundations

Sandra Baum[1] and Konstantinos Manikas[1,2(✉)]

[1] Computer Science Department, IT -University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark
{sanb,koma}@itu.dk
[2] Accenture Consulting, Bohrsgade 35, 1799 Copenhagen V, Denmark

Abstract. Artificial intelligence (AI) has the potential to revolutionize healthcare in the EU by addressing challenges, such as shortages of healthcare personnel and more effective diagnosis and care. However, the safety concerns surrounding AI-based medical devices have been a major roadblock to the technology's wider adoption. This study aims to further investigate these concerns in the European context by analysing the AI-enabled Medical devices currently available in the European Union market along with their potential safety risks. We do this by applying a combination of three research methods: (1) a survey of the safety risks of AI-enabled Medical Devices published between 2012 and 2023, (2) an analysis of AI-based medical devices in the EUDAMED database, and (3) a survey on the perceptions of the EU Medical AI ecosystem stakeholders. Our study analyzed the state-of-the-art with a literature body of 29 papers and summarized a number of risks related to the use of AI in medical devices along with the reported mitigation strategies. Furthermore, we analyzed the approved medical devices (71 devices) that use AI in the EUDAMED database and found that there is a lack of transparency in whether the devices use AI along with the lack of crucial information necessary to asses the devices' safety risks, such information on training data. Finally, when we survey a number of medical device stakeholders (7 out of 130 respondents) we find that there is a disconnect between the industry and regulators: the medical device representatives emphasize the need for better guidance on post-market surveillance while the regulation representatives feel that they lack expertise in AI.

Keywords: Artificial intelligence · Medical device regulation · Literature survey · Medical device survey

1 Introduction

Artificial intelligence (AI) solutions, like ChatGPT are increasingly entering various aspects of our lives. This tendency is arguably also occurring in the medical

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 51–69, 2024.
https://doi.org/10.1007/978-3-031-59717-6_4

Privacy, Ethics and Regulations

33. Tran, Y., Wijesuriya, N., Tarvainen, M., Karjalainen, P., Craig, A.: The relationship between spectral changes in heart rate variability and fatigue. J. Psychophysiol. **23**(3), 143–151 (2009). https://doi.org/10.1027/0269-8803.23.3.143
34. Vanneste, P., et al.: Towards measuring cognitive load through multimodal physiological data. Cogn. Technol. Work **23**(3), 567–585 (2021). https://doi.org/10.1007/s10111-020-00641-0
35. Visnovcova, Z., Mestanik, M., Gala, M., Mestanikova, A., Tonhajzerova, I.: The complexity of electrodermal activity is altered in mental cognitive stressors. Comput. Biol. Med. **79**, 123–129 (2016). https://doi.org/10.1016/j.compbiomed.2016.10.014
36. Volden, F., De Alwis Edirisinghe, V., Fostervold, K.-I.: Human gaze-parameters as an indicator of mental workload. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (eds.) IEA 2018. AISC, vol. 827, pp. 209–215. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-96059-3_23
37. Wu, C., Liu, Y., Guo, X., Zhu, T., Bao, Z.: Enhancing the feasibility of cognitive load recognition in remote learning using physiological measures and an adaptive feature recalibration convolutional neural network. Med. Biol. Eng. Comput. **60**(12), 3447–3460 (2022). https://doi.org/10.1007/s11517-022-02670-5
38. Xu, J., Wang, Y., Chen, F., Choi, E.: Pupillary response based cognitive workload measurement under luminance changes. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS, vol. 6947, pp. 178–185. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23771-3_14
39. Xu, Q., Nwe, T.L., Guan, C.: Cluster-based analysis for personalized stress evaluation using physiological signals. IEEE J. Biomed. Health Inform. **19**(1), 275–281 (2015). https://doi.org/10.1109/JBHI.2014.2311044
40. Yeragani, V.K., Krishnan, S., Engels, H.J., Gretebeck, R.: Effects of caffeine on linear and nonlinear measures of heart rate variability before and after exercise. Depress. Anxiety **21**(3), 130–134 (2005). https://doi.org/10.1002/da.20061
41. Yüce, A., Gao, H., Cuendet, G.L., Thiran, J.P.: Action units and their cross-correlations for prediction of cognitive load during driving. IEEE Trans. Affect. Comput. **8**(2), 161–175 (2017). https://doi.org/10.1109/TAFFC.2016.2584042

19. Li, Y., Li, K., Wang, S., Li, Y., Chen, J., Wen, D.: Towards safer flights: a multimodality fusion technology-based cognitive load recognition framework. In: 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), pp. 525–530 (2022). https://doi.org/10.1109/ICCASIT55263.2022.9986937
20. Luck, S.J.: An Introduction to the Event-Related Potential Technique. MIT Press, Cambridge (2014)
21. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017). https://doi.org/10.5555/3295222.3295230
22. Makowski, D., et al.: NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav. Res. Methods **53**(4), 1689–1696 (2021). https://doi.org/10.3758/s13428-020-01516-y
23. Oppelt, M.P., et al.: Adabase: a multimodal dataset for cognitive load estimation. Sensors **23**(1) (2023). https://doi.org/10.3390/s23010340
24. Orru, G., Longo, L.: The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2018. CCIS, vol. 1012, pp. 23–48. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14273-5_3
25. Pejović, V., Matkovič, T., Ciglarič, M.: Wireless ranging for contactless cognitive load inference in ubiquitous computing. Int. J. Hum.-Comput. Interact. **37**(19), 1849–1873 (2021). https://doi.org/10.1080/10447318.2021.1913860
26. Prajod, P., André, E.: On the generalizability of ECG-based stress detection models. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 549–554 (2022). https://doi.org/10.1109/ICMLA55696.2022.00090
27. Saganowski, S., Kunc, D., Perz, B., Komoszyńska, J., Behnke, M., Kazienko, P.: The cold start problem and per-group personalization in real-life emotion recognition with wearables. In: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 812–817 (2022). https://doi.org/10.1109/PerComWorkshops53856.2022.9767233
28. Solhjoo, S., et al.: Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. Sci. Rep. **9**(1), 14668 (2019). https://doi.org/10.1038/s41598-019-50280-3
29. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 421425 (2009). https://doi.org/10.1155/2009/421425
30. Sweller, J., Ayres, P., Kalyuga, S.: Measuring Cognitive Load, pp. 71–85. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-8126-4_6
31. Tervonen, J., Nath, R.K., Pettersson, K., Närväinen, J., Mäntyjärvi, J.: Cold-start model adaptation: evaluation of short baseline calibration. In: Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2023 ACM International Symposium on Wearable Computing. UbiComp/ISWC '23 Adjunct, pp. 417–422. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3594739.3610731
32. Tervonen, J., Pettersson, K., Mäntyjärvi, J.: Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors. Electronics **10**(5) (2021). https://doi.org/10.3390/electronics10050613

4. Champseix, R.: Heart Rate Variability analysis (2018). https://github.com/Aura-healthcare/hrv-analysis. Accessed 20 June 2023
5. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 19, pp. 785–794. ACM, New York, NY, USA, August 2016. https://doi.org/10.1145/2939672.2939785, https://dl.acm.org/doi/10.1145/2939672.2939785
6. Dalmaijer, E.S., Mathôt, S., Van der Stigchel, S.: Pygaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. Behav. Res. Methods **46**(4), 913–921 (2014). https://doi.org/10.3758/s13428-013-0422-2
7. Delliaux, S., Delaforge, A., Deharo, J.C., Chaumet, G.: Mental workload alters heart rate variability, lowering non-linear dynamics. Front. Physiol. **10** (2019). https://doi.org/10.3389/fphys.2019.00565
8. Ehrmann, D.E., et al.: Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nat. Med. **28**(7), 1331–1333 (2022). https://doi.org/10.1038/s41591-022-01833-z
9. Feradov, F., Ganchev, T., Markova, V.: Automated detection of cognitive load from peripheral physiological signals based on Hjorth's parameters. In: 2020 International Conference on Biomedical Innovations and Applications (BIA), pp. 85–88 (2020). https://doi.org/10.1109/BIA50171.2020.9244287
10. Gjoreski, M., et al.: Datasets for cognitive load inference using wearable sensors and psychological traits. Appl. Sci. **10**(11) (2020). https://doi.org/10.3390/app10113843
11. Gjoreski, M., et al.: Cognitive load monitoring with wearables-lessons learned from a machine learning challenge. IEEE Access **9**, 103325–103336 (2021). https://doi.org/10.1109/ACCESS.2021.3093216
12. Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J.M., Van den Bergh, O.: Respiratory changes in response to cognitive load: a systematic review. Neural Plast. **2016**, 8146809 (2016). https://doi.org/10.1155/2016/8146809
13. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Advances in Psychology, vol. 52, pp. 139–183. North-Holland (1988). https://doi.org/10.1016/S0166-4115(08)62386-9
14. Herbig, N., et al.: Investigating multi-modal measures for cognitive load detection in e-learning. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20, pp. 88–97. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340631.3394861
15. Herbig, N., Pal, S., Vela, M., Krüger, A., van Genabith, J.: Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. Mach. Transl. **33**(1), 91–115 (2019). https://doi.org/10.1007/s10590-019-09227-8
16. Hussain, M.S., Calvo, R.A., Chen, F.: Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. Interact. Comput. **26**(3), 256–268 (2013). https://doi.org/10.1093/iwc/iwt032
17. Jiménez-Guarneros, M., Gómez-Gil, P.: Custom domain adaptation: a new method for cross-subject, EEG-based cognitive load recognition. IEEE Sig. Process. Lett. **27**, 750–754 (2020). https://doi.org/10.1109/LSP.2020.2989663
18. Khanam, F., Hossain, A.A., Ahmad, M.: Electroencephalogram-based cognitive load level classification using wavelet decomposition and support vector machine. Brain-Comput. Interfaces **10**(1), 1–15 (2023). https://doi.org/10.1080/2326263X.2022.2109855

RMSSD and maximum amplitude of EMG, was more convoluted. Here, too, there may be some individual differences. Indeed, several features are clustered around zero, meaning that for those observations the current feature had a small impact, and long tail(s) denoting observations for which the current feature had a larger impact. However, the coloring of e.g. RMSSD and EMG_max_amplitude show that higher values were related to both decreased and increased cognitive load. These may be related to some spurious events during the completion of the task, changing of cognitive load during the two-minute window, or the relation between self-reports and some physiological parameters may be unsystematic.

Investigation and understanding this relationship and finding features with systematic behavior under varying cognitive load is a necessity for robust cognitive load detection. To keep focus on user calibration and normalization options, this analysis is left for future work, together with finding the best type of a model, hyperparameter optimization, feature window duration optimization, feature selection, and signal modality selection, all of which are important steps to consider in model development. Based on Fig. 4, the developed model fit to the data reasonably well, but the described steps may help in improving the model performance.

6 Conclusions

Overcoming the cold-start situation in model personalization is a necessity for future human state detection applications. In this study, using short baseline measurements to normalize features was proposed as the solution in detecting continuous cognitive load. The experiments showed that user calibration always performed better than the general model but worse than a model with participant-wise normalization with full dataset. The optimal baseline duration was found to be 3–3.5 min and there were no differences between the tested normalization functions. A SHAP feature importance analysis revealed that the developed model found physiologically correct patterns, increasing trust to it. Future studies are needed for different mental states to further validate the proposed user calibration approach.

References

1. Albaladejo-González, M., Ruipérez-Valiente, J.A., Gómez Mármol, F.: Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. J. Ambient. Intell. Humaniz. Comput. (2022). https://doi.org/10.1007/s12652-022-04365-z
2. Biondi, F.N., Cacanindin, A., Douglas, C., Cort, J.: Overloaded and at work: investigating the effect of cognitive workload on assembly task performance. Hum. Factors **63**(5), 813–820 (2021). https://doi.org/10.1177/0018720820929928
3. Bozkir, E., Geisler, D., Kasneci, E.: Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 1834–1837 (2019). https://doi.org/10.1109/VR.2019.8797758

load participant-wise, as was done in the original paper [23], but not here since it would be unfeasible in a cold-start scenario.

Curiously, the general model performed the same as predicting the average score. Since the general model has no knowledge of any of the individual components, this further highlights the need for personalization especially in the cold-start case: if there was no prior information from a new individual, using the general, non-personalized model for them would be as good as guessing the average score.

Obtaining this prior information is a crucial step in cold-start model personalization for mental state estimation: without some, one cannot personalize. Having the people rest for a few minutes, as in this study, is one of the less burdensome techniques, but it contains no information about how people might react to or perceive cognitively loading stimuli. While it is already a more tiring option, conducting a mini-protocol of short cognitive tasks might be an alternative to include reactions and variation in the baseline data. However, any baseline measurement should be repeated periodically since e.g. caffeine intake [40] or fatigue [33] may cause changes to human physiology. Therefore, one could alternatively opt for collaborative filtering [29] based methods by e.g. measuring user similarity via background questionnaire, such as demographics or personality traits. Although the big five personality traits did not improve classification results in [23], they could still be useful in cold-start personalization.

Due to subjectiveness, determining the ground truth in detection of cognitive load and other mental states is also a challenge of its own. The current study used non-normalized cognitive load self-reports as labels for a regression model. One could also choose to normalize the labels participant-wise (not feasible in a cold-start scenario) or choose to classify between self-reported low/medium/high load or even classify based on task labels. The latest would be an objective measure when looking at the data labelling, but still physiology would reflect subjective load and some people might not experience cognitive load in tasks labelled under high load. Ultimately the choice of the ground truth comes down to the targeted use case: what is most sensible given the context? The target used in the current study is suitable for applications where the interest is in subtle subjective changes in cognitive load. A continuous target and regression analysis applies to a more (time-wise) continuous modelling and allows developing methods to detect moments when the person is just heading towards cognitive overload, unlike classification, which provides a more definite outcome.

The dataset used in this study contained a simulated driving task and the n-back task conducted in a controlled laboratory environment. Since this was the first inspection of cold-start model personalization in cognitive load detection, such dataset with clear and likely high cognitive load periods was chosen to be able to focus on the cold-start issue. Since the results were encouraging, the method should be evaluated next on different datasets from the health domain to improve its ecological validity.

