

Keynote: Healthy and Sustainable Food Recommendations Exploiting Natural Language Processing and Large Language Models

Cataldo Musto¹

¹University of Bari, Italy

Abstract

The growing focus on healthy and sustainable eating requires innovative tools to support personalized recommendations. This talk will examine how Natural Language Processing (NLP) and Large Language Models (LLMs) can enhance food recommendation systems, enabling them to provide personalized suggestions that promote both individual well-being and environmental responsibility. In particular, we will first show the effectiveness of knowledge-aware recommendation models that encode information about healthy food consumption. Next, we emphasize the importance of natural language processing techniques, which can be used to nudge toward healthier food choices through automatically generated explanations. Finally, we will show recent advances aiming also to include the concept of sustainability in the design and development of conversational food recommenders. In particular, we will discuss a pipeline based on LLMs that identifies healthier and more sustainable food alternatives that can be suitable for the user. We will conclude the presentation by sketching several future directions of this exciting research line.

Keywords

Health recommender systems, Large Language Models, healthy living, health and care, Recommender Systems

Bio

Cataldo Musto is an Associate Professor at the Department of Computer Science, University of Bari. His research focuses on the adoption of NLP, LLMs, and semantic content representation strategies in knowledge-aware recommender systems and AI algorithms. He authored around 90 scientific articles, and he is one of the authors of the textbook “Semantics in Adaptive and Personalized Systems: Methods, Tools and Applications”, edited by Springer. He is also involved in the organization of conferences such as ACM UMAP and ACM RecSys as Student Volunteers Chair in 2019, Social Chair in 2020, Poster and Demo Chair in 2022, Doctoral Symposium Chair in 2023, and Workshops Chair in 2024. In 2025, he will be the Program Chair of ACM UMAP conference. Since 2016, he has given several tutorials at UMAP and ESWC conferences about the exploitation of semantics-aware representation in content-based personalized systems. Since 2019, he has organized a series of workshops on Explainable User Modeling (ExUM), Knowledge-aware Recommender Systems (KARS) and Explainable Artificial Intelligence (XAI).

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Tailoring Health: Contextual Variables In Health Recommender Systems*

Felix Reinsch^{1,*}, Thure Georg Weimann¹ and Jeannette Stark¹

¹TUD Dresden University of Technology, Dresden, Germany

Abstract

hrs have emerged as a crucial tool in personalized healthcare, offering tailored recommendations to promote healthy behaviors and prevent diseases. The effectiveness of these systems hinges on their ability to accurately personalize recommendations based on contextual variables. This research investigates the contextual variables currently employed by Health Recommender Systems (HRSs), addressing two key research questions: (1) Which contextual variables are currently used in HRSs? and (2) How can these variables be categorized? Through an extensive systematic literature review, we identified 24 commonly utilized contextual variables across existing HRSs. To provide a structured approach for understanding, we organized the variables with a framework that classifies contextual variables into four distinct categories: objective-static, objective-dynamic, subjective-static, and subjective-dynamic. Our findings highlight the diverse yet uneven distribution of these variables within the framework, emphasizing the need for a balanced consideration of both objective and subjective data in developing comprehensive HRSs. The proposed framework serves as a robust foundation for future advancements, aiming to enhance the personalization capabilities of HRSs and ultimately improve health outcomes.

Keywords

Health Recommender Systems, Personalized Medicine, Contextual Variables, Data Categorization, Context-Aware Systems

1. Introduction & Background

Noncommunicable diseases (NCDs), including diabetes, obesity, and cardiovascular diseases, represent the leading cause of mortality, contributing to 74% of deaths worldwide [1, 2]. A substantial number of these deaths are linked to lifestyle choices, with over 20% being related to dietary habits, such as the consumption of sugar-sweetened beverages, which elevate metabolic risks [3]. Preventing NCDs and related deaths requires individuals to adopt lifestyle changes [4].

Previous studies have shown that personalized mobile health interventions can be an effective approach for facilitating habit formation and, consequently, promoting lifestyle changes in individuals [5]. The effectiveness of these interventions is largely driven by their ability to tailor recommendations to the individual, a capability often enabled by Recommender Systems (RSs) [6, 7, 8, 9].

These personalized recommendations are achieved through a comprehensive understanding of the user, informed by the contextual variables collected by the system. This precise tailoring enhances the relevance of the recommendations to the individual user, leading to higher user engagement. As a result, there is a greater likelihood that the user will adopt the recommended behavior changes into daily life, fostering healthier habits and ultimately contributing to the prevention of NCDs [10, 11].

Health Recommender Systems (HRSs) serve as a prominent example of personalized (mobile) health interventions. The effectiveness of HRSs is closely tied to their ability to deliver personalized recommendations tailored to the individual user [12, 13]. Therefore, it is crucial to examine the contextual variables that these systems collect to enable such personalization [11, 14]. Understanding which variables are gathered and how they are utilized can provide valuable insights into the factors that drive

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ felix.reinsch@mailbox.tu-dresden.de (F. Reinsch); thure.weimann@tu-dresden.de (T. G. Weimann); jeannette.stark@tu-dresden.de (J. Stark)

🌐 www.felixreinsch.de (F. Reinsch)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the accuracy and relevance of recommendations. This research investigates the contextual variables currently employed by HRSs and, hence, addresses the following research questions 1:

RQ1: Which contextual variables are currently used in HRSs?

Furthermore, we aim to gain deeper insights into the role and relevance of these variables for personalized recommendation in general, as well as, for specific health goals such as improving mental health or sleep as well as weight loss. This endeavor is reflected in RQ2:

RQ2: How can the currently used variables be categorized?

To answer RQ1, we conducted a systematic literature review, identifying 24 commonly utilized contextual variables within the existing HRSs landscape. To address RQ2, we propose a framework that classifies both existing contextual variables and those that may be identified in the future into four distinct categories. This framework aims to provide a structured approach for understanding and organizing the variables that influence the personalization capabilities of HRSs in general. For specific recommendation types, we further display the distribution of contextual variables for different recommendation types (e.g., healthy diet, mental health, sleep, weight loss).

2. Method

We conducted a systematic literature review in accordance with the PRISMA statement, complemented by a forward and backward citation search [15, 16]. The search strings for this systematic literature review were developed collaboratively by all authors. Key concepts central to the study, such as habits, health, personalization, and RSs, were identified, with a focus on contextual variables supporting healthier behavior change. These core concepts guided the selection of relevant terms, which were refined through multiple iterations to ensure comprehensive coverage of the literature. The final search strings, validated by all co-authors, were applied across PubMed, Scopus, and Web of Science—databases. Notably, Scopus was queried with two distinct search strings due to the initial search returning only two sources. The search strings and respective databases are presented in Table 1 of the [online appendix](#).

The systematic literature review process was facilitated by the online tool CADIMA to ensure reproducibility [17, 18].

The review exclusively incorporated peer-reviewed, open accessible English-language journal articles, conference proceedings, and detailed project descriptions. Studies that did not investigate RSs, did not personalize recommendations or suggestions in a health-related or habit-changing context, or did not indicate input variables were excluded. The context was considered health-related if it encompassed promoting physical, mental, or emotional well-being. Habit-related contexts were defined as those aiming to influence or modify behaviors and routines to improve overall health outcomes.

A graphical representation of the literature selection process is shown in Figure 1. In total, $n = 414$ results were obtained (391 after duplicate removal) and initially screened based on the title, abstract, and exclusion criteria mentioned above. After the full-text screening, 48 articles were considered eligible. Among the 48 sources, seven were identified via PubMed, two via Scopus, and eleven via the Web of Science database. Consequently, 48 articles were finally included in the review. From these articles, 267 variable instances were identified. These instances denote individual mentions of a variable; for example, the variable “Age” might be cited multiple times across different papers, with each occurrence contributing to the total of 267 instances. Consequently, 24 unique variables were extracted.

Following the systematic literature review, the authors organized the identified variables into the framework, presented in Figure 2, which comprises four distinct categories. Moreover, the authors classified the identified papers from the literature review into eight distinct categories based on their application. The initial categorization of variables and papers was conducted by one author, after which the other authors reviewed it, leading to a multilateral discussion that refined and finalized the categorization.

Figure 1: Systematic Literature Review Flow Diagram

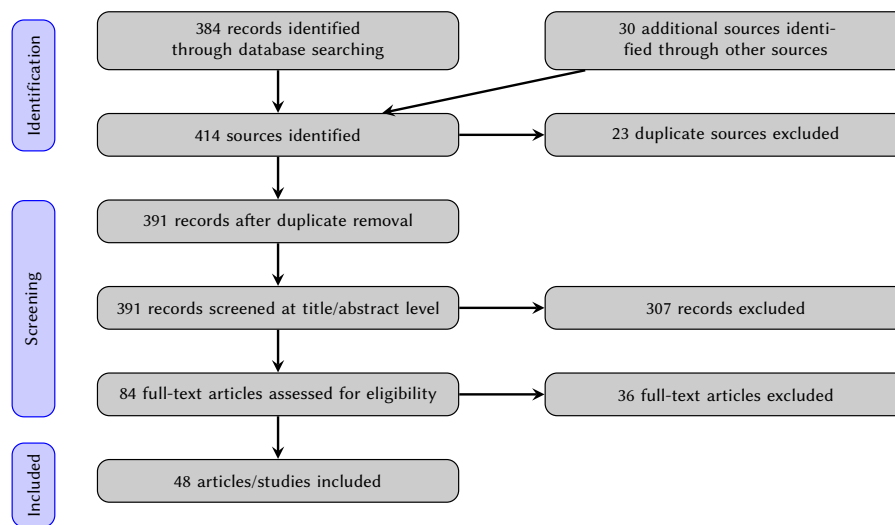
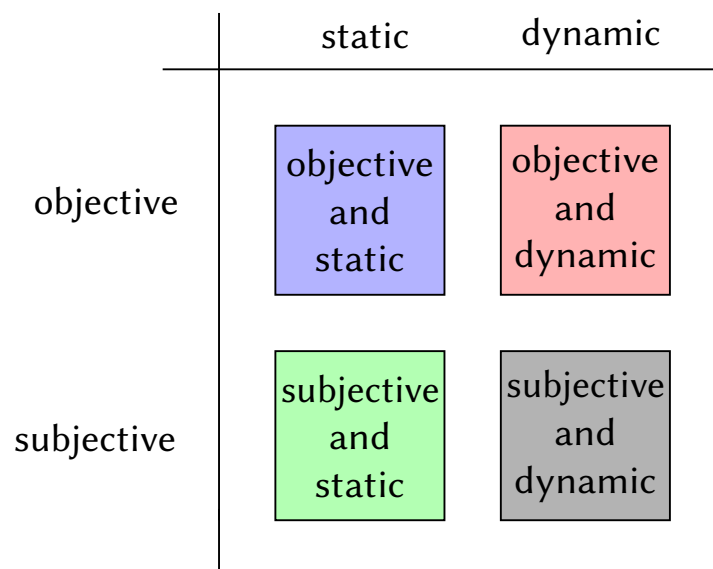


Figure 2: Framework for Structuring Data Collection in HRSs Development



3. Results

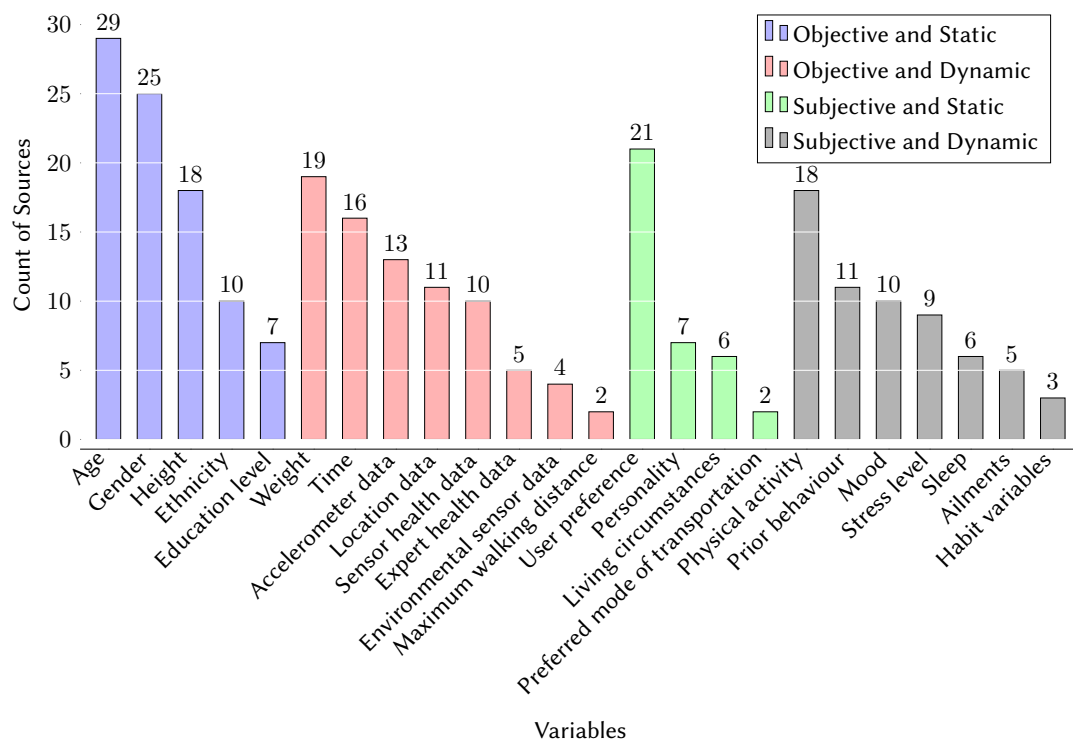
To facilitate a straightforward analysis, the authors propose a framework delineated by two principal axes: “objective-subjective” and “static-dynamic”. This segmentation builds upon previous unidimensional approaches used to organize variables for HRSs, extending them into a second dimension to provide a more nuanced categorization [19, 20]. The vertical axis is grounded in the work of [19], who categorized data collection mechanisms into passive and active sensing, which can be adapted to the objective subjective spectrum. Objective data, such as accelerometer readings, is gathered without direct user input, while subjective data captures users’ feelings or opinions, which are not directly measurable as external inputs. The horizontal axis of the framework draws on the concept of temporal context introduced by [20], distinguishing between static and dynamic variables. For instance, variables like gender may remain constant, while others, such as a user’s weight, can vary over time. The colors chosen in Figure 2 symbolize the same categories as depicted in Figure 3.

The framework distinguishes between four different categories of contextual variables: First, there is *objective and static*, which includes quantifiable and unchanging variables, such as gender or ethnicity.

Second, *objective and dynamic* refers to quantifiable variables that fluctuate over time, like weather data or measurements from health sensors, such as smartwatches. Third, *subjective and static* encompasses personal and consistent preferences or perceptions, such as personality traits. Finally, *subjective and dynamic* involves personal factors that are both individual and variable, including mood or current stress levels. We have classified the variables into five within objective and static, eight within objective and dynamic, four within subjective and static, and seven within subjective and dynamic.

In Table 2 of the [online appendix](#), the 24 distinct variables identified from the systematic review and the frequency of occurrence across the 48 analyzed papers are displayed. The frequency of variable references varies as shown in Figure 3, with “Age” being the most frequently cited variable at 29 references, whereas “Preferred mode of transportation” and “Maximum walking distance” are the least cited with only two references. This diversity in variables encompasses a broad spectrum, ranging from demographic data and physical conditions to mental and emotional states, as well as environmental data such as weather conditions measured by environmental sensors. This highlights the extensive and multifaceted nature of the data considered in HRSs.

Figure 3: Presence of Variables for HRSs in Literature



The heatmap in Table 1 illustrates the distribution of variables across different recommendation types. Each entry in the table represents the absolute frequency with which a particular variable is considered for a given recommendation type, with darker shades indicating higher frequencies. Additionally, the percentage column displays the proportion of each recommendation type relative to the 48 analyzed papers.

Table 1 reveals several insights. Firstly, the majority of sources fall within the categories of physical activity recommendations or the broader domain of dietary recommendations. As expected, variables closely tied to specific recommendation types, such as “Physical activity” reported by the user, are most commonly found in their corresponding RSs. Interestingly, however, variables that do not have an obvious connection to a particular recommendation type are also utilized across various recommendation types. For example, variables that measure a user’s stress level are utilized not only in mental health and stress management RSs but also in physical activity and weight loss recommendations, as stress levels can directly impact a user’s weight [21]. Secondly, some variables appear to be more prominent

Table 1
Distribution of Recommendation Types

Recommendation Type	Percentage	Variables																									
		Accelerometer data	Age	Ailments	Education level	Environmental sensors	Ethnicity	Expert health data	Gender	Habit parameters	Height	Living circumstances	Location data	Maximum walking distance	Mood	Personality	Physical Activity	Preferred mode of transportation	Prior behaviour	Sensor health data	Sleep	Stress level	Time	User preference	Weight		
Healthy Diet and Physical Activity Promotion	8.33%	0	3	1	1	1	1	0	3	0	4	1	1	1	0	0	1	0	1	0	1	0	1	4	4		
Healthy Diet	8.33%	0	3	2	0	0	1	1	3	0	2	0	0	0	0	1	1	0	0	1	0	0	0	2	2		
Healthier Lifestyle (Unspecific)	8.33%	0	3	0	1	0	2	1	3	1	1	1	0	0	0	2	1	0	1	1	0	0	2	3	1		
Mental Health & Stress Management	25.00%	4	3	0	2	1	1	1	3	1	0	0	5	0	6	2	1	0	3	4	3	7	6	2	0		
Physical Activity	27.08%	6	8	2	2	1	3	1	6	0	5	2	3	0	4	2	8	2	2	0	0	1	4	4	6		
Smoking Cessation	8.33%	0	4	0	0	0	1	1	4	1	1	2	1	0	0	0	1	0	2	1	0	0	0	3	1		
Sleep	4.17%	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	2	0	1	2	1	1	1	1	0		
Weight Loss	10.42%	2	5	0	1	0	1	0	3	0	5	0	0	0	0	0	3	0	1	1	1	0	2	2	5		

overall. For example, “Age” and “Gender” are consistently important across all categories, while “Stress level” is particularly prevalent in mental health and stress management RSs.

4. Discussion

There is considerable variation not only in the number of contextual variables employed across different sources but also in the way these variables are utilized. A prominent observation is the preference in current HRSs for collecting objective rather than subjective data. Objective variables represent 169 out of the 267 variable instances, accounting for 63.3%. This preference likely stems from the relative ease of collecting objective data. For instance, when acquiring accelerometer data, timestamps are often included and can be obtained with minimal effort, whereas collecting data on a user’s personality may necessitate substantial resources, such as administering comprehensive questionnaires [22]. This trend is corroborated by the findings of [23], who also noted a diminished use of variables associated with mental health in HRSs, pointing to a potential gap in the integration of psychological and emotional factors. In particular, the integration of subjective data beyond the user’s preferences and prior behavior is comparatively scarce. These subjectively collected factors are, however, essential for personalizing recommendations to align with individual mental states and are difficult to objectively collect via sensors.

Yet, there is a discernible increase in efforts to integrate subjective measures, particularly psycho-

logical variables, into RSs. Notably, initiatives such as Zenspace and Studentlife have pioneered the incorporation of mental health metrics into their platforms [24, 25]. Additionally, the study by [23], which includes a personality measure within their RS, underscores a growing trend toward the adoption of subjective data in these systems.

The analysis further indicates a wide-ranging but imbalanced use of variables in the literature, emphasizing the necessity of comprehensive inclusion in HRSs. The prevalence of certain variables points to their perceived importance and utility in the field, while the less frequently cited variables may represent opportunities for innovation in HRSs research and practice. Given this context, our proposed framework offers a practical step for researchers to evaluate and enhance their HRSs. By assessing the variables they currently use against the framework's categories, researchers may realize that they are predominantly utilizing variables from a single category. This awareness encourages them to reconsider their variable selection, exploring the inclusion of variables from other categories that could be beneficial. For example, an HRS that primarily relies on objective, static data might be significantly improved by integrating subjective or dynamic variables, thereby capturing a more comprehensive picture of the user's context. This approach not only enriches the personalization capabilities of the system but also addresses the identified imbalance in variable utilization, contributing to the development of more effective and user-centric HRSs.

5. Limitations

Not all 24 identified variables are likely to be equally essential for improving the quality of recommendations. A similar conclusion was reached by [26], who distinguished between various observed contextual features in their study. Variables that do not improve the quality of recommendations can complicate compliance with privacy regulations and increase the burden on users (e.g., due to repetitive subjective assessments), ultimately leading to inefficiencies. It also increases the demand for computing power, particularly when the HRS runs on a local device, and heightens the risk of data breaches. Therefore, there is a clear need to adopt a more judicious approach to data collection [27]. This cautious stance is critical to ensure compliance with legal standards and to protect individual privacy, particularly when it is uncertain whether the collected data will provide significant utility for the application or study. This guideline acts as a safeguard, suggesting that data should only be gathered if it is both essential for the intended purpose and can be securely managed.

Moreover, the variables identified in this study are not exhaustive and vary widely in their definitions and applications, which could affect the generalizability of the findings. For example, the variable "Time" can be recorded in various formats, such as absolute (e.g., "8 AM") or relative (e.g., "after lunch"). This variability may lead to inconsistencies in how data is interpreted across different systems. Similarly, subjective variables like "Personality" could yield different results depending on the used scales, affecting the comparability and reliability of the data collected.