While some features showed a rather clear behavior pattern in Fig. 5, like pupil diameter and skin temperature, the behavior of other features, such as

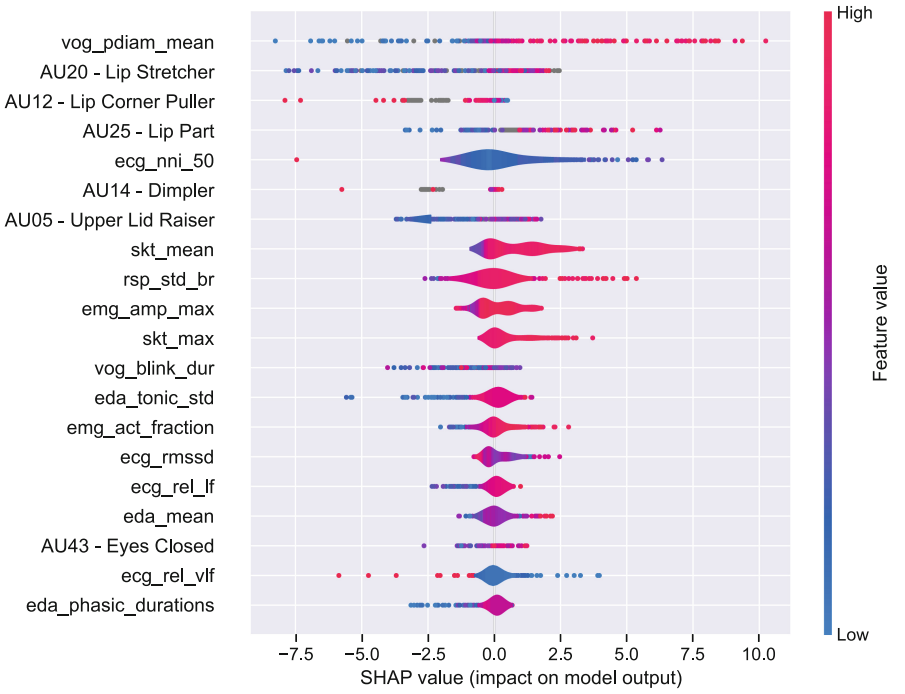


Fig. 5. Beeswarm plot of the SHAP values of the features in the user calibration model with 3.5 min baseline with features normalized with averaging.

5 Discussion

The presented analysis proved the usefulness of user calibration, since it always outperformed the general model regardless of the choice of the normalization function. Thus, collecting 3–3.5 min of baseline resting data from a new user allows making better predictions since the beginning. The proposed approach has some limitations which form the ground for future work: it only captures individual differences at the level of basic physiology, determining the correct ground truth is challenging, the used dataset was collected in a constrained environment with a large set of physiological signals, and only one modelling approach was evaluated. Each of these points are further discussed next.

The data collection protocol exhibited individual variations in three levels: basic physiology, reactions to stimuli, and self-reporting. The basic physiology between the participants differed, which was up to some extent caught by user calibration. Different people react in different ways to similar external stimuli, which in turn was caught by participant-wise normalization. Finally, different people may experience the same task in various ways: some see the task demanding while others find it enjoyable, which is reflected in the subjective reports. Such information could be included in the model by normalizing the reported

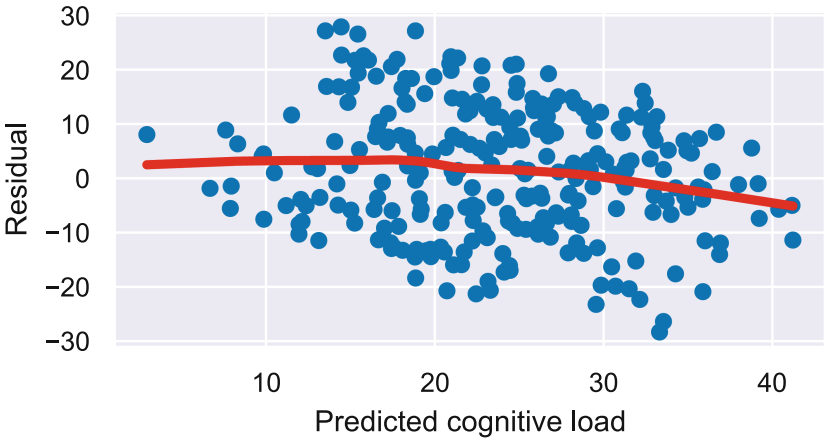


Fig. 4. Scatter plot of model residuals vs. predictions with a LOWESS trend curve, produced over test data folds of the best performing user calibration model with 3.5 min baseline with features normalized with averaging.

when predicting the load. The best performance was observed with 3 min baseline duration with standard scaling, and 3.5 min with averaging and min-max normalization. However, the differences between the different durations were not large.

According to the residual plot in Fig. 4, there were no clear signs of heteroscedasticity or outliers. The trend curve shows slight elevation at lower cognitive loads, and a minor decreasing trend towards the higher predicted cognitive load estimates. Thus, the model may overestimate lower cognitive load and underestimate higher one, but based on the figure the effect should not be large.

Shapley additive explanations (SHAP) [21] were computed to assess the importance of different signal modalities and features. Figure 5 displays a beeswarm plot of the top-20 features with highest average absolute SHAP values in a decreasing order, drawn over the best performing user calibration model. Judging by the number of features from different modalities in the plot, the most influential signal modalities were facial activations and ECG, and each signal modality had at least one feature within the top-20.

Although the direction of changes is a little confused for some features, certain conclusions can be made from the figure. The most influential feature was mean pupil diameter, with its higher values corresponding to higher cognitive load, and vice versa for lower values. Lower/higher values in AU20 (lip stretcher), AU25 (lip part), and AU05 (upper lid raiser) corresponded to lower/higher cognitive load, respectively. Moreover, higher skin temperature, variation in breathing rate, blink duration, trapezoidal EMG activity, and electrodermal activity, and lower heart rate variability all corresponded to higher cognitive load, according to the developed model. These are roughly in line with previous observations [2,7,35,36] but there was mixed evidence of skin temperature and respiration changes under cognitive load [10,12,34].

Table 2. Regression results of predicting subjective cognitive load. Dur stands for best calibration duration in minutes.

Normalization	Personalization	Dur	MAE	MSE	RMSE
Standard scaling	Participant-wise	–	8.20 (1.12)	107.57 (31.86)	10.27 (1.47)
	Calibration	3.0	10.09 (1.69)	148.26 (51.37)	12.01 (2.02)
Averaging	Participant-wise	–	9.93 (1.34)	147.27 (42.49)	12.01 (1.72)
	Calibration	3.5	9.72 (1.54)	136.89 (35.87)	11.60 (1.56)
Min-Max	Participant-wise	–	8.57 (1.75)	114.74 (41.81)	10.52 (2.02)
	Calibration	3.5	9.78 (1.74)	146.99 (47.09)	11.95 (2.05)
No scaling	General model	–	11.03 (1.80)	175.80 (51.65)	13.11 (2.01)
Average prediction	General model	–	10.99 (1.38)	160.78 (34.89)	12.60 (1.42)

Table 3. Related samples T-test results of comparing the MAE's of the user calibration model to those of other models'. P-values were corrected with the Benjamini-Hochberg method.

Normalization	Participant-wise		No scaling		Average prediction	
	T	p	T	p	T	p
Standard scaling	−3.08	0.020	2.49	0.039	−21.64	<0.001
Averaging	0.52	0.618	3.69	0.011	−22.08	<0.001
Min-Max	−2.74	0.029	3.49	0.012	−20.67	<0.001

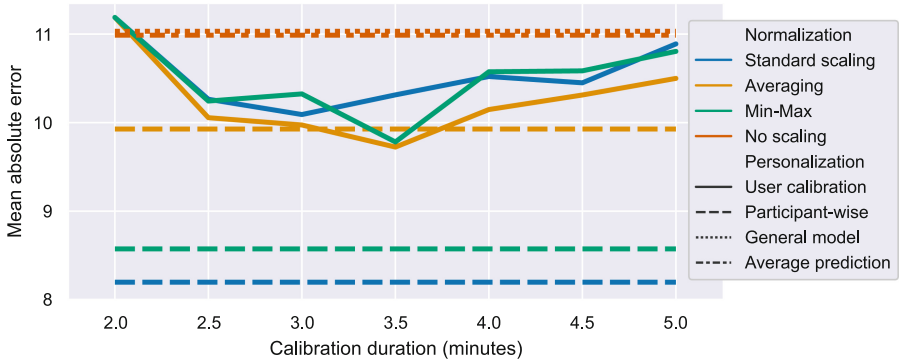


Fig. 3. Model performance at different durations of baseline measurement. The line colors refer to the normalization approach and line style to personalization approach.

three normalization functions when using user calibration were observed (test results not shown). The non-personalized model performed similarly to average prediction.

Figure 3 shows the MAE of user calibration with different baseline durations: a MAE of e.g. 10 would denote that an error of 10 units was made on average

3.4 Experimental Protocol

The extracted features were used to detect cognitive load as a continuous variable using extreme gradient boosting regressor (XGBoost) [5]. The regressor was selected since extreme gradient boosting has been shown to have good performance in different domains with tabular data, and since it natively handles missing data which is prevalent in the current dataset due to some participants opting out from facial data collection.

The k-drive tasks had a duration of about five minutes and so the experienced cognitive load may have varied over the course of the task. Still, subjective ratings were given only after the task. To ensure that the physiological data is timely and best reflects the given rating, only the last two minute window from each task was used for training the model. This choice has also a balancing effect between the two task types, as the n-back tasks lasted for about two minutes.

Following the original paper [23], the adopted cross-validation strategy was leave-three-users-out, resulting in 10 folds total. For each fold, the data of three randomly selected participants was left out and the model was trained with the remaining participants data and tested on the left out fold. The model performance was assessed in terms of mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). The performance was compared against a baseline of predicting the average cognitive load for each observation.

Three modelling tasks were defined: i) without feature normalization; ii) with participant-wise feature normalization; and iii) user calibration. The evaluation without feature normalization sets a baseline which user calibration should try and exceed to be useful, and participant-wise feature normalization sets an upper limit of what to expect with normalization-based personalization approaches.

In addition, the duration of the needed baseline measurement was evaluated by training the user calibration model with varying duration of baseline data, ranging from two to five minutes in 30 s increments. Shorter than two minute calibration was not considered since the used feature window length was two minutes. Moreover, three different normalization functions as specified in Sect. 3.1 were experimented with.

The statistical significance of the differences between the personalization and normalization approaches were assessed with related samples T-test, since the cross-validation errors were found to be normally distributed (Shapiro-Wilk test). P-values were corrected with the Benjamini-Hochberg method to adjust the false discovery rate for multiple testing.

4 Results

User calibration performed better than the general and the average prediction model with each normalization function tested (see Table 2) and the difference was statistically significant (Table 3). As expected, participant-wise scaling performed better than user calibration when features were normalized with standard scaling or min-max normalisation, but the two approaches performed the same when features were averaged. No statistically significant differences between the

Cognitive load was assessed after each phase of the protocol with the NASA-TLX self-report questionnaire, assessing mental, physical and temporal demand, performance, effort and frustration. Each dimension has a weighting factor to compute the final score as a sum of weighted self-report components. This final score serves as the metric for cognitive load in this study. The authors in the original paper [23] suggest transforming the score individually to a value between 0 and 1 through min-max normalisation. Since this transformation is impossible in a cold-start scenario, it was decided to opt for the unscaled metric in this study.

3.3 Data Processing

Features were extracted from each of the available data sources except for the PPG signal, which was thought redundant since ECG was available. Following the original paper [23], features were extracted with a sliding window of two minutes with a five second window slide. The physiological signals were mostly processed using the NeuroKit2 software package [22], but the heart rate variability features were extracted with the hrv-analysis library [4] and saccades, fixations, and blinks were detected from the VOG data with the PyGaze library [6]. The sum of frames with each facial activation was used as the features for the facial data; see [23] for a description of the activation units. The rest of the extracted features are listed in Table 1.

Table 1. Extracted features from each signal.

Signal	Extracted features
ECG	mean_HR, std_HR, mean_nni, sdnn, nni50, pnni50, rmssd, vlf_power, lf_power, hf_power, lf/hf ratio, total power, lfnu, hfnu, vlf_relative_power, lf_relative_power, hf_relative_power
EMG	rms, n_onsets, fraction_high_activity, max_amplitude
EDA	eda_mean, eda_std, eda_min, eda_max,eda_ slope, eda_range, tonic_mean, tonic_std, tonic_correlation_with_time, phasic_n_peaks, phasic_peak_amplitude, phasic_peak_duration, phasic_peak_area
RSP	breathing_rate_mean, breathing_rate_std, phase_ratio_mean
SKT	skt_mean, skt_std, skt_min, skt_max, skt_slope
VOG	pupil_diam_mean, pupil_diam_std, pupil_diam_slope, blink_rate, blink_duration, time_between_blinks, fix_rate, fix_duration, time_between_fix, n_fix_with_dur_>100ms, n_fix_with_dur_66-150ms, n_fix_with_dur_300-500ms, n_fix_with_dur_>1000ms, sac_rate, sac_duration, time_between_sac, sac_amplitude

ECG = electrocardiogram, EMG = electromyogram, EDA = electrodermal activity, RSP = respiration, SKT = skin temperature, VOG = video-oculography, diam = diameter, fix = fixations, sac = saccades

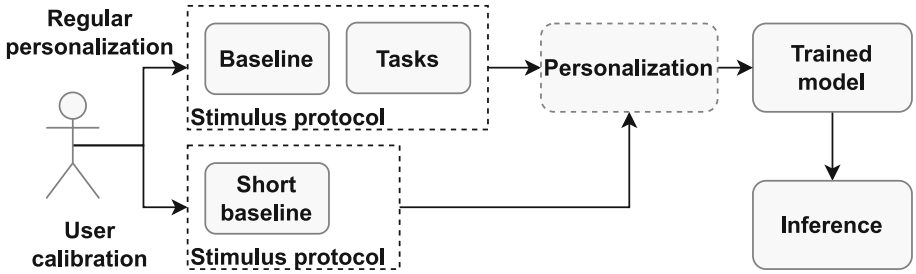


Fig. 2. The cold-start process for mental state estimation of a new user in inference mode. The boxes tagged "Stimulus protocol" display the tasks to be completed before the trained model can be personalized and applied for inferring the state of the user.

3.2 Dataset

The ADABase dataset [23] was adopted for the cognitive load detection experiments. The dataset consists of two tasks aimed at inducing cognitive load: the n-back task and a simulated driving task called k-drive. Simulators and standard cognitive tests provide a controllable and quantifiable environment and thus the induced cognitive states are more homogeneous. Thus, the dataset comprises a good basis to study algorithm development for the cold-start case.

An n-back task consists of a sequence of stimuli and the study participant must indicate when the current stimulus matches the one presented n steps earlier. The n-back task conducted in ADABase consisted of a single (visual stimulus) and a dual stimuli (visual and auditive stimuli) test with three difficulty levels (i.e. $n \in \{1, 2, 3\}$). The k-drive task consisted of watching an autonomous simulator playing a driving game and indicating, on three difficulty levels, whether the car was 1) passing another car, 2) being overtaken, or 3) accelerating or decelerating rapidly; each level was incremental and in each subsequent level the participant had to indicate the events of the previous level(s) as well. Additionally, the participant solved a secondary task of searching and adding songs to a playlist during levels 2 and 3.