Furthermore, the distinction between static and dynamic variables in health-related applications can vary based on the specific context and time horizon of the application. For example, weight may be considered a static variable in contexts involving habitual behaviors, such as reading or brushing teeth, where it is not expected to change significantly. In contrast, in applications focused on weight loss or fitness tracking, weight becomes a dynamic variable that changes over time. This demonstrates how the categorization is influenced by the time horizon and targeted outcomes of the application; longer time horizons can lead to more variables being classified as dynamic due to their potential for change. Finally, while variables such as gender and education level are often considered static due to their relative stability over time, they are not inherently unchangeable. Therefore, recognizing that these variables can also change suggests that distinguishing between variables that are generally static and those that are dynamic, depending on context and time horizon, could enhance the robustness of the framework.

6. Future research

The relevance of HRSs is expected to increase in the near future, particularly for applications in digital therapeutics [28, 19]. This is especially pertinent in light of upcoming regulations, such as the European Union's Artificial Intelligence Act, which will likely shape the development and implementation of AI-driven healthcare solutions.

One critical avenue for future research is determining which contextual variables contribute the most to effective recommendations for each HRS type. While our framework provides valuable insights and a structured approach for evaluating and categorizing variables, the specific impact of each variable on recommendation accuracy and user experience were not investigated. However, our categorization could provide a starting point for this deeper meta-analysis. Addressing this gap involves analyzing the effectiveness of individual variables across different HRSs to identify those that significantly enhance performance. This focus not only helps optimize the recommendation process but also prevents the unnecessary overcollection of information, thereby respecting user privacy and adhering to data minimization principles.

Another important focus for future work is understanding how these variables are currently measured and reported. Consistency and reliability in data collection are essential for the scalability and practical implementation of HRSs. Our review indicates that variables are measured using diverse methodologies, which can lead to inconsistencies and hinder comparability across studies. For instance, the variable "Personality" can be measured through various personality scales, each differing in relevance and predictive utility for HRSs outcomes. Conducting a comparative analysis of these measurement methods would help identify the most appropriate tools for data gathering. Establishing standardized measurement approaches would bridge the gap to the technological aspects of HRSs, facilitating the integration of these variables into system designs and promoting interdisciplinary collaboration. Likewise, our work may represent a further step toward the integration of HRSs into low-/no-code development platforms for mHealth applications. While first platforms emerged in recent years, providing generic mechanisms for easily developing (self-)adaptive interventions is still an underexamined aspect [29, 30]. By identifying important variables in HRSs across different health domains, our work may provide further guidance in this regard. In summary, future research should focus on pinpointing the key variables that most significantly enhance recommendation effectiveness for each HRS type and on standardizing the methods used to measure and report these variables. This dual emphasis will improve the practical applications of HRSs and contribute to the ethical and secure management of personal health data. Ultimately, such efforts will advance the prevention and management of NCDs through more personalized, engaging, and effective health interventions.

References

- [1] WHO, Non communicable diseases, 2023. URL: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>.
- [2] WHO, The top 10 causes of death, 2020. URL: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>.
- [3] GBD 2017 Diet Collaborators, Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017, *Lancet* (London, England) 393 (2019) 1958–1972. doi:10.1016/S0140-6736(19)30041-8.
- [4] H. Aboumatar, S. Pitts, R. Sharma, A. Das, B. M. Smith, J. Day, K. Holzhauer, S. Yang, E. B. Bass, W. L. Bennett, Patient engagement strategies for adults with chronic conditions: an evidence map, *Systematic Reviews* 11 (2022) 39. URL: <https://doi.org/10.1186/s13643-021-01873-5>. doi:10.1186/s13643-021-01873-5.
- [5] Y. Wei, P. Zheng, H. Deng, X. Wang, X. Li, H. Fu, Design Features for Improving Mobile Health Intervention User Engagement: Systematic Review and Thematic Analysis, *Journal of Medical Internet Research* 22 (2020) e21687. URL: <https://www.jmir.org/2020/12/e21687>. doi:10.2196/

- 21687, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [6] F. Ricci, L. Rokach, B. Shapira (Eds.), *Recommender Systems Handbook*, Springer US, Boston, MA, 2015. URL: <https://link.springer.com/10.1007/978-1-4899-7637-6>. doi:10.1007/978-1-4899-7637-6.
 - [7] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: *Proceedings of the 1994 ACM conference on Computer supported cooperative work, CSCW '94*, Association for Computing Machinery, New York, NY, USA, 1994, pp. 175–186. URL: <https://dl.acm.org/doi/10.1145/192844.192905>. doi:10.1145/192844.192905.
 - [8] P. Resnick, H. R. Varian, Recommender systems, *Communications of the ACM* 40 (1997) 56–58. URL: <https://dl.acm.org/doi/10.1145/245108.245121>. doi:10.1145/245108.245121.
 - [9] R. Burke, Hybrid Web Recommender Systems, in: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization*, Springer, Berlin, Heidelberg, 2007, pp. 377–408. URL: https://doi.org/10.1007/978-3-540-72079-9_12. doi:10.1007/978-3-540-72079-9_12.
 - [10] J. Stark, T. Weimann, F. Reinsch, E. Hickmann, M. Kählig, C. Gißke, P. Richter, Building Habits in the Digital Age: Incorporating Psychological Needs and Knowledge from Practitioners to Inform the Design of Digital Therapeutics, *Wirtschaftsinformatik 2023 Proceedings* (2023). URL: <https://aisel.aisnet.org/wi2023/36>.
 - [11] H. Schäfer, S. Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, C. Trattner, Towards Health (Aware) Recommender Systems, in: *Proceedings of the 2017 International Conference on Digital Health, DH '17*, Association for Computing Machinery, New York, NY, USA, 2017, pp. 157–161. URL: <https://doi.org/10.1145/3079452.3079499>. doi:10.1145/3079452.3079499.
 - [12] S. Katariya, J. Bose, V. Mopuru, A. Sharma, S. Tappashetty, A Personalized Health Recommendation System Based on Smartphone Calendar Events, 2018, pp. 110–120. doi:10.1007/978-3-319-94523-1_10.
 - [13] P. Venkatachalam, S. Ray, How do context-aware artificial intelligence algorithms used in fitness recommender systems? A literature review and research agenda, *International Journal of Information Management Data Insights* 2 (2022) 100139. URL: <https://www.sciencedirect.com/science/article/pii/S2667096822000829>. doi:10.1016/j.ijime.2022.100139.
 - [14] R. D. Croon, L. V. Houdt, N. N. Htun, G. Štiglic, V. V. Abeele, K. Verbert, Health Recommender Systems: Systematic Review, *Journal of Medical Internet Research* 23 (2021) e18035. URL: <https://www.jmir.org/2021/6/e18035>. doi:10.2196/18035, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
 - [15] J. Webster, R. T. Watson, Analyzing the Past to Prepare for the Future: Writing a Literature Review, *MIS Quarterly* 26 (2002) xiii–xxiii. URL: <https://www.jstor.org/stable/4132319>, publisher: Management Information Systems Research Center, University of Minnesota.
 - [16] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement., *PLoS medicine* 6 (2009) e1000097. doi:10.1371/journal.pmed.1000097, place: United States.
 - [17] C. Kohl, E. J. McIntosh, S. Unger, N. R. Haddaway, S. Kecke, J. Schiemann, R. Wilhelm, Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools, *Environmental Evidence* 7 (2018) 8. URL: <https://doi.org/10.1186/s13750-018-0115-5>. doi:10.1186/s13750-018-0115-5.
 - [18] C. Kohl, E. J. McIntosh, S. Unger, N. R. Haddaway, S. Kecke, J. Schiemann, R. Wilhelm, Correction to: Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools, *Environmental Evidence* 7 (2018) 12. URL:

<https://doi.org/10.1186/s13750-018-0124-4>. doi:10.1186/s13750-018-0124-4.

- [19] I. Sim, Mobile Devices and Health, *New England Journal of Medicine* 381 (2019) 956–968. URL: <http://www.nejm.org/doi/10.1056/NEJMra1806949>. doi:10.1056/NEJMra1806949.
- [20] G. Adomavicius, B. Mobasher, F. Ricci, A. Tuzhilin, Context-Aware Recommender Systems, *AI Magazine* 32 (2011) 67–80. URL: <https://onlinelibrary.wiley.com/doi/10.1609/aimag.v32i3.2364>. doi:10.1609/aimag.v32i3.2364.
- [21] M. Kivimäki, J. Head, J. E. Ferrie, M. J. Shipley, E. Brunner, J. Vahtera, M. G. Marmot, Work stress, weight gain and weight loss: evidence for bidirectional effects of job strain on body mass index in the Whitehall II study, *International Journal of Obesity* 30 (2006) 982–987. URL: <https://www.nature.com/articles/0803229>. doi:10.1038/sj.ijo.0803229, publisher: Nature Publishing Group.
- [22] A. Buyalskaya, H. Ho, A. Duckworth, X. Li, K. Milkman, C. Camerer, Supplementary Materials for: What can machine learning teach us about habit formation? Evidence from exercise and hygiene (2023).
- [23] K. Eldeswky, F. Elazab, A. E. Bolock, S. Abdennadher, Character-Based Habit Recommender System, in: D. Durães, A. González-Briones, M. Lujak, A. El Bolock, J. Carneiro (Eds.), *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection, Communications in Computer and Information Science*, Springer Nature Switzerland, Cham, 2023, pp. 104–115. doi:10.1007/978-3-031-37593-4_9.
- [24] M. McDaniel, M. Anwar, Zen_space: A Smartphone App for Individually Tailored Stress Management Support for College Students, in: H. Chen, D. D. Zeng, E. Karahanna, I. Bardhan (Eds.), *Smart Health, Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2017, pp. 123–133. doi:10.1007/978-3-319-67964-8_12.
- [25] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, A. T. Campbell, StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones, *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014) 3–14. URL: <https://dl.acm.org/doi/10.1145/2632048.2632054>. doi:10.1145/2632048.2632054, conference Name: UbiComp '14: The 2014 ACM Conference on Ubiquitous Computing ISBN: 9781450329682 Place: Seattle Washington Publisher: ACM.
- [26] I. Coppens, T. De Pessemier, L. Martens, Connecting physical activity with context and motivation: a user study to define variables to integrate into mobile health recommenders, *User Modeling and User-Adapted Interaction* (2023). URL: <https://doi.org/10.1007/s11257-023-09368-9>. doi:10.1007/s11257-023-09368-9.
- [27] R. Bejtlich, New cybersecurity mantra: “If you can’t protect it, don’t collect it”, 2015. URL: <https://www.brookings.edu/articles/new-cybersecurity-mantra-if-you-cant-protect-it-dont-collect-it/>.
- [28] D. Fürstenau, M. Gersch, S. Schreiter, Digital Therapeutics (DTx), *Business & Information Systems Engineering* 65 (2023) 349–360. URL: <https://doi.org/10.1007/s12599-023-00804-z>. doi:10.1007/s12599-023-00804-z.
- [29] S. Liu, H. La, A. Willms, R. E. Rhodes, A “No-Code” App Design Platform for Mobile Health Research: Development and Usability Study, *JMIR Formative Research* 6 (2022) e38737. URL: <https://formative.jmir.org/2022/8/e38737>. doi:10.2196/38737, company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [30] T. Weimann, C. Gißke, Unleashing the potential of reinforcement learning for personalizing behavioral transformations with digital therapeutics: a systematic literature review, in: *Proceedings of the 17th international joint conference on biomedical engineering systems and technologies - HEALTHINF, SciTePress / INSTICC*, 2024, pp. 230–245. doi:10.5220/0012474700003657, iSSN: 2184-4305.

A. Online Resources

The online appendix for this submission can be found [here](#).

Explaining Decision-Making between Exploration and Repetition: Key Factors for Physical Activity Recommendations*

Ine Coppens^{1,*}, Toon De Pessemier¹ and Luc Martens¹

¹imec - WAVES - Ghent University, iGent - Technologiepark-Zwijnaarde 126, Ghent, Belgium

Abstract

A challenge in promoting physical activity with recommender systems lies in balancing repeat recommendations to create habits, with exploration to prevent boredom. This study aims to identify the key variables influencing users' decision-making between the two. Through the analysis of data from an eight-week Micro-Randomized Trial conducted via a mobile health app, using random forest variable importance measures and SHAP analyses, we identified factors affecting these decisions. Our findings reveal that participants were more likely to explore new activities during the first two weeks of the intervention, in the afternoons and evenings, on Sundays, and when activities involved specific locations or workouts. These findings provide valuable insights into the transition from exploration to repetition, contributing to more effective recommender systems for physical activity promotion.

Keywords

recommender systems, physical activity, decision-making, variable importance analysis, random forest, repetition, exploration,

1. Introduction

Insufficient physical activity (PA) is associated with adverse health outcomes [1]. To support motivation and long-term health benefits, individuals with low activity levels may benefit from engaging in enjoyable activities. Recommender Systems (RSs) can automatically personalize such enjoyable PAs by utilizing a wide range of information about users and their preferences [2, p 9].

As these physically inactive people might not have established healthy habits yet, the RS should repeat some activities to create habits, because repetition makes the behavior more automatic [3], increasing the chances for long-term engagement and positive health outcomes. However, the RS should also provide sufficient opportunity to explore new and varied activities to prevent boredom [4, 5, 6].

In addition to their recommended minimum of 150 minutes of moderate-to-vigorous PA (MVPA) per week, the World Health Organization (WHO) also recommends a mix of aerobic and muscle-strengthening activities [1]. As such, both exploration for variety and repetition for habit formation are important for healthy behavior. The challenge lies in determining when an RS should prioritize one approach over the other.

This study focuses on identifying the key variables that impact users' decision-making process to choose either an exploration or repetition PA item, arguing that future RSs can integrate these variables in their own decision-making processes as well. The research question is:

RQ: Which variables affect users' decision to choose an exploration or repetition PA recommendation, and in which conditions are they more likely to choose one over the other?

By analyzing data from an eight-week Micro-Randomized Trial (MRT) conducted with a mobile health (mHealth) app, we aim to uncover the most important factors, determined with Random Forests' (RFs) variable importances [7] and SHapley Additive exPlanations (SHAP) [8] analyses. RFs are a machine

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ Ine.Coppens@UGent.be (I. Coppens); Toon.DePessemier@UGent.be (T. De Pessemier); Luc1.Martens@UGent.be (L. Martens)

ORCID: 0000-0002-3051-506X (I. Coppens); 0000-0002-3920-7346 (T. De Pessemier); 0000-0001-9948-9157 (L. Martens)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

learning technique used for classification or regression to predict an outcome variable based on several input variables by aggregating the results from multiple decision trees. RFs are widely used for data exploration and understanding using variable importance measures [7].

This paper merges two study waves: one that ran from October 2023 until January 2024 (started in autumn), and one from March until June 2024 (spring). Our previous research was solely conducted on the data from the study starting in autumn [9]. To increase participation and dataset, we replicated the study in spring and combined both study's data. In a previous paper [10], we already conducted a first analysis on the subjective perceptions and preferences measured with star rating of the repetition vs. exploration PA recommendations. The current paper dives deeper in the decision-making of the users using a set of factors that might influence their decision, contributing to the design of more effective RS algorithms in the PA domain.

The rest of the paper is organized as follows. Section 2 covers previous work on RFs in the domain of PA. The methods are discussed in Sect. 3. Next, the results and their discussion are elaborated in Sect. 4, followed by the conclusion in Sect. 5.

2. Related work

Previous work with RFs in the domain of PA mainly focused on the classification and recognition of the activity using sensors, such as accelerometers [11, 12, 13]. In those RFs, the input data consists of the data collected from the sensors and the predicted outcome variable is the detected type of PA. In our study, the input data is a combination of manual information (e.g., the self-reported motivation of the user at that time) and automatic information about the context (e.g., the weather).

Other studies used PA as an input variable in the RF to predict a certain outcome variable. In [14], body mass index and depressive symptoms are predicted with sedentary time and MVPA. Another study investigated severity of menopausal symptoms with an RF regression and found that PA level was in the top four of variable importances [15]. In [16], COVID-19 death rates were estimated with 29 socioeconomic and health-related factors with an RF.

PA amount was also used as outcome variable in several studies. Meeting the guidelines for sufficient aerobic PA in a target group of adults with autism was assessed in [17]. Another study found that greenery in the streets impacts the duration of light intensity PA the highest for older adults [18]. In [19], an RF algorithm performed best to predict the probability of achieving a daily steps threshold. In our study, the predicted outcome class is whether the user decides to engage in an exploration or repetition item, thus resulting in a binary classification.

Previous work has also investigated the exploration and repetition decision in next basket recommendations [20] and sequential recommendations [21]. While these are mainly focused on e-commerce and grocery shopping, our study is situated in the health domain and PA promotion with mHealth interventions.

3. Methods

An Android app was created to display the personalized PA recommendations as shown in Fig. 1. Our PA dataset is based on activities from the compendium of physical activities [22], for which we distinguish between *workout* and *location* PA types. In our dataset, we also integrated *general* PA tips from the Belgian website for health (www.gezondleven.be), which contains small-effort activities people can integrate in all four PA situations of their daily life (*free time*, *during work*, *household task*, *active transport*). The resulting dataset contains 237 PA items, connected to the corresponding PA type and situation. For the content-based RS algorithm, all PA items were manually extended with 24 binary attributes to describe their content. For a full implementation description of content-based RS algorithm, we recommend reading our previous paper [23]. The exact same RS algorithm was used in this study, with an adjustment of the dataset and output list that now shows six recommendations.

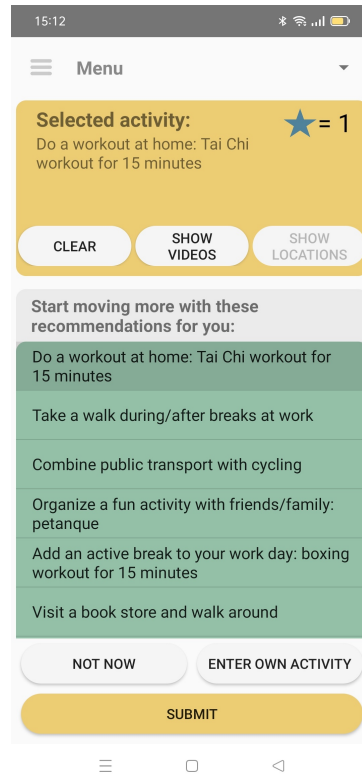


Figure 1: The main screen of the app displays six recommendations in a random order, corresponding to the six combinations of exploration versus repetition and the PA type.

A list of six items was chosen because six combinations can be made in our MRT study design: 2 from the outcome variable (repetition vs. exploration) x 3 from the PA type (general vs. location vs. workout). A repetition item refers to a personalized recommendation for an activity that the user has previously submitted during the study, while an exploration item is a personalized recommendation for an activity that the user has not yet submitted. By randomly positioning each of the six combination in the list for every participant at every delivery time, the micro-randomization of this study is achieved [24] and position bias effects are prevented. As a result, the user can freely choose between a repetition or exploration item of all three PA types at any time.

We recruited healthy adults (between 18 and 65 years old), who have less than 150 min of MVPA per week, via the Sona Platform and Facebook groups for paid studies of Ghent University. For both study waves (starting in autumn and spring), participants were asked to use the app for at least eight consecutive weeks, after which they receive 30 EUR. To prevent false submissions about their PA behavior, participants were informed that they do not receive more money for submitting more PAs, but only qualify for the incentive if they actively use the app for eight weeks and complete all questionnaires. At the start and after eight weeks, they answer the pre-test and post-test questionnaire, respectively, both containing the European Health Interview Survey - Physical Activity Questionnaire (EHIS-PAQ) to measure weekly PA [25] and a question about their age group (18 – 44 or 45 – 65 years). The study received ethical approval from the Ethical Committee (www.ugent.be/pp/en/research/ec) on August 22, 2023 (reference number: 2023-061A).

An RF classifier is tuned with Randomized Search and trained using scikit-learn (scikit-learn.org/) version 1.5.1. Table 1 provides an overview of all 15 input variables of the RF model. The predicted output of the RF is whether the user engaged in a repetition or exploration item (binary classification). SHAP provides an explanation for this output by assigning feature importance values to each variable for a specific prediction [8]. For the variables with the highest feature importance, we apply the dependence plot function from SHAP package version 0.46.0 to show the relation between the input variable and the corresponding SHAP value for the RF's prediction [8]. For an additional statistical analyses on the

EHIS-PAQ answers, SPSS Statistics v. 28 is used for the analyses with a Linear Mixed Model (LMM) using the MIXED procedure [26].

Table 1

We put 15 input variables in the RF, which can be categorized in eight categories and originate from various sources.

category	type	input variables for the RF	source
Time	continuous	day in study	device's clock
		hour of day	
	categorical	start season: spring / autumn	
		day of week	
Weather	continuous	weekend / week day	GPS + weather API (OpenWeatherMap)
		outdoor temperature	
	categorical	clear sky	
		clouds	
Situation	categorical	rain	dataset
		free time / work / household / transport	
	categorical	general (e.g., walk during breaks) / location (e.g., minigolf) / workout (e.g., pilates)	
		alone / with a buddy	
PA type	categorical	score on 4	dataset
Company	categorical	indoors / outdoors	self-reported
Motivation	continuous	amount of steps already detected that day	self-reported
Location	categorical		self-reported
Step count	continuous		accelerometer

4. Results and Discussion

Of the 62 participants who started the study (100% in the age group of 18 - 44 years), 34 continued for at least eight weeks. Throughout the study, the participants submitted 457 recommended items, of which the amounts are categorized in Table 2.