The test participants' physiology was monitored with a Biopac MP160 system measuring electrocardiogram (ECG), electromyogram (EMG, trapezius muscle), electrodermal activity (EDA), respiration (RSP), skin temperature (SKT) and photoplethysmogram (PPG). In addition, video-oculography (VOG) was recorded with Tobii Pro Fusion and facial cues with a BASLER camera.

The published version of dataset contains 30 participants, 12 of whom refused the collection of facial video data. The order of n-back and k-drive was randomized, and in the public version 18 participants first completed the n-back task. The measurement set-up was the same for all participants, but it was adjusted for handedness and the time of day varied. The baseline measurement used in this study for user calibration is taken from the resting baseline that occurred before the first stimulus, whether it was n-back or k-drive.

personalization approaches include e.g. custom domain adaptation [17], where a transfer learning approach adapts the neural network to each individual. In a related context of stress detection, users have been clustered first to train the model based on similar users' data [39].

Each of these approaches require a substantial amount of data from each user, most of them a full set similar to that of the users in the training data. In a cold-start case, such data is not available, and the developed model is unapplicable for new users.

To the best of our knowledge, no earlier studies exist in cognitive load detection addressing this challenge. However, in emotion recognition, Saganowski *et al.* [27] proposed to apply transfer learning with accumulating data in real-life scenario, and in stress detection, participant-wise feature normalization has been implemented with just the baseline data [1,26]. A similar approach for stress and affect detection was examined in [31] who also analyzed the duration of baseline measurement needed. In neuroscientific research, eliminating the individual variations with baseline data is a standard procedure [20].

To set the current study apart from related work the following differences are outlined: i) continuous cognitive load is detected with a regression model as opposed to classification, ii) cold-start is addressed by personalizing based on short baseline measurement, iii) different baseline durations and normalization functions are evaluated.

3 Methodology

3.1 User Calibration

In general terms, a normalization function transforms given input data according to some normalization parameters into a representation that is better suited for a machine learning model. The normalization parameters should be computed from the training and applied for the testing data but the same parameters are used for both splits of the dataset.

In participant-wise scaling, the normalization parameters are computed separately for each participant from a full measurement protocol. Instead, the parameters could be computed from a short baseline measurement, called user calibration in this study, and applied for all subsequent data from that person. In a real-life use case, a new person should sit still and relax for a couple of minutes, allowing the collection of baseline data, which is a much less burdening option than completing the whole protocol. Figure 2 highlights the differences in the inference process between regular personalization and user calibration.

Three normalization functions are applied in this study: averaging $X_{avg} = X - mean(X)$, standard scaling $X_{std} = X_{avg}/std(X)$ and min-max transformation $X_{minmax} = \frac{X-min(X)}{max(X)-min(X)}$, for a dataset X. In participant-wise scaling, the normalization parameters (mean, std, min, max) are computed separately for each participant across the whole measurement protocol, and in user calibration they are computed separately for each participant from short baseline measurement of varying duration.

2 Related Work

2.1 Methods in Cognitive Load Detection

Cognitive load or mental workload refers to the amount of mental resources used to perform a task [24]. Several approaches to measure cognitive load as a continuous variable exist, like self-report questionnaires (e.g. the NASA-TLX [13]), performance measures of cognivitely demanding tasks [30], and physiological triggers, since cognitive load is reflected to e.g. pupillary responses [36,38], heart rate variability [7,28], electrodermal activity [34,35], and facial expressions [16,41]. The link between cognitive load and physiology has led to the development of automated tools to detect cognitive load based on (wearable) sensor data. Figure 1 shows a general machine learning pipeline for the detection task based on sensor data processing.

Previous works attempting cognitive load detection with machine learning methods have primarily focused on a classification setup. Most works have detected high cognitive load from low or no load [3,9,32,37] while some have had three or more levels based on estimated difficulty of the task [18,19]. Although cognitive load can be considered a continuous measure and treating it as a continuous variable allows for a more fine grained analysis, few works attempt to detect it with a regression model. Herbig *et al.* [14,15] have recognized cognitive load in an e-learning and a machine translation task. Pejović *et al.* [25] developed a non-contact sensor to detect cognitive load during elementary cognitive tasks. Lastly, Oppelt *et al.* [23], who introduced the dataset used for experiments in this study, presented results also for regression modelling. Each work selected self-reported cognitive load as the regression target.

2.2 Cold-Start Model Personalization

Baseline physiology, physiological reactions and task perception are individual, which calls for model personalization when detecting cognitive load. Still, several previous studies aim for fully person-independent detection [3,9,19]. When personalization was considered, the most prevalent approach has been some version of participant-wise feature normalization, used in e.g. [14,15,25,32], which normalizes features separately for each person using their full dataset. In addition, it was the only personalization approach considered in a contest to detect cognitive load from wearable sensor data, applied by 5 teams from 12 [11]. Other

Fig. 1. A machine learning pipeline generally used for training a state detection model. Personalization is not always included, which is depicted with a dashed box and it can overlap with either preprocessing or model training.

aiming to improve user health, wellbeing, and performance. For example, a virtual physiotherapist could detect person's activities to assist in physical rehabilitation, detected stress or emotional state could trigger interaction in mental coaching or when recovering from a trauma, measuring alertness or drowsiness with interventions could help if a person has trouble sleeping, and the detected states could be used in clinical decision support as additional information.

To unlock the full potential of these applications, they should work automatically, close to real-time, and adapt to each individual, even new, unknown ones. One major drawback in current state detection approaches is that they fail to properly account for individual differences especially in a cold-start scenario. Basic physiology, reactions to external stimuli and perceptions of varying situations are individual-specific, which should be accounted for in the modelling procedure: the detection model should be personalized. Typically, the physiological features used for state detection are normalized participant-wise, using a whole dataset from each person to do so. This solution accounts for the differences in individual baselines and individual reactions to the different stimuli. However, applying it requires a complete set of data from each participant before the developed model can be applied for them. When a new user starts using the system, i.e. a cold-start occurs, completing a lengthy calibration protocol with different stimuli is burdening for them and may lead to demotivation and even giving up with the system before even properly beginning.

The current study investigates the cold-start problem in the context of cognitive load detection. Monitoring cognitive load is important in safety-critical fields such as flight control, and healthcare professionals working e.g. in the emergency room or the first aid unit, but also in everyday life like driving a vehicle, and in training and education applications for improved learning. It may also help in detecting early signs of cognitive impairments. Furthermore, it has been suggested that cognitive load of medical professionals should be monitored when considering the use of artificial intelligence assisted decision making tools in healthcare [8].

Specifically, using a few minutes of baseline data for model personalization is proposed. Different normalization functions and baseline durations are investigated, and self-reported cognitive load is detected as a continuous variable with a regression model. Additionally, model behavior is explained with a feature contribution analysis with SHAP values. The approach is evaluated on an open-source dataset ADABase [23] having several physiological signals measured in a controlled laboratory protocol consisting of simulated driving and n-back tasks. Such a dataset offers clear signals and tasks which likely results with a rather high cognitive load, making it suitable for the first evaluation of the proposed approach. The main contributions of the study are listed as follows:

- Different normalization strategies are evaluated to use short baseline period for cold-start model personalization in detecting continuous cognitive load.
- Minimal baseline duration for optimal performance is estimated.
- Feature importance and contribution analysis is provided to assess which factors increase and decrease experienced cognitive load.

Baseline User Calibration for Cold-Start Model Personalization in Mental State Estimation

Jaakko Tervonen[1(✉)], Rajdeep Kumar Nath[2], Kati Pettersson[1], Johanna Närväinen[2], and Jani Mäntyjärvi[3]

[1] VTT Technical Research Centre of Finland, Tekniikantie 1, Espoo, Finland
{jaakko.tervonen,kati.pettersson}@vtt.fi
[2] VTT Technical Research Centre of Finland, Microkatu 1, Kuopio, Finland
{rajdeep.nath,johanna.narvainen}@vtt.fi
[3] VTT Technical Research Centre of Finland, Kaitoväylä 1, Oulu, Finland
jani.mantyjarvi@vtt.fi

Abstract. Robust human state detection based on analysis of physiological signals requires model personalization since physiological reactions are individual. Personalization requires prior information, which is not available for a new, unknown person, i.e. in a cold-start. To overcome this, the current study proposes user calibration, which uses easily obtainable short baseline measurements to normalize physiological variables individually. Experiments were conducted on a cognitive load detection use case to determine effectiveness of the approach, required baseline duration, and the most suitable normalization function. In addition, the behavior of the model was analyzed with Shapley additive explanations to assess its trustworthiness. The results showed that user calibration always beat the non-personalized model, the optimal baseline duration was 3–3.5 min, and there were no differences between the different normalization functions. The model paid the greatest attention to the physiological phenomena found to be indicative of cognitive load in previous studies. The results encourage further evaluation of user calibration in different use cases for smart healthcare.

Keywords: cold-start · physiology · cognitive load · personalization

1 Introduction

Recent advances in sensor technology have enabled pervasive monitoring of people's physiology, which facilitates real-time detection of stress, and mental and cognitive state of the user, to name a few. Knowledge of the user's state can be utilized in the design and implementation of novel interactive applications

The work was funded by VTT and the Academy of Finland under GrantNos: 334092, 351282, 355693.

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 34–48, 2024.
https://doi.org/10.1007/978-3-031-59717-6_3

41. Saha, S., Chant, D., Welham, J., McGrath, J.: A systematic review of the prevalence of schizophrenia. PLoS Med. **2**(5), e141 (2005)
42. Silver, L.: Smartphone ownership is growing rapidly around the world, but not always equally (2019)
43. Torous, J., Kiang, M.V., Lorme, J., Onnela, J.P., et al.: New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health **3**(2), e5165 (2016)
44. Wander, C.: Schizophrenia: opportunities to improve outcomes and reduce economic burden through managed care. Am. J. Manag. Care **26**, S62–S68 (2020)
45. Wang, R., et al.: Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In: 2016 ACM Int. Joint Conf. Pervasive & Ubiquitous Comput., pp. 886–897 (2016)
46. Wang, R., et al.: Predicting symptom trajectories of schizophrenia using mobile sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 1–24 (2017)
47. Wang, W., et al.: Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2020)

22. Insel, T.R.: Digital phenotyping: technology for a new science of behavior. JAMA **318**(13), 1215–1216 (2017)
23. Insel, T.R.: Digital phenotyping: a global tool for psychiatry. World Psychiatry **17**(3), 276 (2018)
24. Jacobson, N.C., Feng, B.: Digital phenotyping of generalized anxiety disorder: using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life. Transl. Psychiatry **12**(1), 336 (2022)
25. Jacobson, N.C., Summers, B., Wilhelm, S.: Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J. Med. Internet Res. **22**(5), e16875 (2020)
26. Kamath, J., Barriera, R.L., Jain, N., Keisari, E., Wang, B.: Digital phenotyping in depression diagnostics: Integrating psychiatric and engineering perspectives. World J. Psychiatry **12**(3), 393 (2022)
27. Lillie, E.O., Patay, B., Diamant, J., Issell, B., Topol, E.J., Schork, N.J.: The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Pers. Med. **8**(2), 161–173 (2011)
28. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11) (2008)
29. McCutcheon, R.A., Marques, T.R., Howes, O.D.: Schizophrenia-an overview. JAMA Psychiatry **77**(2), 201–210 (2020)
30. Melcher, J., Hays, R., Torous, J.: Digital phenotyping for mental health of college students: a clinical review. BMJ Ment Health **23**(4), 161–166 (2020)
31. Mohr, D.C., Shilton, K., Hotopf, M.: Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. NPJ Digital Med. **3**(1), 45 (2020)
32. Mohr, D.C., Zhang, M., Schueller, S.M.: Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu. Rev. Clin. Psychol. **13**, 23–47 (2017)
33. Morriss, R., Vinjamuri, I., Faizal, M.A., Bolton, C.A., McCarthy, J.P.: Training to recognise the early signs of recurrence in schizophrenia. Cochrane Database of Systematic Reviews (2013)
34. Nahum-Shani, I., Smith, S.N., Spring, B.J., Collins, L.M., Witkiewitz, K., Tewari, A., Murphy, S.A.: Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann. Behav. Med. **52**(6), 446–462 (2018)
35. National Collaborating Centre for Mental Health (UK and others): Psychosis and schizophrenia in adults: treatment and management. London: National Collaborating Centre for Mental Health (2014)
36. Onnela, J.P.: Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology **46**(1), 45–54 (2021)
37. Patel, K.R., Cherian, J., Gohil, K., Atkinson, D.: Schizophrenia: overview and treatment options. Pharm. Ther. **39**(9), 638 (2014)
38. Perez-Pozuelo, I., Spathis, D., Clifton, E.A., Mascolo, C.: Wearables, smartphones, and artificial intelligence for digital phenotyping and health. In: Digital Health, pp. 33–54. Elsevier (2021)
39. Punja, S., Bukutu, C., Shamseer, L., Sampson, M., Hartling, L., Urichuk, L., Vohra, S.: N-of-1 trials are a tapestry of heterogeneity. J. Clin. Epidemiol. **76**, 47–56 (2016)
40. Rhemtulla, M., Fried, E.I., Aggen, S.H., Tuerlinckx, F., Kendler, K.S., Borsboom, D.: Network analysis of substance abuse and dependence symptoms. Drug Alcohol Depend. **161**, 230–237 (2016)