Table 2

The amount of submits per situation and per PA type, including the percentage of the situation.

situation	PA type	amount	% of situation
transport	exploration	17	94%
	repetition	1	6%
during work	exploration	45	75%
	repetition	15	25%
household	exploration	73	55%
	repetition	59	45%
free time	exploration	153	62%
	repetition	94	38%

Firstly, the hyperparameters of the model were tuned with Randomized Search with 5-fold cross validation and a max depth of 6 to prevent overfitting. The resulting optimal hyperparameters are: n estimators = 493, min samples split = 10, min samples leaf = 10, max features = sqrt, max depth = 6, bootstrap = True, resulting in a train accuracy of .764 and test accuracy of .663. The corresponding feature importances are shown in Fig. 2, in which the time in study showed the highest relevance.

Figure 3 shows the SHAP dependence plots that visualize the relation between the six input variables with the highest feature importances and their corresponding SHAP value. We added a color legend to show possible interaction effects with days in study, as this variable ranked highest on the variable importances.

4.1. The input variables

4.1.1. Day in study

The first dependence plot shows that days in the beginning of the study are associated with a higher prediction for an exploration item (above the y=0 line). The threshold at which the model switches

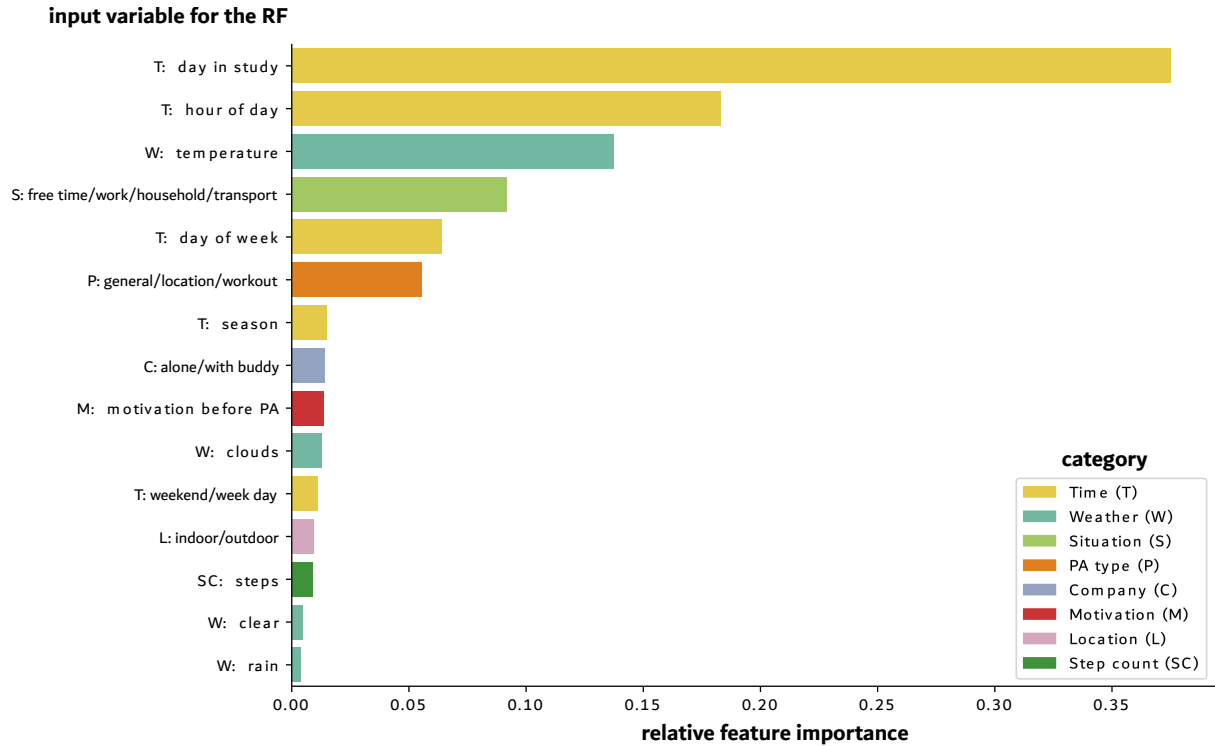


Figure 2: The feature importances show that day in study, hour of day, outdoor temperature, PA situation, day of week, and PA type score the highest on the feature importances in the RF's exploration/repetition prediction.

from predicting exploration to repetition lies around day 15 in the study.

The switch from exploration to repetition after two weeks was also found in our previous study, which showed a significant interaction effect on the star rating feedback after this two-week mark for general PAs [10]. In this previous study [10], we referred to the lower effort and complexity of general PAs as the reason for the quicker shift to repetition, based on [3].

4.1.2. Hour of day

The second dependence plot reveals that exploration predictions are associated with afternoon and evening hours, but not with mornings or nights, with this effect becoming more pronounced later in the study. The highest predictions for explorations are between 4 and 10 pm, especially for later days in the study. An explanation for this pattern could be that our participants (who all belong in the age group of 18 - 44 years) prefer fixed routines in the morning before they go to work or school, and are more likely to explore new activities when they finished their daytime commitments.

This pattern could also be attributed to people's circadian rhythm and the distinction between morning chronotypes who rise and peak early in the day, and evening chronotypes who experience arousal in the afternoon or evening [27]. However, as we did not collect the chronotype of our participants, nor an exact age or occupation, we cannot derive a clear explanation from this.

4.1.3. Outdoor temperature

The third plot suggests a larger prediction for exploration when the outdoor temperature is around 10 degrees Celsius, with two threshold values around 5 and 15 degrees. According to [28], the first and last warm days of the year may motivate people for more PA, which could explain why they are more open to exploration in this temperature range, unlike higher temperatures above 20 degrees.

4.1.4. PA type and situation

High variable importance was also found for PA type and situation, of which the dependence plots are shown at the bottom of Fig. 3. The higher association with repetition in free-time, household, and general contexts can be due to established habits in these contexts. For example, exploration could be less preferred in household tasks because people already have their own routine for these.

Exploration was mostly associated with activities for transportation, during work, at a location, and as a workout. The results of our previous analysis [10] showed that exploration for general PAs was only preferred in the first two weeks of the study, after which higher star ratings were given to repetitions. This can explain why general PAs are mostly associated with repetition prediction in the SHAP dependence plot. Similarly, the star rating was consistently higher for exploration of location and workout PAs in the LMM [10], which corresponds to the conclusions of this SHAP analysis.

As 94% of the active transport and 75% of during work submits were an exploration, as displayed in Table 2, this explains the higher association with exploration. Nonetheless, submits in the situations of active transport and during work are limited, thwarting reliable analyses for these situations.

4.1.5. Day of week

The fifth plot shows that exploration predictions are more associated with Sundays, likely because more people have time off then, allowing more time for exploration. As Saturdays are not associated with more exploration, this can explain the lower variable importance of the “weekend/week day” variable.

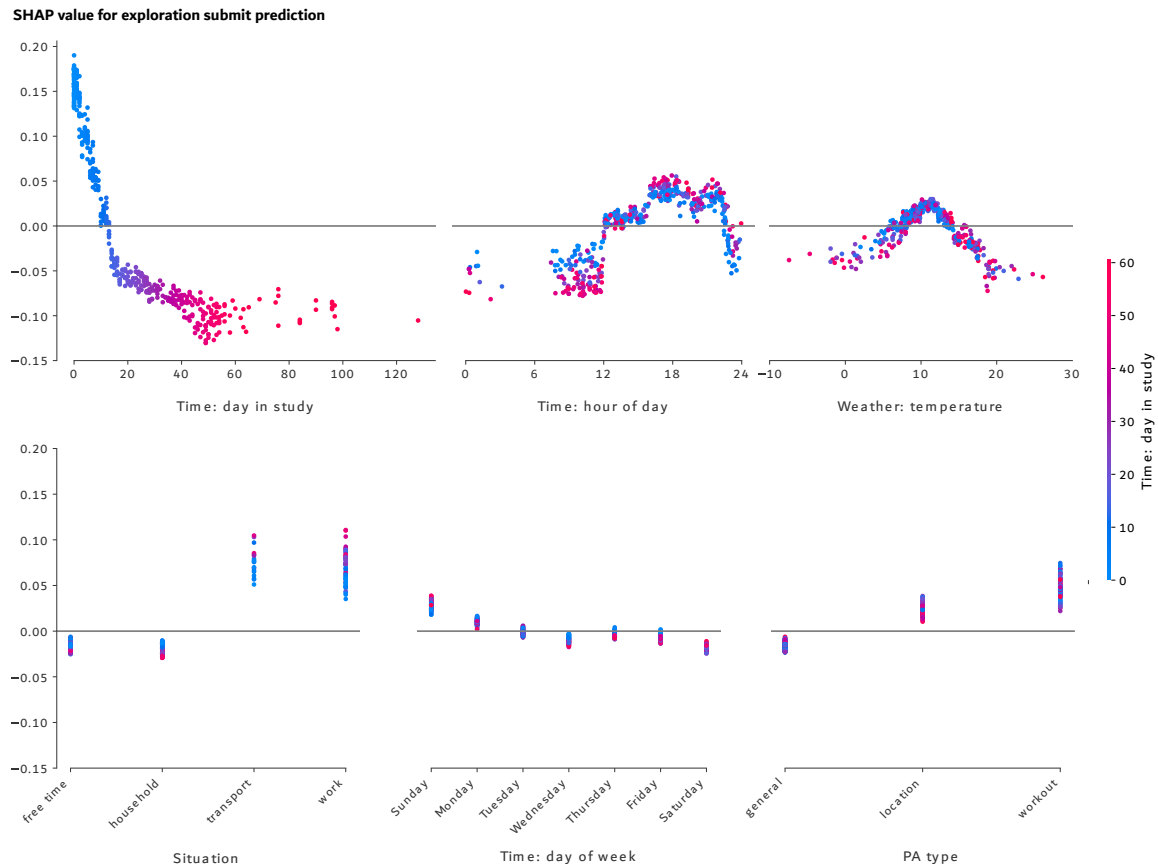


Figure 3: The dependence plots from SHAP show a switch in predicting an exploration item (above the $y=0$ line) to a repetition item (under the $y=0$ line).

4.1.6. Season

The season in which the participant started the study (autumn vs. spring) ranked seventh in the variable importances. We did not create a dependence plot for this variable, but conducted an LMM analysis with this variable to investigate the effect on the amount of weekly MVPA.