3. Ascher-Svanum, H., et al.: The cost of relapse and the predictors of relapse in the treatment of schizophrenia. BMC Psychiatry **10**, 1–7 (2010)
4. Bak, M., Drukker, M., Hasmi, L., van Os, J.: An n= 1 clinical network analysis of symptoms and treatment in psychosis. PLoS ONE **11**(9), e0162811 (2016)
5. Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., Onnela, J.P.: Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology **43**(8), 1660–1666 (2018)
6. Beard, C., Millner, A.J., Forgeard, M.J., Fried, E.I., Hsu, K.J., Treadway, M.T., Leonard, C.V., Kertz, S., Björgvinsson, T.: Network analysis of depression and anxiety symptom relationships in a psychiatric sample. Psychol. Med. **46**(16), 3359–3369 (2016)
7. Ben-Zeev, D., et al.: Crosscheck: integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. Psychiatr. Rehabil. J. **40**(3), 266 (2017)
8. Benoit, J., Onyeaka, H., Keshavan, M., Torous, J.: Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. Harv. Rev. Psychiatry **28**(5), 296–304 (2020)
9. Birchwood, M., Spencer, E., McGovern, D.: Schizophrenia: early warning signs. Adv. Psychiatr. Treat. **6**(2), 93–101 (2000)
10. Borsboom, D., et al.: Network analysis of multivariate data in psychological science. Nature Rev. Methods Primers **1**(1), 58 (2021)
11. Bradbury, J., Avila, C., Grace, S.: Practice-based research in complementary medicine: could n-of-1 trials become the new gold standard? In: Healthcare, vol. 8, p. 15. MDPI (2020)
12. Brown, L.A., et al.: Digital phenotyping to improve prediction of suicidal urges in treatment: study protocol. Aggress. Violent. Beh. **66**, 101733 (2022)
13. Canas, J.S., Gomez, F., Costilla-Reyes, O.: Counterfactual explanations and predictive models to enhance clinical decision-making in schizophrenia using digital phenotyping. arXiv preprint arXiv:2306.03980 (2023)
14. Chalmers, T.C., et al.: A method for assessing the quality of a randomized control trial. Control. Clin. Trials **2**(1), 31–49 (1981)
15. Chong, H.Y., Teoh, S.L., Wu, D.B.C., Kotirum, S., Chiou, C.F., Chaiyakunapruk, N.: Global economic burden of schizophrenia: a systematic review. Neuropsychiatric disease and treatment, pp. 357–373 (2016)
16. Davidson, B.I.: The crossroads of digital phenotyping. Gen. Hosp. Psychiatry **74**, 126–132 (2022)
17. Emsley, R., Chiliza, B., Asmal, L., Harvey, B.H.: The nature of relapse in schizophrenia. BMC Psychiatry **13**, 1–8 (2013)
18. Fisher, A.J., Medaglia, J.D., Jeronimus, B.F.: Lack of group-to-individual generalizability is a threat to human subjects research. Proc. Natl. Acad. Sci. **115**(27), E6106–E6115 (2018)
19. Fonseca-Pedrero, E., Al-Halabí, S., Pérez-Albéniz, A., Debbané, M.: Risk and protective factors in adolescent suicidal behaviour: a network analysis. Int. J. Environ. Res. Public Health **19**(3), 1784 (2022)
20. He-Yueya, J., Buck, B., Campbell, A., Choudhury, T., Kane, J.M., Ben-Zeev, D., Althoff, T.: Assessing the relationship between routine and schizophrenia symptoms with passively sensed measures of behavioral stability. NPJ Schizophr. **6**(1), 35 (2020)
21. Hevey, D.: Network analysis: a brief overview and tutorial. Health Psychol. Behav. Med. **6**(1), 301–328 (2018)

participants we observe a reduced difference between distributions, participant 14 returns a significantly larger result ($t = 112.87$), suggesting that visiting locations outside of the home has an influence over how this individual reports their EMAs. Continued analysis of each behavioral context at a positive and negative EMA level would yield insights into whether or not this influence is specific to one set of EMAs more so then the other.

6 Discussion

This paper presents an application of an n-of-1 network analysis, leveraging qualitative EMA data whilst factoring in changes and differences in behavioral context measured via sensor data. Specifically, our method allows researchers and clinicians to study, in both exploratory and confirmatory ways, to which degree mental health variables collected via EMA as well as the relation among variables in networks differ across situations. As such, the proposed method provides an inroad to combining multimodal data sources in clinical research and practice, with the goal to enable the potential development of bespoke treatment/management pathways. As an example, the results produced in Fig. 4 reveal that, for this participant at least, not leaving the home significantly changes the way this person reports on his/her positive EMA questions. This new insight coupled with the structure of generated networks could give qualitative actionable information, enabling timely and adaptive interventions for this person's care going forward [34]. The next step from this approach will be to expand this methodology and to represent this information in a format this is comprehensible and explainable to the larger psychiatry community.

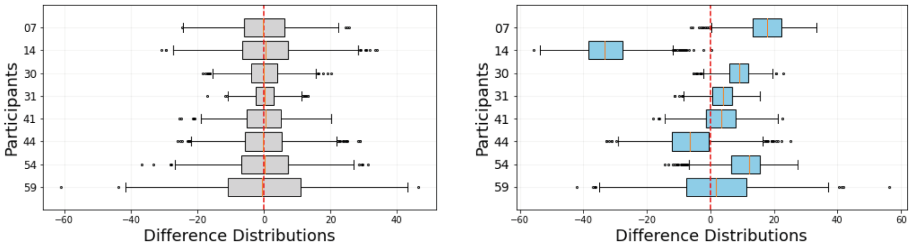
In summary, the task of how to meaningfully analyse multimodal behavioral sensor data with a complex array of concurrent qualitative self reported data is not well understood, particularly for SMI applications. In this analysis we demonstrate an n-of-1 network analysis approach applied solely to self-reported contextual behavioral data. Networks generated from these distinct periods of behavioral context reveal differences in self-reporting habits, differences beyond chance. This is a first stage indicative approach for datasets similar to CrossCheck which is computationally inexpensive, easily deployed and may lead to actionable clinical insights. However further studies are required to better understand how such insights can be utilized in practice, which are clinically effective but that are also compliant of ethical, regulator and legal requirements.

References

1. American Psychiatric Association, A., Association, A.P., et al.: Diagnostic and statistical manual of mental disorders: DSM-IV, vol. 4. American psychiatric association Washington, DC (1994)
2. American Psychiatric Association, D., Association, A.P., et al.: Diagnostic and statistical manual of mental disorders: DSM-5, vol. 5. American psychiatric association Washington, DC (2013)

ing the hypothesis that for this participant activities associated with sociability and social engagement have a marked influence on their self-reported EMAs.

Daily Number of Locations Visited: The following results are from 8 participant's who returned distributions for this behavioral context across all 10 EMAs.



(a) Baseline (b) Locations Visited

Fig. 7. Results across for **all 10 EMAs** for participants who only returned results for **Daily Number of Locations Visited**. Plot color varies depending on statistical significance when using pair-sampled t-testing to compare with the baseline distribution - blue if $p < 0.05$ otherwise plot remains grey. (Color figure online)

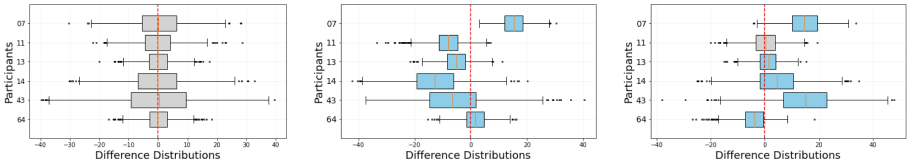
The side-by-side plots in Fig. 7a visualize both the baseline and behavioral context distributions for each valid participant, again, we observe statistical significance ($p < 0.05$) across each individual. Table 5 provides a numerical breakdown of these distributions.

Table 5. Results across for **all 10 EMAs** for participants who returned results for **Daily Number of Locations Visited**. Results include; mean ($\bar{x}$), standard deviation (σ), t-score (t), and p-value (p).

ID	Baseline		Daily Number of Locations Visited		
	$\bar{x}$	σ	$\bar{x}$	σ	t
07	−0.07	8.76	17.42	6.40	−55.33 *
14	0.44	10.28	−32.71	8.16	112.87 *
30	0.09	5.60	8.96	4.15	−56.95 *
31	0.10	4.43	3.80	4.39	−21.34 *
41	−0.03	7.84	3.34	7.14	−6.05 *
44	−0.09	8.21	−6.26	8.56	22.57 *
54	0.09	10.32	10.81	6.80	−35.19 *
59	0.13	16.00	1.54	13.96	−2.25 **

* $p < 0.001$, ** $p < 0.05$

As with previous behavioral contexts, the results presented in Table 5 further demonstrate the unique behavioral patterns of each participant. Whilst in certain



(a) Baseline (b) Calls Made (c) Conversations

Fig. 6. Results across for **all 10 EMAs** for participants who returned results for both selected behavioral contexts - **Daily Number of Calls Made** and **Daily Number of Detected Conversations**. Plot color varies depending on statistical significance when using pair-sampled t-testing to compare with the baseline distribution - blue if $p < 0.05$ otherwise plot remains grey. (Color figure online)

As expected, the baseline distributions in Fig. 6a indicate a mean close to 0, demonstrating the consistency of randomized sampling. However following the same process as with the previous individual's results, a comparison between each participant's baseline distribution and behavioral context distribution demonstrates varying degrees of difference - further highlighting the individuality of participant data.

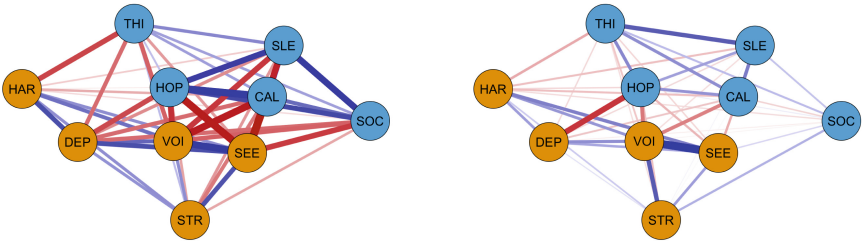
Table 4. Participant results for paired-sample t-testing between their baseline distribution and **both** Number of Calls Received and Calls Made. For each behavior results include; mean ($\bar{x}$), standard deviation (σ), t-score (t), and p-value (p).

ID	Baseline		Daily Number of Calls Made			Daily Number of Conversations		
	$\bar{x}$	σ	$\bar{x}$	σ	t	$\bar{x}$	σ	t
07	−0.18	8.75	15.45	4.97	−49.65 *	14.67	6.57	−46.80 *
11	−0.02	6.81	−8.47	5.43	44.06 *	0.24	5.41	−1.24
13	−0.01	5.33	−4.73	4.90	19.93 *	1.00	4.09	−4.91 *
14	−0.27	10.62	−12.40	9.58	39.43 *	3.88	9.51	−13.01 *
43	−0.42	13.63	−5.45	12.12	12.91 *	14.43	11.44	−37.31 *
64	−0.11	4.88	1.55	4.61	−9.79 *	−4.16	5.00	23.47 *

* $p < 0.001$

Table 4 provides a breakdown of results for each participant visualized in Fig. 6. Whilst in most cases analysis of these two behaviors produces a statistically significant p-value ($p < 0.001$), there is one instance where this is not the case - participant 11 for daily number of detected conversations. A population level analysis, or an unfiltered analysis would have obscured this outlier individual given the variability in distributions across each participant. Moreover participant 07, the same participant analysed previously (see Sect. 5.1), presents significant differences for each of these behavioral contexts; further strengthen-

two networks that each visualize one of the two categories for variations in the daily number of locations this individual participant has visited.



(a) Days with 0 locations visited (b) Days with 1 or more locations visited

Fig. 5. Cross-Sectional Networks for **Daily Number of Locations Visited**. Nodes: **DEP**ressed, **HAR**m, **SEE**ing Things, **STR**essed, Hearing **VOI**ces, **CAL**m, **HOP**e, **SLE**ep, **SOC**ial, **THI**nking Clearly. Thicker edges denote stronger relationships with blue edges indicating positive correlations and red negative. As expected, one can observe positive relations between positive EMAs (i.e., when this person reports one positive EMA, they are more likely to report others, too); positive relations between negative EMAs; and negative relations between positive and negative EMAs. (Color figure online)

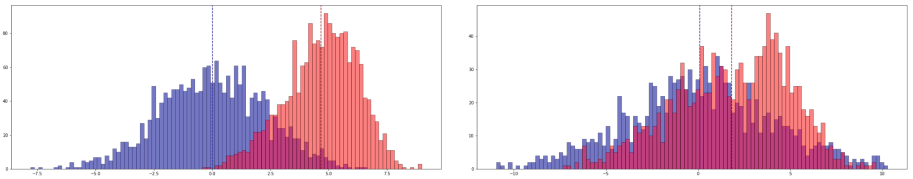
Both networks in Fig. 5 present strong positive relationships between auditory and visual hallucinations, however, on days where this participant remained at home (see Fig. 5a) we see a much more complex network with the presence of stronger edges in greater numbers. A visual analysis of these two networks suggests that, for this particular participant, there is a beneficial link between social engagement (interacting with locations away from home) and improvements in self-reported EMAs. In particular, Fig. 5a illustrates strong negative relationships between each hallucinatory node and feelings of calm and hopeful, this suggests that increased instances of one has a detrimental impact on the other. For example, hallucinations experienced at home could have a more noticeable detrimental impact on this participant's ability to feel calm and hopeful; likewise it could also suggest that feeling calm within this participant's own home reduces the likelihood of them experiencing increased hallucinations.

5.2 By Behavioral Context

In the interest of brevity the following results focus on the analysis of all 10 EMAs for three specific behavioral contexts. The daily number of calls made, daily number of conversations detected and daily number of locations visited.

Daily Number of Calls Made & Detected Conversations: Fig. 6 presents the distributions of 6 participants who each returned results for both the daily number of calls made and number of conversations detected.

number of locations visited, there is not only an observable difference between behavior and baseline, but there is a notable difference when comparing positive and negative EMAs.



(a) Positive EMAs Only (b) Negative EMAs Only

Fig. 4. More detailed analysis of participants results for both their baseline distribution (Blue) and **Number of Locations Visited** distribution (Red). Figure 4a clearly shows an observable difference between the baseline distribution and the behavioral context distribution for positive EMAs. (Color figure online)

Figure 4 provides a more detailed visualization of this participant's baseline distribution (in blue) and variations in their daily number of locations visited (in red). Upon visual inspection, there is a clear observable difference in distributions produced using positive EMAs (Fig. 4a) when compared to those produced using negative EMAs (Fig. 4b). This suggests that, for this participant, there is a noticeable impact on their self-reported positive EMAs when factoring in variations in the daily number of locations visited. Table 3 provides a breakdown of these results across all EMA groups, with a consistent p-value < 0.001 indicating these results are statistical significant. Moreover, we see a more sizeable t-score for this participant's positive EMAs (-72.78), further suggesting that variations in this behavioral context impacts this individual's self-reporting habits - particularly for their positive EMAs.

Table 3. Participant's baseline and **Daily Number of Locations Visited** test results, including mean ($\bar{x}$), standard deviation (σ), t-score (t) and p-value (p)

	Baseline		Daily Number of Locations Visited		
	$\bar{x}$	σ	$\bar{x}$	σ	t
All EMAs	-0.44	9.06	17.30	6.22	-55.62 *
Positive EMAs	0.05	2.28	4.61	1.55	-72.72 *
Negative EMAs	-0.05	3.80	1.83	3.07	-13.21 *

* $p < 0.001$

Whilst numerically these results indicate a statistical significance for this particular behavioral context; network generation provides a visualization of the relationship between EMAs. Figure 5 provides a side-by-side comparison between

network matrices from each other. In probability theory, Central Limit Theorem (CLT) suggests a sample size of approximately 30, however, to maximise the number of viable participants a reduced sample size of 25 days is used throughout the permutation testing process. Repeated 2000 times, a new observed difference in network connectivity is calculated for each permutation, from which a distribution of these differences is then produced. Figure 2 demonstrates this process using the daily number of locations visited as an example.

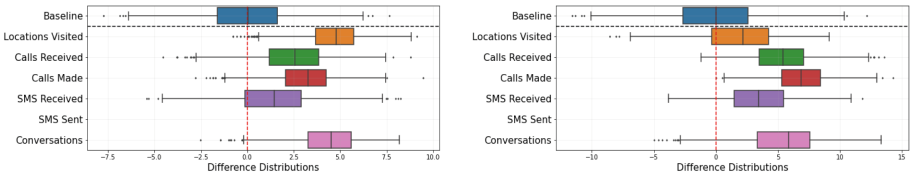
The resulting distribution of differences in network connectivity for a given behavior can then be compared with a baseline distribution. This baseline undergoes the same testing process but is generated using randomly sampled data, and serves as an empirical distribution from which comparisons can then be made. To measure statistical significance, paired-sample t-testing is used to compare a given behavioral distribution with the empirical baseline distribution. The resulting t-score and p-value can then be used to confirm or reject the null hypothesis that a selected sensor based behavioural context has no discernible influence over an individual's network connectivity.

5 Results

In this section we present the findings of our analysis; first detailing results for a single participant, and then at a wider level for multiple participants.

5.1 Example of an Individual CrossCheck Participant

Having recorded 330 days of usable data, this individual returned results across 5 of the 6 selected behavioral contexts. Figure 3 visualizes the results produced during permutation testing for each behavior according to networks generated using only positive, and only negative EMAs (see Table 1).



(a) Positive EMAs Only (b) Negative EMAs Only

Fig. 3. Side-by-side comparison of permutation test results for a single individual across all selected behavioral contexts as well a baseline test - networks generated during testing process used positive EMAs only versus negative EMAs only

In each instance, we observe differences when comparing each behavior to a baseline distribution, for example in Fig. 3a, we observe a baseline mean of 0 compared to a mean of 4.7 in detected conversations. Focusing in on daily

4.2 Network Structure Estimation and Description

Although participant data is initially chronologically ordered, filtering according to a selected behavioral category, either social isolation or sociability, results in time-series segmentation. This segmentation necessitates the identification of a statistical model that can account for this lack of temporal consistency. Correlation network models applied to cross-sectional data deal well with this lack of consistency, as they are effective at visualizing relationships between variables at a specific point in time [10]. As a result we can use this model to generate networks for each person that represent an aggregated average for a sample taken from each selected behavioral category. This allows us to compare network structures of EMAs when participants are, for instance, spending time socially isolated (e.g. remaining at home) or engaging socially (e.g. visiting locations outside of their home). Structurally, EMAs are represented as network nodes with edges between nodes visualizing the linear relationship between each. The strength and sign of any given relationship is defined by the correlation coefficient. Numerically, correlations between nodes range between -1 and 1; with -1 indicating a perfect negative relationship, 1 a perfect positive relationship, and 0 no relationship at all. Pearson's R Correlation Coefficient is used to calculate the associations between all 10 EMA nodes. For each behavioral context two networks are generated, one for each category within a given behavior (Table 2).

4.3 Permutation Testing

To discern whether variations in each behavioral context and their observed network structures differ in statistically meaningful ways, permutation testing is used. The goal of which is to evaluate the null hypothesis that these variations have no discernible concurrence with a participant's self-reported EMAs.

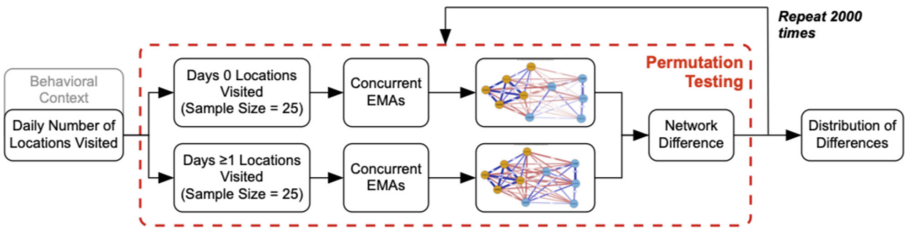


Fig. 2. Permutation testing process for **Daily Number of Locations Visited**, networks are generated using concurrent EMAs from each behavioral category, in turn allowing for the calculation of differences in network connectivity.

Procedurally, permutation testing requires a selected behavior to be filtered according to predefined categories (see Table 2). A network is generated for each category using a 25 day sample, at the end of each permutation the observed difference in network connectivity is calculated by subtracting the sum of the

4 Method

The following section presents the proposed methodology employed in this study, from prerequisite data pre-processing, through to initial network generation and statistical analyses. Data pre-processing requires basic resampling and thresholding, with network generation simply based on the correlations within the 10 set EMA questions (see Table 1). As such giving low computational requirements and ease of re-implementation.

4.1 Data Pre-processing

From CrossCheck's original study arm, 50 participants have been identified from which further analysis can be conducted, these individuals were selected for the quantity and quality of their recorded data. Participants whose engagement was limited, or who recorded inconsistent and unusable data were omitted from further analysis. Whilst both sensory and device usage data is temporally continuous, self-reported EMA responses are only given every 2 to 3 days [45]. However, each EMA question also pertains to days prior to a given response, as such we can retrospectively replicate each EMA score to also be concurrent with sensor data recorded between EMA responses. Any days that fall outside of the 2 day back fill window are omitted from a participant's dataset.

Selection of sensor features was based on their ability to effectively capture defined behavioral contexts, without the need for further processing to map sensor data to a particular context (e.g. it is a reasonable assumption that no location data outside of the primary residence indicates not leaving the home, and that no calls or detected conversations indicates not verbally socializing). Table 2 lists these selected features and their corresponding categories. These categories are created based on 2 factors; first relating to the veracity of the behavioral context that the data represents as stated above, with the second factor being based on well understood behavioral contexts for this SMI, such as social isolation and sociability. Along side these features random sampling of unfiltered data is also conducted from which an empirical distribution is generated, this serves as a baseline from which comparisons can then be made.

Table 2. Selected behavioral contexts and their corresponding categories.

Behavioral Feature in a 24 h period	Periods of Social Isolation	Periods of Sociability
Baseline	*Random Unfiltered Sample*	*Random Unfiltered Sample*
Locations Visited	*No locations visited*	≥ 1 *locations visited*
Calls Made	*No calls made*	≥ 1 *calls made*
Calls Received	*No calls received*	≥ 1 *calls received*
SMS Messages Sent	*No SMS messages sent*	≥ 1 *messages sent*
SMS Messages Received	*No SMS messages received*	≥ 1 *messages received*
Conversations Detected	*No detected conversations*	≥ 1 *detected conversations*

3 Dataset

The CrossCheck [45] dataset originally consisted of 150 participants, each of whom met the criteria for schizophrenia as defined in the DSM-IV [1] and DSM-V [2], whilst also meeting CrossCheck's inclusion criteria [7,45]. Organised into two groups of 75, participants within the CrossCheck study arm were each issued with a smartphone that continuously recorded a range of behavioral, sensory and self-reported EMA data over a 12 month period [7,45,47]. Embedded smartphone sensors passively record daily behaviours and activities continuously, whilst EMAs were self-reported every 2 to 3 days [45]. Of particular relevance to this paper are the following features:

Ecological Momentary Assessment (EMA): EMAs afford a viable way of capturing real-time psychological data within a natural environment. Every 2 to 3 days, CrossCheck administered a 10-item self-reporting assessment designed to measure schizophrenia-related thoughts, feelings, and behaviors [7,45]. Each question (see Table 1) was answered on a scale from 0 ("Not at all") to 3 ("Extremely"). For ease of analysis and understanding, each EMA is grouped according to either its positive or negative association.

Table 1. CrossCheck EMA Questions

Positive EMAs	Negative EMAs
+ Have you been feeling **CALM**?	- Have you been **DEPRESSED**?
+ Have you been **SOCIAL**?	- Have you been feeling **STRESSED**?
+ Have you been **SLEEPING** well?	- Have you been bothered by **VOICES**?
+ Have you been able to **THINK** clearly?	- Have you been **SEEING THINGS** other people can't see?
+ Have you been **HOPEFUL** about the future?	- Have you been worried about people trying to **HARM** you?

Behavioural Sensing: Study smartphones continuously collected a wide range of behavioral features for each participant, however only the following behavioral features are relevant to this paper due to their close association with sociability and social isolation. - *Geo-spatial Activity:* refers to timestamped locational data derived using a combination of device GPS, Wifi and cellular network towers [7,45]. - *Speech Frequency & Duration:* periods of human speech was inferred from ambient sound using the inbuilt device microphone [7]. - *Calls & SMS:* The frequency and duration of incoming and outgoing calls, as well as the number of incoming and outgoing SMS messages is passively logged and recorded by the CrossCheck application [7,45].

[45]. These findings were further supported through research into behavioral stability using the same dataset, this stability index drew on participant's passively recorded features and behaviours to assess the extent to which a diagnosed participant adheres to a stable routine. This study identified correlations between the stability index of recorded features and symptomatic severity, the results of which demonstrated that greater periods of stability in social activities - such as calls and SMS messages - was associated with reduced symptoms. In contrast, increased stability in periods of inactivity - time spent still - exhibited an association with increased symptom severity [20]. The findings of these studies highlight not only the highly person-centric nature of CrossCheck's multimodal data, but also the close association between daily behaviors and symptom severity; both of which are of particular importance to this study as we seek to analyse the impact recorded behavioral contexts have on self-reported EMAs at an individual level.

In recent years the use of network analysis within psychological research has become an important tool in the estimation and visualization of psychological data, and can be used to identify multivariate patterns and relationships [10,21]. Within these networks, nodes represent variables such as mood states collected via EMAs, with edges between nodes denoting statistical relationships between said nodes [21]. The process by which these associations and relationships are calculated can vary depending on the initial dataset and selected statistical model [10]. In recent years, network analysis techniques have been applied to a large number of datasets in order to gain deeper insights into a range of mental health problems, including drug and alcohol dependency [40], suicidal behavior in adolescents [19], depression and anxiety [6], and the treatment of psychosis [4]. Network analysis consists of three stages; network structure estimation, network description, and network stability analysis [10]. Network structure estimation refers to the process by which the underlying structure of a network is inferred, involving the selection of relevant nodes and edges as well as selecting an optimal statistical model. Network description is the characterisation of a network which involves understanding network topology and node centrality. Finally, network stability analysis refers to the examination of a network's robustness, consistency and the accuracy of edge weights [10]. Recently, estimating network models on time-series data with numerous repeated observations using EMA data has gained traction, however this presents three distinct challenges. First, most work is estimated at the group level, ignoring potential variation across participants in network structures. Second, networks are stationary, i.e., one network is obtained throughout the time period, assuming network structure does not vary by context. Third, it is unclear how network analysis ought to deal with multimodal (e.g., sensor and EMA) data. Here we tackle all three challenges, by estimating n-of-1 networks according to a selected behavioral (sensor-based) context prior to EMA network generation, as a result enabling more nuanced, context-dependent qualitative networks.

clinical environment. However, the pervasiveness of mobile technology [42] in everyday life is affording researchers and clinicians access to vast quantities of moment-by-moment, in-situ, individual-level data captured by personal digital devices; the granular level quantification of which is referred to as digital phenotyping [32,43]. These personal phenotypes [5,36] provide a digital fingerprint from which psychological, cognitive and behavioural characteristics can be measured and assessed [16,22,23,31]; providing valuable insights into symptomatic markers and effective psychiatric treatments [38]. Within mental health research, digital phenotyping has been employed in a number of studies, including student mental health [30], depression [26], prediction of suicidal urges [12], anxiety disorders [24], social anxiety [25], and psychosis spectrum illnesses [8].

With an increasing emphasis on patient-centred healthcare and individualised medicine [11], digital phenotyping lends itself to move away from the population level [18] and instead conduct n-of-1 trials (or single subject trials); these focus on an individual patient as the sole unit of observation throughout a study [27]. Typically, n-of-1 studies have been used within both clinical and research settings to assess pharmaceutical efficacy and treatment viability within individual participants [39]. The focus of these trials enables the identification of observations or characteristics that may not be evident in a collective population-level analysis. However for larger population samples, the insights gained from n-of-1 trials can contribute to larger-scale Randomized Control Trials (RCTs).

The CrossCheck collection emulated a Randomized Control Trial (RCT) design [14] that explored the viability of continuous remote patient monitoring through a multimodal sensing system; the core aim of which sought to accurately predict indicators of symptomatic and psychotic relapse in Schizophrenia Spectrum Disorders (SSD) [45]. CrossCheck's digital phenotyping dataset identified unique digital indicators of psychotic relapse; for some participants changes in self-reported EMAs provided actionable descriptors of symptom exacerbation, whilst in others, passively recorded behavioural and sensory data proved useful in identifying changes in established behaviors or daily functioning [7]. A recent study demonstrated the detection of decreases in symptoms using change-point algorithms and counterfactual explanations [13]. Additional research using the CrossCheck dataset mapped features on a two-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE), a technique used for dimensionality reduction that projects each high-dimensional data point to a two-dimensional data point [28]. Using t-SNE, CrossCheck visualized data points that represented a participant's behavioural features used to predict EMA responses; when plotted, these data points clustered according to each specific study participant. This clearly demonstrated that there are observable differences between study participants and that CrossCheck's sensor data is highly person dependent [45]. At a population level, the initial study found significant associations between recorded behavioral features and changes in mental health indicators; in particular decreased levels of physical activity and sociability was associated with negative mental health, whilst improvements in established sleep patterns and getting up earlier collated with positive mental health

personal and economic burden at an individual, familial and societal level [15,44]. Symptoms can include hallucinations (both visual and auditory), disordered and delusional thinking, impaired cognitive ability, disorganized speech and behavior [37], as well as increased social isolation, withdrawal and amotivation [29]. Although characterised as a chronic condition, the disease course is not static, with diagnosed individuals typically fluctuating between periods of partial remission and periods of symptomatic relapse [17,35,46]. Studies have identified symptomatic and behavioral changes that can manifest prior to relapse [3,9,17,20], however, these changes often remain undetected until the occurrence of significant negative consequences [45]. Evidence further suggests that timely clinical intervention poses an effective strategy in the prevention of further deterioration, and the transition into a state of full relapse [33,45].