To investigate the increase in weekly PA, we compare the total MVPA measured by the EHIS-PAQ ($Q4 \times Q5 + Q7$) in the pre-test and post-test questionnaires [25]. An LMM was fitted with the time (pre-test vs. post-test), the season in which the participant started (autumn vs. spring), and their interaction as fixed effects [26, 29]. To account for possible variations between users, the user ID was considered as a random effect in a random coefficient model with a random slope [26].

We found a significant interaction effect, of which the mean fixed predicted values are shown in Fig. 4, illustrating that the group of participants who started the study in spring had a higher increase in weekly MVPA, on average ($F(1,46.094)=4.079, p=.049$). While a general increase in MVPA in both study groups was expected [30, 31], the stronger increase in the spring group can be explained because PA is higher in warmer months [28].

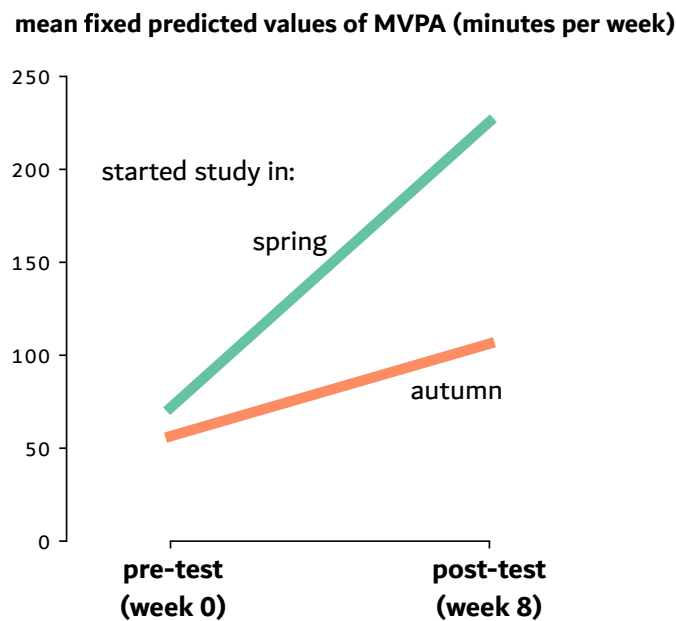


Figure 4: This interaction plot shows the significant interaction effect of the starting season on the increase in weekly MVPA from pre-test to post-test.

4.2. Limitations and future work

As only 34 participants finished the full eight-week study, a large amount of data of the later weeks are missing. Additionally, the study was not conducted year-round or across different climate zones, limiting reliable conclusions about the impact of outdoor temperature. Given the limited submissions for transportation and work PAs, we suggest future research to include longer studies with a larger participant pool.

Our results show that the variables company, location, and motivation had lower importance. However, this company variable could be extended with the presence of a human trainer, as their supervision and guided workout plans, often containing repetition of activities, increase engagement in trainings delivered via mobile apps [32].

Nonetheless, our results indicate that time-related factors primarily drive the decision to explore new activities. We suggest integrating these time-related variables in future RSs for PA promotion. Although we identified specific conditions for exploration, we propose to tailor these conditions to the user. For example, an RS could learn at what moments a user prefers to explore a new activity.

Additional information could be integrated in the system, such as the user's chronotype to take into account moments at which the user is most active [27].

5. Conclusion

Two eight-week user studies with a total of 62 physically inactive participants (<150 minutes MVPA/week) were conducted to investigate factors influencing people's decision to either repeat or explore a PA recommendation. In the MRT, repetition and exploration PAs were provided to the user in an mHealth app at random positions, allowing participants to freely choose between the options at each delivery time.

The RF and SHAP approach identified key factors and conditions influencing the likelihood of exploring a new activity: in the first two weeks of the mHealth intervention, in afternoons and evenings, on Sundays, and for activities at a location or as a workout. Other factors scored lower on the variable importances of the decision between an exploration or repetition item, such as the season, company, motivation, and location. However, we did find a significant higher increase in weekly MVPA for the group that started the study in spring, suggesting that the season might not largely influence the exploration/repetition decision, but does affect the amount of PA.

By defining moments and contexts at which people are more open to explore new activities, this study contributes to future PA recommenders to balance between repeating favorite activities and introducing new ones.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

- [1] World Health Organization, WHO guidelines on physical activity and sedentary behaviour., 2020. URL: <http://www.ncbi.nlm.nih.gov/books/NBK566045/>, oCLC: 1237095892.
- [2] F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, New York, NY, 2022. URL: <https://link.springer.com/10.1007/978-1-0716-2197-4>. doi:10.1007/978-1-0716-2197-4.
- [3] P. Lally, C. H. M. Van Jaarsveld, H. W. W. Potts, J. Wardle, How are habits formed: Modelling habit formation in the real world, *European Journal of Social Psychology* 40 (2010) 998–1009. URL: <https://onlinelibrary.wiley.com/doi/10.1002/ejsp.674>. doi:10.1002/ejsp.674.
- [4] A. Anderson, R. Kumar, A. Tomkins, S. Vassilvitskii, The dynamics of repeat consumption, in: Proceedings of the 23rd international conference on World wide web, ACM, Seoul Korea, 2014, pp. 419–430. URL: <https://dl.acm.org/doi/10.1145/2566486.2568018>. doi:10.1145/2566486.2568018.
- [5] K. Kapoor, K. Subbian, J. Srivastava, P. Schrater, Just in Time Recommendations: Modeling the Dynamics of Boredom in Activity Streams, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, Shanghai China, 2015, pp. 233–242. URL: <https://dl.acm.org/doi/10.1145/2684822.2685306>. doi:10.1145/2684822.2685306.
- [6] Y. Zhang, Variety-Seeking Behavior in Consumption: A Literature Review and Future Research Directions, *Frontiers in Psychology* 13 (2022) 874444. URL: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.874444/full>. doi:10.3389/fpsyg.2022.874444.
- [7] A. Verikas, A. Gelzinis, M. Bacauskiene, Mining data with random forests: A survey and results of new tests, *Pattern Recognition* 44 (2011) 330–349. doi:10.1016/j.patcog.2010.08.011.
- [8] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777. doi:10.5555/3295222.3295230.

- [9] I. Coppens, T. De Pessemier, L. Martens, Repeating my workouts or exploring new activities? a longitudinal micro-randomized user study for physical activity recommender systems, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 176–182. URL: <https://doi.org/10.1145/3631700.3664867>. doi:10.1145/3631700.3664867.
- [10] I. Coppens, T. De Pessemier, L. Martens, Balancing habit repetition and new activity exploration: A longitudinal micro-randomized trial in physical activity recommendations, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1147–1151. URL: <https://doi.org/10.1145/3640457.3691715>. doi:10.1145/3640457.3691715.
- [11] K. Lee, M.-P. Kwan, Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results, *Computers, Environment and Urban Systems* 67 (2018) 124–131. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0198971517301977>. doi:10.1016/j.compenvurbsys.2017.09.012.
- [12] M. T. Mardini, C. Bai, A. A. Wanigatunga, S. Saldana, R. Casanova, T. M. Manini, Age Differences in Estimating Physical Activity by Wrist Accelerometry Using Machine Learning, *Sensors* 21 (2021) 3352. URL: <https://www.mdpi.com/1424-8220/21/10/3352>. doi:10.3390/s21103352.
- [13] Z. Junjie, C. Shenghao, X. Jie, Y. Hua, Random forest-based physical activities recognition by using wearable sensors, *Industria Textila* 73 (2022) 27–33. URL: http://revistaindustriatextila.ro/images/2022/1/004%20ZHANG%20JUNJIE%20Industria%20Textila%201_2022.pdf. doi:10.35530/IT.073.01.20215.
- [14] G. E. Duncan, A. R. Avery, S. Tsang, N. F. Watson, B. D. Williams, E. Turkheimer, The pillars of health: influence of multiple lifestyle behaviors on body mass index and depressive symptoms in adult twins, *BMC Public Health* 22 (2022) 1487. URL: <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-022-13901-7>. doi:10.1186/s12889-022-13901-7.
- [15] S. Wu, Y. Shi, Q. Zhao, K. Men, The relationship between physical activity and the severity of menopausal symptoms: a cross-sectional study, *BMC Women's Health* 23 (2023) 212. URL: <https://bmcmenshealth.biomedcentral.com/articles/10.1186/s12905-023-02347-7>. doi:10.1186/s12905-023-02347-7.
- [16] G. Grekousis, Z. Feng, I. Marakakis, Y. Lu, R. Wang, Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach, *Health & Place* 74 (2022) 102744. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1353829222000053>. doi:10.1016/j.healthplace.2022.102744.
- [17] S. Healy, B. Brewer, E. Hoopes, A. Paller, S. Mayberry, J. Maguire, J. Daly, P. Laxton, F. Patterson, Identifying the most proximal multi-level factors associated with meeting each of the 24-h movement behavior recommendations in a sample of autistic adults, *Disability and Health Journal* 15 (2022) 101367. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1936657422001170>. doi:10.1016/j.dhjo.2022.101367.
- [18] P. Zang, H. Qiu, F. Xian, L. Yang, Y. Qiu, H. Guo, Nonlinear Effects of the Built Environment on Light Physical Activity among Older Adults: The Case of Lanzhou, China, *International Journal of Environmental Research and Public Health* 19 (2022) 8848. URL: <https://www.mdpi.com/1660-4601/19/14/8848>. doi:10.3390/ijerph19148848.
- [19] T. B. Dijkhuis, F. J. Blaauw, M. W. Van Ittersum, H. Velthuis, M. Aiello, Personalized Physical Activity Coaching: A Machine Learning Approach, *Sensors* 18 (2018) 623. URL: <https://www.mdpi.com/1424-8220/18/2/623>. doi:10.3390/s18020623.
- [20] M. Li, S. Jullien, M. Ariannezhad, M. de Rijke, A next basket recommendation reality check, *ACM Trans. Inf. Syst.* 41 (2023). URL: <https://doi.org/10.1145/3587153>. doi:10.1145/3587153.
- [21] M. Li, A. Vardasbi, A. Yates, M. De Rijke, Repetition and Exploration in Sequential Recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Taipei Taiwan, 2023, pp. 2532–2541. URL: <https://dl.acm.org/doi/10.1145/3539618.3591914>. doi:10.1145/3539618.3591914.
- [22] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, N. Meckes, D. R. Bassett, C. Tudor-Locke, J. L. Greer,

- J. Vezina, M. C. Whitt-Glover, A. S. Leon, 2011 Compendium of Physical Activities: A Second Update of Codes and MET Values, *Medicine & Science in Sports & Exercise* 43 (2011) 1575–1581. URL: <https://journals.lww.com/00005768-201108000-00025>. doi:10.1249/MSS.0b013e31821ece12.
- [23] I. Coppens, T. De Pessemer, L. Martens, Exploring the added effect of three recommender system techniques in mobile health interventions for physical activity: a longitudinal randomized controlled trial, *User Modeling and User-Adapted Interaction* (2024). URL: <https://link.springer.com/10.1007/s11257-024-09407-z>. doi:10.1007/s11257-024-09407-z.
- [24] J. Xu, X. Yan, C. Figueroa, J. J. Williams, B. Chakraborty, A flexible micro-randomized trial design and sample size considerations, *Statistical Methods in Medical Research* 32 (2023) 1766–1783. URL: <http://journals.sagepub.com/doi/10.1177/09622802231188513>. doi:10.1177/09622802231188513.
- [25] J. D. Finger, J. Tafforeau, L. Gisle, L. Oja, T. Ziese, J. Thelen, G. B. M. Mensink, C. Lange, Development of the European Health Interview Survey - Physical Activity Questionnaire (EHIS-PAQ) to monitor physical activity in the European Union, *Archives of Public Health* 73 (2015) 59. URL: <http://archpublichealth.biomedcentral.com/articles/10.1186/s13690-015-0110-z>. doi:10.1186/s13690-015-0110-z.
- [26] SPSS Inc., Linear Mixed-Effects Modeling in SPSS: An Introduction to the MIXED Procedure, Technical Report, SPSS Inc., 2005. URL: <https://imaging.mrc-cbu.cam.ac.uk/statwiki/FAQ/multilevel?action=AttachFile&do=get&target=spssegs.pdf>.
- [27] C. P. May, L. Hasher, K. Healey, For Whom (and When) the Time Bell Tolls: Chronotypes and the Synchrony Effect, *Perspectives on Psychological Science* 18 (2023) 1520–1536. URL: <http://journals.sagepub.com/doi/10.1177/17456916231178553>. doi:10.1177/17456916231178553.
- [28] M. Bélanger, K. Gray-Donald, J. O’loughlin, G. Paradis, J. Hanley, Influence of weather conditions and season on physical activity in adolescents, *Annals of Epidemiology* 19 (2009) 180–186. URL: <https://www.sciencedirect.com/science/article/pii/S1047279708003670>. doi:<https://doi.org/10.1016/j.annepidem.2008.12.008>.
- [29] G. Molenberghs, G. Verbeke, A Model for Longitudinal Data, in: *Linear Mixed Models for Longitudinal Data*, Springer New York, New York, NY, 2009, pp. 19–29. URL: http://link.springer.com/10.1007/978-1-4419-0300-6_3. doi:10.1007/978-1-4419-0300-6_3, series Title: Springer Series in Statistics.
- [30] A. Romeo, S. Edney, R. Plotnikoff, R. Curtis, J. Ryan, I. Sanders, A. Crozier, C. Maher, Can Smartphone Apps Increase Physical Activity? Systematic Review and Meta-Analysis, *Journal of Medical Internet Research* 21 (2019) e12053. URL: <http://www.jmir.org/2019/3/e12053/>. doi:10.2196/12053.
- [31] L. Laranjo, D. Ding, B. Heleno, B. Kocaballi, J. C. Quiroz, H. L. Tong, B. Chahwan, A. L. Neves, E. Gabarron, K. P. Dao, D. Rodrigues, G. C. Neves, M. L. Antunes, E. Coiera, D. W. Bates, Do smartphone applications and activity trackers increase physical activity in adults? systematic review, meta-analysis and metaregression, *British Journal of Sports Medicine* 55 (2021) 422–432. URL: <https://bjsm.bmj.com/content/55/8/422>. doi:10.1136/bjsports-2020-102892. arXiv:<https://bjsm.bmj.com/content/55/8/422.full.pdf>.
- [32] L. Boratto, S. Carta, F. Mulas, P. Pilloni, An e-coaching ecosystem: design and effectiveness analysis of the engagement of remote coaching on athletes, *Personal and Ubiquitous Computing* 21 (2017) 689–704. URL: <http://link.springer.com/10.1007/s00779-017-1026-0>. doi:10.1007/s00779-017-1026-0.

Personalized Music Recommendation for People with Autism Spectrum Disorder

Liliana Ardissono¹, Federica Cena¹ and Noemi Mauro¹

¹Computer Science Department, University of Torino, Corso Svizzera 185, Torino, I-10149, Italy

Abstract

The project "ACCESS: Accessibility to Clinical Care for People with ASD through Anxiety Management Using Personalized Applications and IoT" aims to develop innovative technologies to manage anxiety in individuals with Autism Spectrum Disorder (ASD) when they undergo dental and otolaryngology treatments. The premises of this project are the potential of multimedia content to help people with ASD relax when exposed to stressful conditions in healthcare. The result will be an app that personalizes the selection of music tracks and videos to be played before and during the treatment to help the patient distract and cope with stressful conditions that might cause anxiety. For this purpose, the app will receive information about the patient's arousal state collected by physical sensors and/or the clinical staff. The app will use this data to personalize the selection of multimedia content based on the patient's preferences, arousal level, and level of noise around her or him.

Keywords

Personalized healthcare services, recommender systems, autism

1. Introduction

In this paper, we present the project ACCESS, "Accessibility to Clinical Care for People with ASD through Anxiety Management Using Personalized Applications and IoT". The project aims to develop ICT technologies that support the management of anxiety in individuals with Autism Spectrum Disorder (ASD) before and during clinical treatments, with specific attention to dental care treatments and otolaryngological examinations. Given the peculiarity of ASD, which induces different sensitivities to external stimuli, these treatments challenge patients through stimuli concerning hearing and touch, a social context that includes unknown people, and a possibly high level of surrounding noise. Patients thus need help in coping with anxiety.

ICT offers a powerful means to assist individuals with autism in everyday life, including healthcare support [1, 2]. However, for these tools to be effective, efficient, and satisfactory, they should be (i) accessible and user-friendly for individuals with autism, and (ii) seamlessly integrated into clinical protocols, which need to be adapted to incorporate and leverage this technology. Recent studies [3, 4] show the potential to personalize healthcare for people with autism. However, current ICT-based solutions to healthcare support adopt a one-size-fits-all approach to all patients.

The ACCESS project investigates the benefits of integrating digital personalization techniques to help patients with ASD undergo treatments adapting to individual needs, sensitivities, and the noise in the surrounding context. The project builds on findings showing that people's preferred music is effective in anxiety treatment [5] and custom multimedia content administration has positive effects in dental care [3]. However, in these works the selection of the content to be delivered is not automated. Differently, ACCESS aims to develop an application that plays music and multimedia content to calm patients with ASD by adapting content in real-time based on their arousal level and the surrounding noise, and exploiting their preferences for content selection.

Our work in this project concerns the development of the multimedia content selection module. For this purpose, we will develop a novel interaction model to adapt the elicitation of content preferences to ASD people with different functioning levels. Moreover, we will develop a recommender system that combines context awareness (to consider the patient's preferences, arousal state, and the level of noise in the environment) with constraints imposed by the type of visit to be carried out and the patient's sensitivities.

2. Related Work

Music is considered a powerful relaxation tool. Both the acoustic features of the music and individual user preferences play crucial roles in relaxation [6]. In the clinical domain, Lai-Tan et al. report that the best results in applying music therapy to face depression and anxiety are obtained when people listen to their preferred songs because this induces enjoyment [5]. For instance, some people find heavy metal music (that typically has 100-150 Beats Per Minute (BPM)) relaxing while, according to [7], sedative music used in music therapy is characterized by a slow tempo of 60-80 BPM. Thus, knowing the contents that the patient likes, and her/his musical preferences, is key to selecting promising tracks for relaxation.

Personalized technologies for autism are scarce in healthcare contexts [8]. Few works provide personalized support to people with autism (as well as their caregivers and physicians) in managing medical care [9]. Nicolaidis et al. [10] developed a tool that allows patients to create a personalized report for their healthcare provider, improving communication with healthcare operators. Personalized interventions proved to be very effective, especially in the management of anxiety in people with autism [8, 11].

Bondioli et al. [3] developed personalized digital tools such as photos, videos, and interactive PDFs to familiarize ASD children with dental procedures and environments and teach them to perform proper oral hygiene at home. The results confirmed the potential to personalize the selection of ICT tools to reduce anxiety in professional settings. However, in that work, the creation and selection of the multimedia material was carried out by the medical staff in collaboration with the patient and her/his caregivers, without any automated personalization support.

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

✉ liliana.ardissono@unito.it (L. Ardissono); federica.cena@unito.it (F. Cena); noemi.mauro@unito.it (N. Mauro)

🆔 0000-0002-1339-4243 (L. Ardissono); 0000-0003-3481-3360 (F. Cena); 0000-0001-8234-3266 (N. Mauro)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The ACCESS project differs from the previously cited works because we plan to develop an app that dynamically selects and plays multimedia content based on the patient's preferences, sensitivities, and arousal level, and considering the specific treatment to undergo and the presence of noise in the surrounding environment.

Notice that, concerning music selection, our work differs from musical biofeedback, where sounds are exploited as interaction means to convey positive or negative feedback to patients while they carry out tasks in contexts such as relaxation assistance [12] and stroke rehabilitation [13]. In our case, the patient is expected to listen to the music passively, and we aim to select music tracks that can make her/him relax to cope with a clinical treatment.

3. The Project

ACCESS focuses on dental care and otolaryngologist treatments. It is a 2-years project that involves the University of Pisa (development of the main ACCESS tools), the Polytechnic of Milan (development of sensors to monitor patients' anxiety and stress levels), the University of Torino (development of the multimedia content selection system), and the National Research Council (participatory methodology for the project's design, development, and assessment).

As the literature suggests the benefits of music and videos in helping people with ASD cope with medical treatments, the project will target both media. However, we will first focus on delivering music. For this purpose, we are developing a web app to acquire the patient's musical preferences and support her/him during the medical treatments. The app is used in three separate phases:

1. The patient will be instructed to use the app at home to explore and play the preferred multimedia content. The app will adapt the level of guidance in preference elicitation and content exploration to the patient's autonomy, considering different ASD functioning levels. Given the patient's interaction with the app, the system will collect her/his music preferences regarding genres, authors, and individual preferred tracks. Moreover, it will analyze the acoustic features of the preferred tracks through the Spotify API.¹ As people with ASD have individual tolerance levels to acoustic features, this analysis is key to building an individual user profile that specifies the ranges of values compatible with the individual user and can be used to select the tracks to play before and during medical treatments.
2. Before the medical treatment, e.g., in the waiting room, the app will help the patient listen to her/his favorite multimedia content or guide her/him in the exploration of new content. The goal is to keep the user relaxed by suggesting content that reflects the preferences collected in the previous phase and having features that do not agitate her/him. The app will sense the surrounding environment to estimate the noise level. Based on this information and the user profile, it will select the content to be administered (e.g., whether playing music or showing mute videos to avoid adding further disturbing

factors for the patient). For music selection, we plan to steer the invocation of the Spotify API concerning the recommendations² to retrieve music tracks compatible with the patient's musical preferences and sensitivity, and with the surrounding environment. We will do this by feeding the API with data about authors, genres, and the minimum and maximum values of the acoustic features stored in the patient's user profile.