In this paper, we seek to demonstrate a method which consolidates the complexities of subjective self-reported Ecological Momentary Assessments (EMAs) when accounting for variations in behavioral context. Using the CrossCheck dataset, a first of its kind dataset combining real-world, longitudinal behavioral data, and self-reported EMAs specific to schizophrenia [45]; we aim to demonstrate the effectiveness of using sensor-based data to identify periods of various behavioral context, from which network analysis can be applied to observe and compare differences in network connectivity and the relationships between corresponding EMAs. Specifically, we focus on behaviors that can be categorised according to periods of sociability and social isolation, both noted symptoms associated with schizophrenia symptom severity [20]. Figure 1 provides a high-level overview of this process, from individual-level (n-of-1) data through to behavioral filtering and network analysis. Ultimately, the goal of this framework is to reveal behavioral contexts and their resulting impact on self-reported EMAs at an n-of-1 level, in particular providing insights into symptomatic improvement or disease exacerbation.

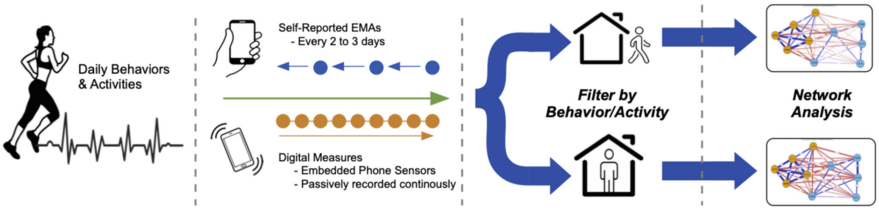


Fig. 1. Overview of the processes involved in this study; networks are generated using EMA responses concurrent with specific sensor-based behavioral contexts.

2 Related Work

Conventional research into human behavior often relies on time and resource intensive data collection through face-to-face engagement in a controlled or

Individual Behavioral Insights in Schizophrenia: A Network Analysis and Mobile Sensing Approach

Andy Davies[1(✉)], Eiko Fried[2], Omar Costilla-Reyes[3], and Hane Aung[1]

[1] School of Computer Sciences, University of East Anglia, Norwich NR9 7TJ, UK
{andy.davies,Min.aung}@uea.ac.uk
[2] Faculty of Social Sciences, Institute of Psychology, Leiden University, Rapenburg 70, 2311 Leiden, Netherlands
e.i.fried@fsw.leidenuniv.nl
[3] Computer-Aided Programming Research Group, MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA 02139, USA
costilla@mit.edu

Abstract. Digital phenotyping in mental health often consists of collecting behavioral and experience-based information through sensory and self-reported data from devices such as smartphones. Such rich and comprehensive data could be used to develop insights into the relationships between daily behavior and a range of mental health conditions. However, current analytical approaches have shown limited application due to these datasets being both high dimensional and multimodal in nature. This study demonstrates the first use of a principled method which consolidates the complexities of subjective self-reported data (Ecological Momentary Assessments - EMAs) with concurrent sensor-based data. In this study the CrossCheck dataset is used to analyse data from 50 participants diagnosed with schizophrenia. Network Analysis is applied to EMAs at an individual (n-of-1) level while sensor data is used to identify periods of various behavioral context. Networks generated during periods of certain behavioral contexts, such as variations in the daily number of locations visited, were found to significantly differ from baseline networks and networks generated from randomly sampled periods of time. The framework presented here lays a foundation to reveal behavioural contexts and the concurrent impact of self-reporting at an n-of-1 level. These insights are valuable in the management of serious mental illnesses such as schizophrenia.

Keywords: Schizophrenia · CrossCheck · n-of-1 · Digital Phenotyping · Network Analysis · Mobile Sensing

1 Introduction

Schizophrenia is a complex, Serious Mental health Illness (SMI) that develops in approximately 1% of the global population [41] and represents a significant

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 18–33, 2024.
https://doi.org/10.1007/978-3-031-59717-6_2

24. Scahill, L., Riddle, M.A., McSwiggin-Hardin, M., Ort, S.I., King, R.A., Goodman, W.K., Cicchetti, D., Leckman, J.F.: Children's yale-brown obsessive compulsive scale: reliability and validity. J. Am. Acad. Child Adolesc. Psychiatry **36**(6), 844–852 (1997)
25. Selles, R.R., et al.: Effects of treatment setting on outcomes of flexibly-dosed intensive cognitive behavioral therapy for pediatric ocd: a randomized controlled pilot trial. Front Psychiatry **12** (2021)
26. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2015)
27. Steil, J., Huang, M.X., Bulling, A.: Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, pp. 1–9. ACM (2018). https://doi.org/10.1145/3204493.3204538
28. Stewart, S.E., Geller, D.A., Jenike, M., Pauls, D., Shaw, D., Mullin, B., Faraone, S.V.: Long-term outcome of pediatric obsessive-compulsive disorder: a meta-analysis and qualitative review of the literature. Acta Psychiatr. Scand. **110**(1), 4–13 (2004)
29. Thierfelder, A., et al.: Multimodal sensor-based identification of stress and compulsive actions in children with obsessive-compulsive disorder for telemedical treatment. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2976–2982 (2022). https://doi.org/10.1109/EMBC48229.2022.9871899
30. Tichon, J.G., Wallis, G., Riek, S., Mavin, T.: Physiological measurement of anxiety to evaluate performance in simulation training. Cognition, Technol. Work **16**(2), 203–210 (2014)
31. Ward, J.A., Lukowicz, P., Gellersen, H.W.: Performance metrics for activity recognition. ACM Trans. Intell. Syst. Technol. **2**(1), 1–23 (2011). https://doi.org/10.1145/1889681.1889687
32. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
33. World Health Organization: International classification of diseases for mortality and morbidity statistics (11th revision). https://icd.who.int/browse11/l-m/en. Accessed 9 July 2023
34. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361 (2015)

6. Bradley, M.C., Hanna, D., Wilson, P., Scott, G., Quinn, P., Dyer, K.F.W.: Obsessive-compulsive symptoms and attentional bias: an eye-tracking methodology. J. Behav. Ther. Exp. Psychiatry **50**, 303–308 (2016)
7. Brown, M., Gang Hua, Winder, S.: Discriminative learning of local image descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 43–57 (2011). https://doi.org/10.1109/TPAMI.2010.54
8. Chen, S., Epps, J., Ruiz, N., Chen, F.: Eye activity as a measure of human mental effort in HCI. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, pp. 315–318. IUI 2011. ACM, New York (2011). https://doi.org/10.1145/1943403.1943454
9. Douglass, H.M., Moffitt, T.E., Dar, R., McGee, R., Silva, P.: Obsessive-compulsive disorder in a Birth Cohort of 18-Year-Olds: prevalence and predictors. J. Am. Acad. Child Adolesc. Psychiatry **34**(11), 1424–1431 (1995)
10. Heyman, I., Fombonne, E., Simmons, H., Ford, T., Meltzer, H., Goodman, R.: Prevalence of obsessive-compulsive disorder in the British nationwide survey of child mental health. British J. Psychiatry J. Mental Sci. **179**, 324–329 (2001)
11. Hollis, C., Falconer, C.J., Martin, J.L., Whittington, C., Stockton, S., Glazebrook, C., Davies, E.B.: Annual research review: digital health interventions for children and young people with mental health problems - a systematic and meta-review. J. Child Psychol. Psychiatry **58**(4), 474–503 (2017)
12. Hollmann, K., et al.: Internet-based cognitive behavioral therapy in children and adolescents with obsessive-compulsive disorder: a randomized controlled trial. Front Psychiatry **13** (2022)
13. Klein, C.S., et al.: Smart sensory technology in tele-psychotherapy of children and Adolescents with Obsessive-Compulsive Disorder (OCD): a feasibility study. preprint, SSRN (2023). https://doi.org/10.2139/ssrn.4395216
14. Kübler, T.: Look! Technical specifications, Blickschulungsbrille (2021)
15. Lappi, O.: Eye movements in the wild: oculomotor control, gaze behavior & frames of reference. Neurosci. Biobehav. Rev. **69**, 49–68 (2016)
16. Marquart, G., Cabrall, C., De Winter, J.: Review of eye-related measures of drivers' mental workload. Procedia Manuf. **3**, 2854–2861 (2015)
17. Mataix-Cols, D., de la Cruz, L.F., Nordsletten, A.E., Lenhard, F., Isomura, K., Simpson, H.B.: Towards an international expert consensus for defining treatment response, remission, recovery and relapse in obsessive-compulsive disorder. World Psychiatry **15**(1), 80–81 (2016)
18. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
19. Mullen, M., Hanna, D., Bradley, M., Rogers, D., Jordan, J.A., Dyer, K.F.W.: Attentional bias in individuals with obsessive-compulsive disorder: a preliminary eye-tracking study. J. Behav. Cogn. Ther. **31**(2), 199–204 (2021)
20. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
21. Primbs, J., et al.: The SSTeP-KiZ system-secure real-time communication based on open web standards for multimodal sensor-assisted tele-psychotherapy. Sensors **22**(24), 9589 (2022)
22. Recarte, M.A., Nunes, L.M.: Effects of verbal and spatial-imagery tasks on eye fixations while driving. J. Exp. Psychol. Appl. **6**(1), 31–43 (2000)
23. Santini, T., Niehorster, D.C., Kasneci, E.: Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, pp. 1–10. ACM (2019). https://doi.org/10.1145/3314111.3319835

between patients, like the type of conversation during exposure or the patient-therapist relationship.

Our results underline the effectiveness of our method. However, it should be noted that the method assumes that exposure sessions contain a physical exposure to objects or locations. Therefore, this approach is most suitable for manifestations of OCD where obsessive thoughts are connected to a physical counterpart like an action, an object or a specific location.

In general, our findings stress the importance of analysing fixation behaviour during real-life exposure sessions as an extension of controlled eye tracking studies in the laboratory. Especially fixation duration and variability promise to be valuable parameters for therapists to monitor and adapt the patient's stress level and the connected mental load during exposure sessions. Since our analysis is exemplary on four patients, future studies would have to replicate and validate these results with a larger population.

4.1 Conclusion

We proposed a pipeline suitable for analysing gaze behaviour of patients with OCD during exposure sessions in their home environments. Although further research is needed to validate our findings, our work provides a preliminary argument for the usefulness of eyetracking for patients with OCD. Providing feedback about gaze behaviour could therefore support therapists in monitoring stress and mental load of patients, helping them to adapt to the needs of the patient. Next steps will be to use our approach for behavioural feedback to therapists in exposure exercises practised as homework outside of therapy sessions, to ensure correctness and prevent avoidance behaviour. In future research, it will also be interesting to connect the gaze features not only with perceived stress but also with physiological measures of stress such as heart rate.

References

1. Abramowitz, J.S., Taylor, S., McKay, D.: Obsessive-compulsive disorder. Lancet **374**(9688), 491–499 (2009)
2. Ahmadi, N., et al.: Quantifying Workload and Stress in Intensive Care Unit Nurses: Preliminary Evaluation Using Continuous Eye-Tracking. Human Factors (2022)
3. Armstrong, T., Olatunji, B.O.: Eye tracking of attention in the affective disorders: a meta-analytic review and synthesis. Clin. Psychol. Rev. **32**(8), 704–723 (2012)
4. Basel, D., Hallel, H., Dar, R., Lazarov, A.: Attention allocation in OCD: a systematic review and meta-analysis of eye-tracking-based research. J. Affect. Disord. **324**, 539–550 (2023)
5. Behroozi, M., Lui, A., Moore, I., Ford, D., Parnin, C.: Dazed: measuring the cognitive load of solving technical interview problems at the whiteboard. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, pp. 93–96. ACM, Gothenburg Sweden (2018). https://doi.org/10.1145/3183399.3183415

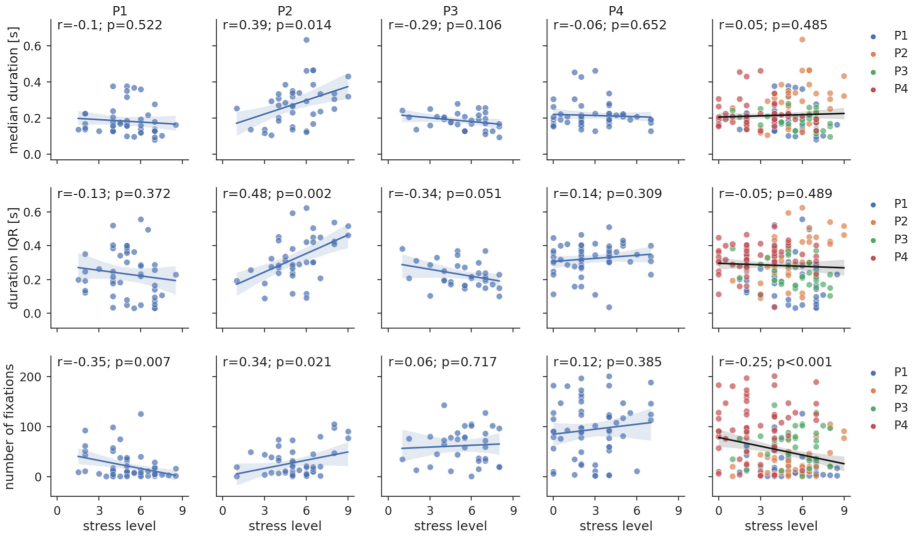


Fig. 6. Correlation of the patients' reported subjective stress levels with the metrics for fixations onto the *therapist* cluster. The first four columns represent a patient each, while the last column presents the results taken across all patients.

movements. We extended the method with unsupervised clustering to automatically identify targets in the real world from the fixated locations. We demonstrated that these clusters have semantic meaning and represent three categories of gaze targets: *exposure-relevant*, *therapist* and *other* locations.

We found that fixation duration onto *exposure-relevant* locations consistently reflected reported stress levels of the patient, i.e., fixation duration correlated positively with higher stress levels. This was accompanied by an increase in fixation variability, suggesting that fixation duration did not change systematically but that only some fixations lasted longer. While similar effects could be observed for *other* locations, these proved smaller and inconsistent across subjects.

Our results reveal that during real-life exposures patients put more attention on exposure-relevant locations if the subjective stress level, and thus the intensity of the exposure, increased. Given that fixation duration has been shown to increase with higher mental load, our results suggest that the patients' mental load and perceived stress level are closely connected.

There was no increase in the amount of fixations, neither onto *exposure-relevant* nor *other* locations, which we would have expected as an effect of the rising stress. In contrast, some patients even showed a decrease in the amount of fixations. Together with the increase in fixation duration, this indicates that reported stress reflects mental effort during exposure rather than anxiety.

Fixation behaviour towards the therapist was highly individual and there was no common effect across patients. This is not surprising given that fixation behaviour towards the therapist can depend on many variables that differ

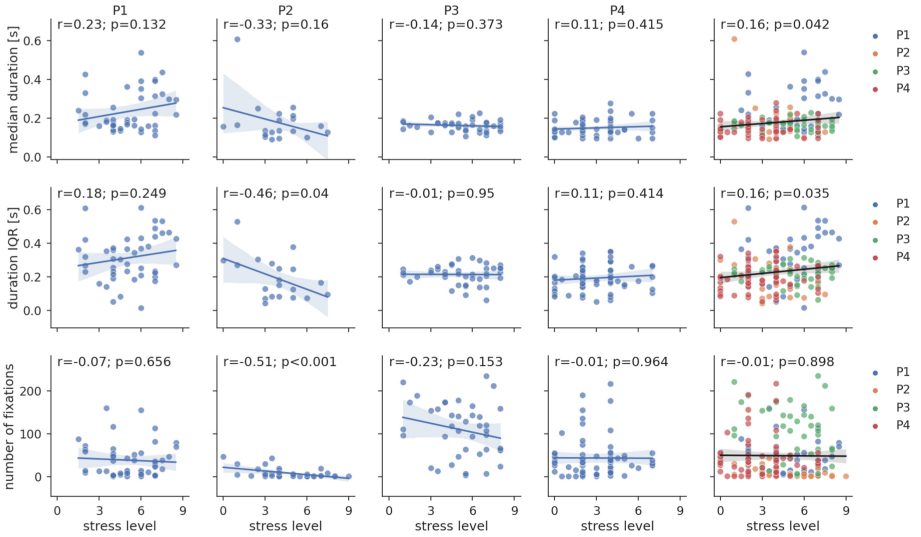


Fig. 5. Correlation of the patients' reported subjective stress levels with the metrics for fixations onto the *other* cluster. The first four columns represent a patient each, while the last column presents the results taken across all patients.

duration variability ($p = 0.04, r = 0.16$), but no correlation with the number of fixations ($p = 0.9$) across all patients. There was no consistent trend of duration increase in single patients and none showed significant results. We observed similar results for duration IQR, which was only significant for P2, where correlation was strongly negative ($p = 0.04, r = -0.46$) as opposed to the positive correlation across all patients. P2 was also the only patient that showed a significant correlation with the number of fixations ($p < 0.001, r = -0.51$).

Correlation results for fixations onto the *therapist* are shown in Fig. 6. For fixation duration and duration variability, there was no correlation with stress levels across all patients ($p > 0.48$) and no common trend across single patients. Correlation was only significant in P2 for fixation duration ($p = 0.01, r = 0.39$) and variability ($p = 0.002, r = 0.48$). The number of fixations across all patients decreased with reported stress level ($p < 0.001, r = -0.25$), which was not consistent across single patients, where the number of fixations decreased with stress for P1 ($p = 0.007, r = -0.35$) but increased for P2 ($p = 0.02, r = 0.34$).

4 Discussion

We proposed a method to analyse the fixation behaviour of children and adolescents with OCD during exposure exercises within video-based CBT. The method was specifically designed for challenges caused by real-life behaviour recorded with mobile eye tracking in home environments. For fixation detection we adapted an approach based on gaze patch similarity that is robust to head

The visualisation of clustering results for P3 along with examples for each cluster is shown in Fig. 3. In the 2D feature space, *therapist* clusters locate at the lower right. The *exposure-relevant* clusters where the pencil holder with the "contaminated" glue was placed in front of the therapist is separated but close to the *therapist* clusters. Clusters connected through the higher semantic level "being in a bathroom" are located at the top left part of the feature space.

3.2 Correlation Results

Results for the correlations of fixation metrics onto *exposure-relevant* locations with the reported stress level are displayed in Fig. 4. There was a significant increase in fixation duration ($p < 0.001, r = 0.41$) and fixation duration variability ($p < 0.001, r = 0.3$) with the perceived stress across all patients, and a small decrease in the number of fixations ($p = 0.01, r = -0.18$). The trend for increased fixation duration was observed in all subjects with statistical significance for P1 ($p = 0.02, r = 0.37$) and P2 ($p = 0.004, r = 0.34$). The increase in the fixation duration IQR did not occur in P1, but was significant for both P2 ($p = 0.016, r = 0.29$) and P3 ($p = 0.01, r = 0.42$). The patient-specific patterns regarding the number of fixations were individually different, and were only significant for P2 ($p = 0.01, r = -0.3$).

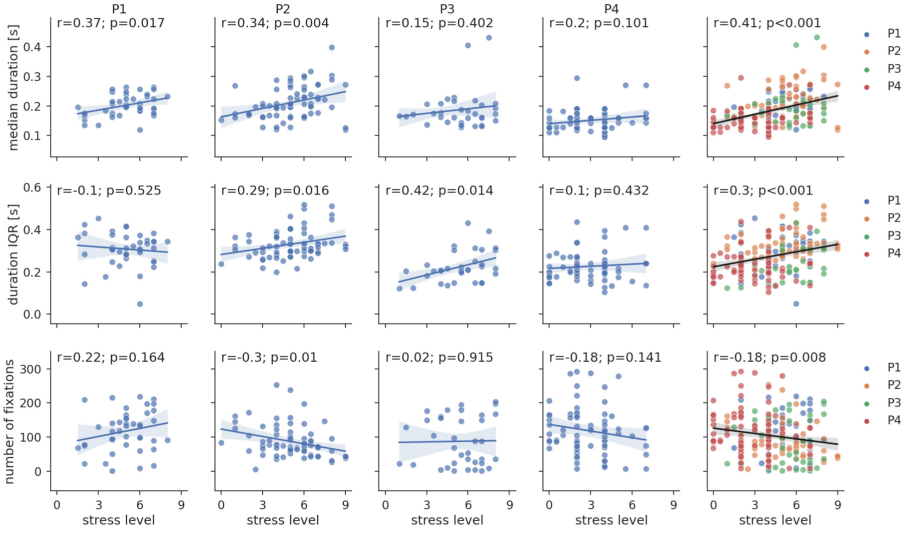


Fig. 4. Correlation of the patients' reported subjective stress levels with the metrics for fixations onto the *exposure-relevant* cluster. The first four columns represent a patient each, while the last column presents the results taken across all patients.

For fixations onto *other* locations, displayed in Fig. 5, we found small correlations of the stress level with fixation duration ($p = 0.04, r = 0.16$) and fixation

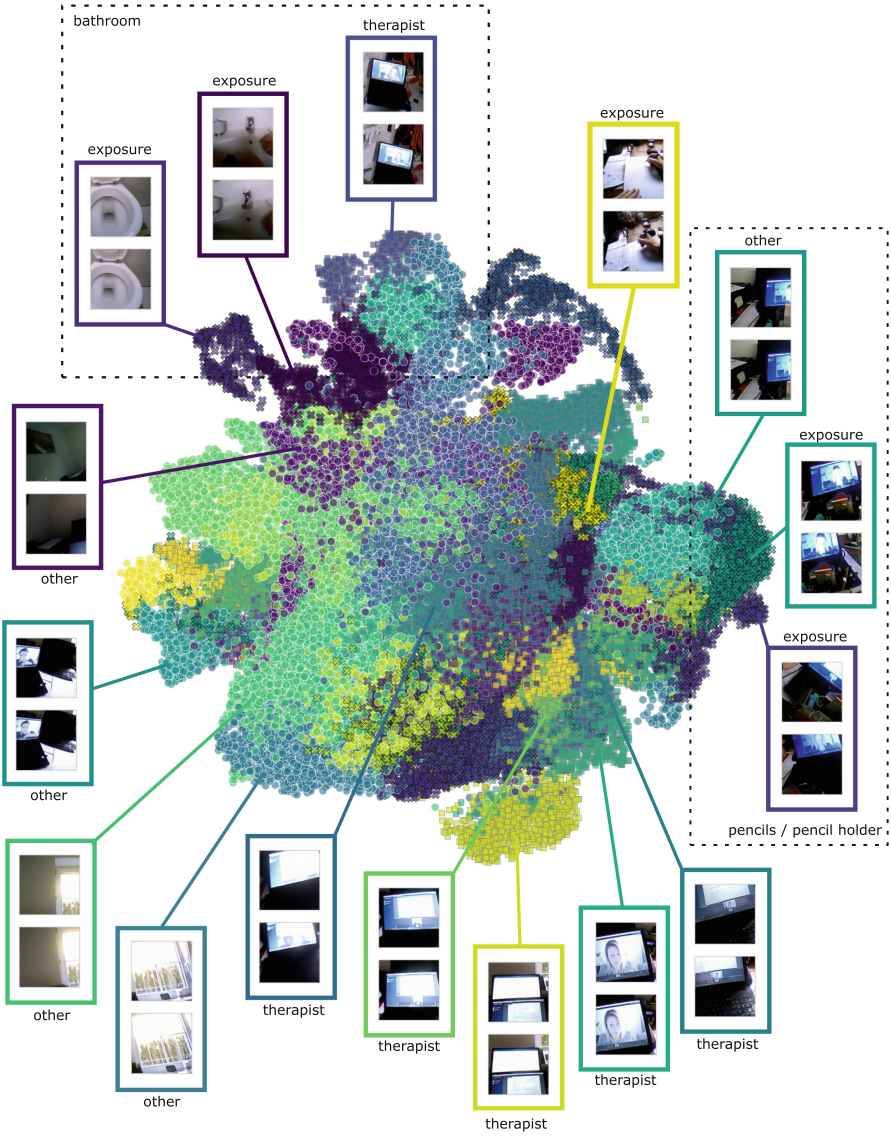


Fig. 3. Visualisation of the clustering results for P3. clusters were projected into a 2D space using UMAP. Clusters are colour-coded, while semantic groups are displayed as different shapes (X-shape: *exposure-relevant*, square: *therapist*, and circle: *other*). For exemplary clusters, the two images closest to the cluster centre are shown for illustration. Semantically meaningful grouped clusters are indicated by dashed lines.

in real-life environments the prediction of mobile eyetracking is not always accurate, cropping a larger window for clustering also reduces the effect of small localisation errors.

The image size was chosen as the optimal input size for the VGG-16 network pretrained on image recognition that we used to extract a feature vector for every representative image [26]. We extracted the output of the last fully-connected layer to get a 4096-dimensional vector containing high-level feature information for each image, resulting in a feature matrix of the size $n_{\text{fixations}} \times 4096$ for each session.

In order to find meaningful clusters across all therapy sessions, we appended the feature matrices of all sessions into one feature matrix. After normalization, the features were clustered with the agglomerative clustering algorithm implemented in the scikit-learn toolbox [20]. Agglomerative clustering is a bottom-up hierarchical clustering algorithm that starts with every sample as a single cluster and subsequently merges the closest clusters until either a maximum distance between clusters or a predefined number of clusters is reached.

The distance between single samples was computed as the euclidean distance and extended to distance between clusters with the Ward linkage criterion [32]. We computed distances between clusters for the full hierarchical tree until all clusters were merged. As stopping criterium, we then defined the maximal distance between clusters as the knee point among the largest 2000 distances.

We visually checked the resulting clusters and grouped them semantically into *exposure-related*, *therapist* and *other* locations. Few clusters showed a mix of different groups and were therefore not assigned. Note here, that the *exposure-related* group can contain very distinct clusters, since the conducted exposure exercises within a patient vary between sessions.

For visualising the results, we calculated a lower-dimensional representation of the features using UMAP [18]. Parameters were chosen to capture both the local and global structure of the data.

2.5 Analysis

For every reported stress level within an exposure exercise, we extracted the fixations within 1.5 min before and after the report. We computed fixation metrics for every semantic group separately, including the number of fixations, the median fixation duration and the interquartile range (IQR) of the fixation durations. Pearson correlation was calculated to assess the relation between the different fixation metrics and the corresponding reported stress level.

3 Results

3.1 Clustering Results

The clustering resulted on average in 31.25 clusters per subject of which on average 25 could be assigned to one of the semantic groups (P1: 25 cluster (21 assigned), P2: 25 (17), P3: 47 (40), P4: 28 (22)).

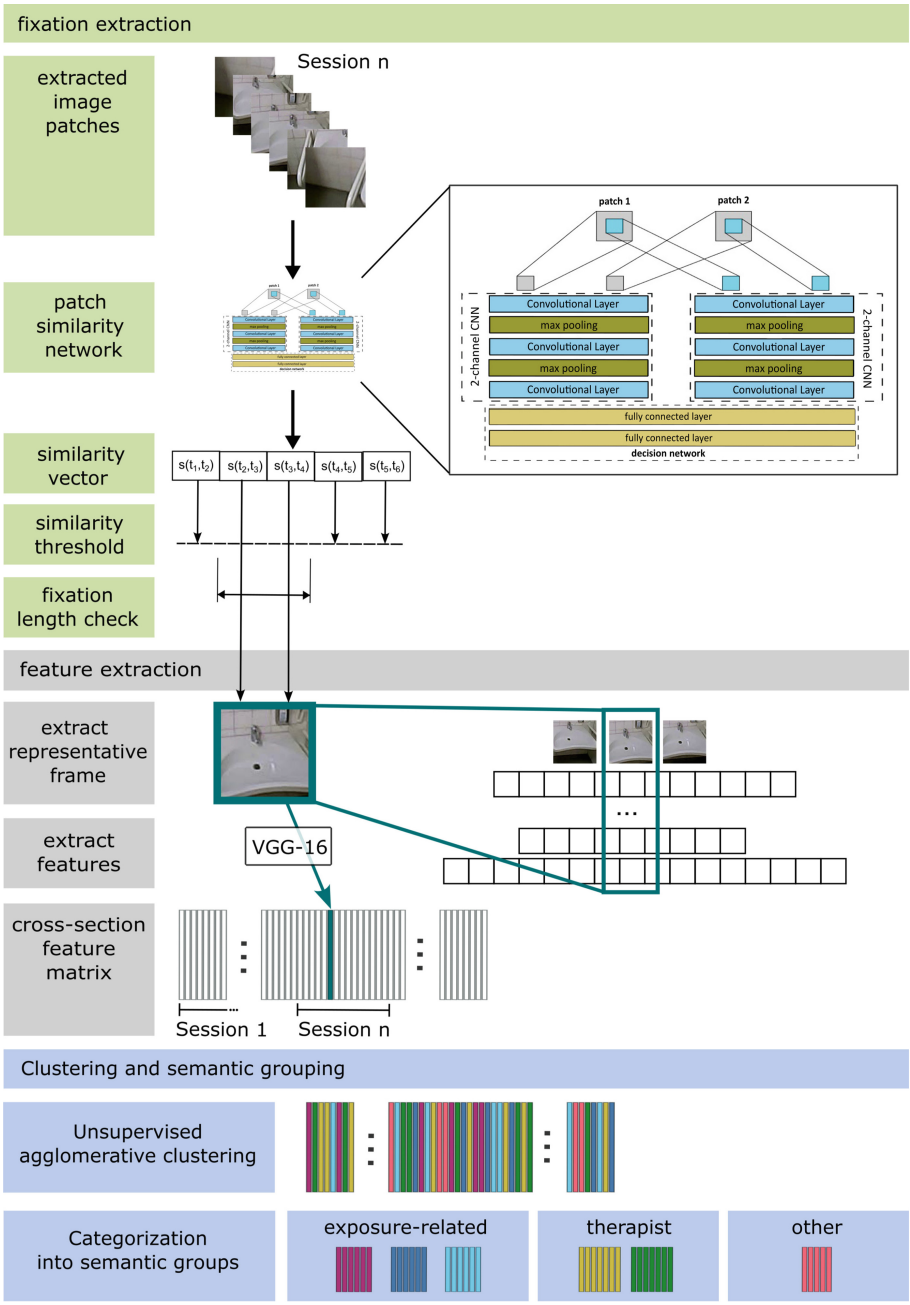


Fig. 2. Visualisation of the fixation processing pipeline, including fixation detection, feature extraction and unsupervised clustering. The network figure is adapted from [34].

therapy with severe symptoms, but showed a surprisingly successful therapeutic effect, managing to reduce symptom severity by 100% to a minimum.

P4 showed a repetition-based manifestation of OCD with the urge to repeat actions with a positive thought until they felt "just right". Exposure sessions consisted mainly of selecting an item once and then performing the action with that item once (e.g., selecting and wearing the clothing item). The exercises were intensified by having to think of a negative event during the one-time execution.

2.4 Fixation Analysis Pipeline

To analyse the fixation behavior in the introduced patients, we adapted a fixation detection method for mobile eye tracking [27], structured the fixations using unsupervised clustering and assigned each cluster to *exposure-relevant*, *therapist* or *other* locations. The complete pipeline can be seen in Fig. 2.

Fixation Detection. First, we reduced noise in the gaze estimation by applying a moving average filter over a window of five frames. We then adapted an approach specifically designed for mobile eye tracking based on the assumption that image patches around the gaze estimation stay similar during a fixation [27].

We cropped an image patch of 50×50 px around the gaze estimation for every frame in the video. Each pair of consecutive frames served as input for a convolutional neural network (CNN) pretrained to predict patch similarity on the liberty dataset [7,34], resulting in a vector containing the patch similarity for each pair of frames. Sequences with similarities above the threshold of 1.3 were kept as fixation candidates. To remove outliers and ensure validity of fixations, we discarded candidates shorter than 3 frames (i.e., 100 ms) or longer than 95% of the data which corresponded to a maximal fixation length of 1 s.

The pipeline including the architecture of the CNN is shown in Fig. 2. It consists of two parallel 2-channel streams, one processing both full patches ("periphery") and one processing the central crops in a higher resolution ("fovea") that are integrated by two fully connected layers. Details on the network architecture can be found in the original publication [34].

To adapt the approach to our data, we tuned three hyperparameters: (1) image patch size, (2) similarity threshold and (3) the dataset for pretraining the patch similarity network. For tuning, we created a validation dataset consisting of a total video length of 15 min (i.e., roughly 27.000 frames) taken from three different subjects of the SSTeP KiZ study. All videos were labeled by at least two and at most three annotators to form the ground truth labels "fixation" and "no fixation" for each frame. The parameters were tuned by evaluating the fixations with event detection performance metrics [31].

Fixation Clustering. For each detected fixation, we extracted the centre frame as a representative and replaced its patch by a larger 256×256 px image cropped around the gaze estimation to obtain more context information. Since especially

Gaze data was recorded using the *Look!* head-mounted eye tracking device [14]. It included a scene camera with a resolution of 640×280 px and two eye cameras with a resolution of 320×240 px each. All videos were recorded at 30 Hz. Gaze estimation was computed using a convolutional neural network designed to be robust against small movements of the eye tracker to reduce the need for frequent recalibration [23]. Patients were asked to calibrate the system regularly, but at least before the first session and towards the middle of therapy.

2.3 Patients

We investigated a sample of four patients that participated in the SSTeP KiZ study. Symptom severity was assessed with the CY-BOCS score (Children's Yale-Brown Obsessive Compulsive Scale) before (t_0) and after treatment (t_1) [24]. A general reference for therapy success is a reduction of the CY-BOCS score by at least 35% [17], however, numerical symptom reduction can differ from personal experience. CY-BOCS values for all patients as well as demographic information and the amount and duration of exposure exercises can be found in Table 1.

Table 1. Overview of the patients included in this work. The table shows the amount of recorded exposure exercises per patient, mean and standard deviation of their durations, and the CY-BOCS score before and after treatment.

Patient			exposure exercises		CY-BOCS		
	age (year)	sex	n	length (min)	t_0	t_1	reduction
P1	17	f	9	38.6 ± 15.6	28	25	10.71%
P2	16	f	10	30.1 ± 11.9	21	13	38.10%
P3	17	m	7	28.9 ± 9.3	28	0	100%
P4	18	f	8	23 ± 6.4	29	12	58.62%

Manifestations of OCD are very heterogeneous across patients, which reflected in the four patients investigated in this work.

P1 showed a manifestation of OCD caused by the thought that certain objects are contaminated, triggering a strong feeling of disgust. To neutralise, these objects as well as both hands were cleaned excessively. Exposure exercises involved physical contact with "contaminated" objects such as doorknobs or contaminating "clean" objects such as one's bed. Subjectively, P1 reported a beneficial impact of the treatment, even though numerical symptom reduction was minor.

P2 showed the a repetition compulsion, repeating actions until they felt "just right", and a counting compulsion, mentally reciting certain number sequences over and over. Exposure for P2 included performing certain actions only once, e.g., closing the lid of a pen, and writing down a number included in the number sequence several times without mentally reciting the entire sequence.

P3 showed a contamination-based manifestation of OCD accompanied by the urge to perform frequent hand-washing. Exposures mainly consisted of touching "disgusting" objects like glue, the bathroom sink or toilet bowl. P3 started

2 Method

2.1 Study Details

Eye tracking data for this work was collected within the SSTeP KiZ study [13]. In this study, different sensor modalities were integrated into video-based CBT for children and adolescents with OCD, allowing patients to receive treatment in their home environments. An overview of the sensor-assisted therapy setup can be seen in Fig. 1. The sensor modalities included eye tracking, heart rate monitoring and hand movements, which have been shown to be promising candidates for measuring stress and compulsive behaviour [29]. The procedure was approved by the local ethics committee (877/2020BO1). In this work, we will focus on the eye tracking recordings of four patients from this study.

Fig. 1. Patients were equipped with a system to record their therapy sessions in their home environment. The data was streamed to the therapist UI, where the therapist had access to the egocentric video including gaze estimation and physiological measures. All icons are attributed to Flaticon.com

Treatment consisted of 14 session of video-based CBT. There was no exposure exercise within the first four sessions, since these were dedicated to building a therapist-patient relationship and psychoeducation in preparation for E/RP exercises. Afterwards, the amount of sessions including an exposure exercise was dependent on condition and therapy progress of the individual patient.

2.2 Data Collection and Labelling

The software architecture to record, transmit and display sensor data was custom designed for the purpose of the study [21]. Sensor data was recorded and synchronised locally and streamed to a therapist User Interface (UI), where the therapist could access the egocentric video of the patient together with the current gaze estimation and physiological parameters.

The therapist UI additionally served as a platform for data labelling where the therapists tagged time points defining the course of the therapy session, including start and end point of the exposure exercise. Throughout therapy, patients were asked to rate their perceived stress level on a scale from 0 to 10, which was also provided as label through the therapist UI.

velocity or gaze dispersion unsuitable for mobile eye tracking [15,27]. We therefore adapt the method from [27] that detects fixations based on the similarity of small regions around each gaze location within the visual scene.

1.2 Eye Tracking in Patients with OCD

Eye tracking is being increasingly used to study attention of patients with psychiatric disorders [3], including OCD [4,6,19]. The recording paradigm closest to real-life gaze behaviour is the free-viewing task in which patients are shown a neutral and an OCD-related stimulus at the same time without instructions on where to look. Studies using the free-viewing paradigm have shown evidence for a maintenance bias [4]: Patients with OCD show sustained attention towards OCD-related stimuli resulting in an increased number and duration of fixations on these stimuli [6,19].

An important marker during exposure exercises is the perceived stress triggered by the confrontation with the obsession. In healthy subjects, several studies investigated the effect of stress on gaze behaviour during real-life situations, e.g., flight simulation [30], interview settings [5] or the work day of ICU nurses [2]. These studies have shown that with increased stress the fixation duration drops [5,30] or, similarly, the number of fixations increases [2].

In these studies, stress has usually been induced by a time-pressured increase in mental effort to create anxiety. During exposure sessions, in contrast, stress is induced by confrontation with the obsessive thought without time constraints. Studies in healthy subjects have shown that an increase in mental effort without time pressure leads to longer fixation duration [8,16] and sometimes a concurrent increase in variability of fixation duration [22]. Longer fixations are thought to reflect the narrowed but increased attention allocated to the fixated goal [16] and therefore to be dependent on the task type [8].

Due to the lack of methods to analyse fixation behaviour in patients with OCD during real-life situations, the study of attention in said patients has been constrained to laboratory settings. While the effect of stress and mental effort on fixation behaviour in real-life situations has been well studied for healthy subjects, studies on their influence on real-life gaze behaviour in patients with OCD are still missing.

In this work we propose a method to detect and cluster fixations in mobile eye tracking during real-life therapy sessions, and analyse the effect of the subjective stress levels on fixation behaviour. Based on existing evidence for maintenance bias from laboratory eye tracking studies and the effect of mental effort on fixation behaviour, we hypothesised that attention on exposure-relevant locations increases with rising stress levels, showing in a higher fixation count and longer fixations towards these locations. If the predominant influence on gaze behaviour are stress and anxiety, the number of fixations onto other objects should also increase but fixations should become shorter.

our method for analysing natural gaze behaviour during exposure sessions. The fixation analysis shows that patients allocate more attention towards exposure-related objects under higher stress levels, suggesting higher mental load. As such, providing feedback on fixation behaviour holds significant promise to support therapists in monitoring intensity of exposure exercises.

Keywords: mobile eye tracking · obsessive-compulsive disorder · sensor-assisted therapy · exposure exercises · real life gaze behaviour

1 Introduction and Related Work

Digital health interventions are becoming increasingly important to ensure that affected patients have easy access to treatment and to personalize said treatment. Video-based online therapy has proven its effectiveness in treating anxiety [11], including obsessive-compulsive disorder (OCD) [12]. OCD is a psychiatric disorder characterized by a combination of obsessions and compulsions [1]. Obsessions are repeated intrusive thoughts or urges that cause anxiety, whereas compulsions are mental or behavioural repetitions or rituals that the person with OCD feels the urge to do in order to reduce anxiety [33]. In children and adolescents, OCD affects around 0.5–4% of the population [9,10] and if not treated in young age, patients are at high risk to develop chronic symptoms [28].

The state-of-the-art treatment for OCD is cognitive behavioural therapy (CBT) based on exposure and response prevention (E/RP) sessions [1]. In the E/RP sessions, patients confront their obsession whilst refraining from the urge to perform the compulsion and instead enduring the anxiety until it, in the desired case, reduces without the compulsion. Given that obsessions are often linked to specific places or objects, which can be challenging to recreate in clinical environments, moving therapy into patients' homes through video-based CBT allows for a more direct confrontation with the obsession, thereby improving the effectiveness of the treatment [25].

A key limitation of such video-based approaches for CBT is the limited field of view of the web camera. Here, eye tracking devices can add valuable information about what is within sight of the patient, what the patient is focusing on at any point in time, and emerging gaze behaviour throughout therapy.

1.1 Fixations in Mobile Eye Tracking

What constitutes a fixation in mobile eye tracking differs from what is typically considered a fixation in laboratory-based eye tracking research [15,27]. There, fixations are temporal episodes during which the eyes remain stationary. In mobile eye tracking this definition is too narrow given that neither the body nor the head is necessarily still: Gaze may follow a moving object (smooth pursuit) or the head moves while the eyes keep fixating a still object. Thus, a fixation is typically defined as an episode during which gaze stays on a fixed object or location. This difference renders common methods based on eye movement

Gaze Behaviour in Adolescents with Obsessive-Compulsive Disorder During Exposure Within Cognitive-Behavioural Therapy

Annika Thierfelder[1(✉)], Björn Severitt[2], Carolin S. Klein[3], Annika K. Alt[3], Karsten Hollmann[3], Andreas Bulling[4], and Winfried Ilg[1]

[1] Hertie Institute for Clinical Brain Research, Section for Computational Sensomotorics, University Hospital Tübingen, Tübingen, Germany
{annika.thierfelder,winfried.ilg}@uni-tuebingen.de
[2] Department of Computer Science, Human-Computer Interaction, University of Tübingen, Tübingen, Germany
bjoern.severitt@uni-tuebingen.de
[3] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital Tübingen, Tübingen, Germany
{carolin.klein,annika.alt,karsten.hollmann}@med.uni-tuebingen.de
[4] Institute for Visualisation and Interactive Systems, University Stuttgart, Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

Abstract. Digital health interventions that involve monitoring patient behaviour increasingly benefit from improvements in sensor technology. Eye tracking in particular can provide useful information for psychotherapy but an effective method to extract this information is currently missing. We propose a method to analyse natural gaze behaviour during exposure exercises for obsessive-compulsive disorder (OCD). At the core of our method is a neural network to detect fixations based on gaze patch similarities. Detected fixations are clustered into *exposure-relevant*, *therapist*, and *other* locations and corresponding eye movement metrics are correlated with subjective stress reported during exposure. We evaluate our method on gaze and stress data recorded during video-based psychotherapy of four adolescents with OCD. We found that fixation duration onto *exposure-relevant* locations consistently increases with the perceived stress level as opposed to fixations onto *other* locations. Fixation behaviour towards the *therapist* varied largely between patients. Taken together, our results not only demonstrate the effectiveness of

This work is funded by the German Federal Ministry of Health (BMG) project SSTeP KiZ (2520DAT700) and the European Research Council (ERC SYNERGY Grant RELEVANCE). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting A. Thierfelder. A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 3–17, 2024.
https://doi.org/10.1007/978-3-031-59717-6_1

Pervasive Mental Health

Developmental Evaluation of an e-Counselling and Learning Application for Parents of Children with Attention Deficit Hyperactivity Disorder 520
Andrea Kerschbaumer, Lisa-Sophie Gstöttner, Erna Schönthaler, Károly Szabó, Peter Putz, Carina Hauser, and Franz Werner

Author Index 525