3. During medical treatment, the app will receive information about the patient's arousal state (e.g., heart and respiration rates collected by sensors, or information provided by the medical staff). Moreover, it will sense the surrounding environment to estimate the noise level and play the appropriate multimedia content based on this data and the user profile. In this case, the type of content selected by the app could change to reflect the patient's arousal state.

The app will be tested in collaboration with the Audiology and Phoniatrics Clinics (OPC) at AOUP Pisa to measure the user experience during the interaction with it, and its effectiveness in supporting patients during medical treatments. So far, we participated in a focus group with 2 caregivers, 1 psychologist, 1 speech therapist, 1 otolaryngologist, 6 researchers and 2 technologists to gather stakeholders' requirements and needs. The focus group confirmed that the use of technology can help the user to prepare before the medical treatment and could also be useful while the treatment is performed. In addition, a personalized approach could improve the user experience if a strong weight is given to user preferences.

Acknowledgments

This study is funded by the NATIONAL RECOVERY AND RESILIENCE PLAN (NRRP) – MISSION 4 COMPONENT 2 INVESTMENT 1.1 – “Fund for the National Research Program and for Projects of National Interest (NRP)” Italian Ministry of Education call PRIN 2022 PNRR D.D. n. 1409 14/09/2022 - Title of the project “ACCESS: Accessibility to clinical care for people with ASD through anxiety management by using personalized applications and IoT”, project number P2022PBTSC. We thank Luca Bonamico for his contribution to the project.

References

- [1] A. J. Doherty, H. Atherton, P. Boland, R. Hastings, L. Hives, K. Hood, L. James-Jenkinson, R. Leavey, E. Randell, J. Reed, L. Taggart, N. Wilson, U. Chauhan, Barriers and facilitators to primary health care for people with intellectual disabilities and/or autism: an integrative review, *BJGP Open* 4 (2020). doi:10.3399/bjgpopen20X101030.
- [2] E. McGhee Hassrick, L. Graham Holmes, C. Sosnowy, J. Walton, K. Carley, Benefits and risks: A systematic review of information and communication technology use by autistic people, *Autism in Adulthood* 3 (2021). doi:10.1089/aut.2020.0048.
- [3] M. Bondioli, S. Pelagatti, M. C. Buzzi, M. Buzzi, C. Senette, Ict to aid dental care of children with

¹<https://developer.spotify.com/documentation/web-api/reference/get-audio-features>. This API allows data to be recovered on the acousticness, danceability, energy, instrumentality, liveness, loudness, speechiness, and tempo of each track.

²<https://developer.spotify.com/documentation/web-api/reference/get-recommendations>

- autism, in: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 321–322. URL: <https://doi.org/10.1145/3132525.3134799>. doi:10.1145/3132525.3134799.
- [4] S. Petty, M.-L. Bergenheim, G. Mahoney, L. Chamberlain, Adapting services for autism: Recommendations from a specialist multidisciplinary perspective using freelistig, *Current Psychology* 42 (2021) 1–12. doi:10.1007/s12144-021-02061-3.
 - [5] N. Lai-Tan, M. G. Philiastides, F. Kawsar, F. Deligianni, Toward personalized music-therapy: A neurocomputational modeling perspective, *IEEE Pervasive Computing* 22 (2023) 27–37. doi:10.1109/MPRV.2023.3285087.
 - [6] D. Mélo, Investigating the role of personalization when creating relaxing playlists, *WebMedia '20*, Association for Computing Machinery, New York, NY, USA, 2020, p. 213–216. URL: <https://doi.org/10.1145/3428658.3431755>. doi:10.1145/3428658.3431755.
 - [7] J. Yu, B. Benatallah, F. Casati, F. Daniel, Understanding mashup development, *IEEE Internet Computing* 12 (2008) 44–52. doi:10.1109/MIC.2008.114.
 - [8] W. Simm, M. Ferrario, A. Gradinar, M. Smith, S. Forshaw, I. Smith, J. Whittle, Anxiety and autism: towards personalized digital health, in: C. Lampe (Ed.), *Proceedings - The 34th Annual CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery (ACM), United States of America, 2015, pp. 1270–1281. doi:10.1145/2858036.2858259.
 - [9] P. M., Y. C., Food and therapy recommendation system for autistic syndrome using machine learning techniques, in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–6. doi:10.1109/ICECCT.2019.8868979.
 - [10] C. Nicolaidis, D. Raymaker, K. McDonald, S. Kapp, M. Weiner, E. Ashkenazy, M. Gerrity, C. Kripke, L. Platt, A. Baggs, The development and evaluation of an online healthcare toolkit for autistic adults and their primary care providers, *Journal of general internal medicine* 31 (2016) 1180–1189. doi:10.1007/s11606-016-3763-6.
 - [11] J. Rodgers, S. Brice, P. Welsh, B. Ingham, C. Wilson, G. Evans, K. Steele, E. Cropper, A. Frpcch, M. Free-ston, J. Parr, A pilot randomised control trial exploring the feasibility and acceptability of delivering a personalised modular psychological intervention for anxiety experienced by autistic adults: Personalised anxiety treatment-autism (pat-a), *Journal of Autism and Developmental Disorders* (2023) 1–16. doi:10.1007/s10803-023-06112-5.
 - [12] B. Yu, M. Funk, J. Hu, L. Feijs, Unwind: a musical biofeedback for relaxation assistance, *Behaviour & Information Technology* 37 (2018) 800–814. URL: <https://doi.org/10.1080/01449290500331131>. doi:10.1080/0144929X.2018.1484515.
 - [13] P. Kantan, E. G. Spaich, S. Dahl, A technical framework for musical biofeedback in stroke rehabilitation, *IEEE Transactions on Human-Machine Systems* 52 (2022) 220–231. doi:10.1109/THMS.2021.3137013.

Recommending News Articles for Public Health Intelligence

Diana F. Sousa^{1,*}, Nicolas Stefanovitch¹ and Luigi Spagnolo¹

¹*European Commission Joint Research Centre, Ispra, Italy*

Abstract

Public Health Intelligence (PHI) is the process of extracting useful information from vast amounts of data to help quickly identify and respond to health threats. Systems that perform PHI are used daily by different national and international organizations. One of the most prominent platforms is the Epidemic Intelligence from Open Sources Initiative (EIOS) platform, which continuously gathers health-related news items. However, the EIOS platform requires users to swift through unrelated information to their domain or work needs, even when using different filtering options. This inefficiency in assessing the relevance of each article creates the need to develop a recommender system that effectively positions each incoming article according to its significance. In this work, we present the first iteration of this system, making use of previous user interactions with the articles already available in the platform and the articles' content and metadata. We investigated various configurations to address the problem of data sparsity by conducting cluster-based harmonization. Our best-performing model reports an NDGC@K of 0.4108 and an F-measure@K of 0.7287, respectively, for $K = 100$ articles.

Keywords

Public Health Intelligence, Recommender Systems, Clustering, User Data, Health News Articles

1. Introduction

Every day, expert analysts swift through tens of thousands of health news articles to identify incoming health threats, such as an outbreak of a disease and other types of relevant health information regarding humans, animals, and plants. To do their work, the analysts use platforms that primarily aim to gather all news articles and reports on health topics. The Epidemic Intelligence from Open Sources (EIOS) platform is the most well-known Public Health Intelligence (PHI) resource. EIOS is an international initiative led by the World Health Organization (WHO) with a unified all-hazards One Health approach to early detection, verification, assessment and communication of public health threats using publicly available information¹.

The analysts working on identifying relevant health information for each of their purposes and domains have to carry out their day-to-day work and often prepare for large mass gatherings, e.g. sports championships or the Olympics games, which present an increased risk of disease outbreaks. Thus, analysts face the daily challenge of processing a high volume of information. EIOS collects 50,000 articles a day; as such, the possibility to organise information by relevance using a recommender system, a feature currently missing in EIOS, would improve analysts' experience by significantly alleviating the time spent identifying which articles are relevant for their purpose.

Health recommender systems are broad and encompass epidemic forecasting tools such as HealthMap [1] and EPIWATCH², which track disease spread by collecting information from various channels, including news and social media [2]. In crises, these recommender systems are pivotal for effectively allocating medical resources and guiding interventions. Moreover, they extend to environmental health monitoring, offering air and water quality advice, and are integrated into Personal Health Records

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ diana.francisco-de-sousa@ec.europa.eu (D. F. Sousa); nicolas.stefanovitch@ec.europa.eu (N. Stefanovitch); luigi.spagnolo@ec.europa.eu (L. Spagnolo)

ORCID 0000-0003-0597-9273 (D. F. Sousa); 0009-0000-2061-3216 (N. Stefanovitch); 0009-0008-0179-7468 (L. Spagnolo)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.who.int/initiatives/eios>

²<https://www.epiwatch.org/>

(PHRs) to suggest health actions [3, 4], such as vaccine recommendation features [5]. Lastly, health applications employ these systems to promote personalized health-related behaviour [6, 7]. Despite their potential, ensuring data privacy, system validation, multilingual adaptability, and ethical use is paramount for maintaining public trust and successfully deploying recommender systems in public health.

To address the need for more efficient identification of relevant articles coming to the EIOS platform, we created a content-based recommender system that is based on three data streams: (1) The content of the article, specifically the first 1000 characters, taking into account complete sentences; (2) The event type labels resulting from the application of a pandemics event classifier; (3) The user interactions with each article (i.e., relevance score), obtained using a scoring function that considers the type and number of interactions, augmented with a clustering procedure to tackle data sparsity. We tested XGBoost [8] with seven different data augmentation procedures.

The article’s main contributions are:

- Usage of an event classifier labels to enrich the recommendation algorithm;
- Introducing a clustering-based approach for user activity harmonization to address data sparsity challenges;
- Development of a content-based system for recommending articles in real-world PHI scenarios.
- Error analysis conducted on example use cases to assess whether the recommender can flag relevant information missed by the users.

The data described and used in this paper was sourced from a live system. As a result, Intellectual Property and Privacy regulations apply, preventing dataset sharing. Nevertheless, the experiments detailed in this article are significant for health recommender systems. They offer valuable insights into implementing AI-based solutions using actual user data.

Section 2 describes the data, mainly the metadata used to train the recommender system. Section 3 describes the cluster-based procedure to perform data harmonization and tackle sparsity. Section 4 presents the recommender system, including model and evaluation metrics. Section 5 presents results, a discussion of the clustering plus recommendation pipeline, and an error analysis of the different clustering modalities. Finally, Section 6 presents the main conclusions and future work.

2. Data

To train and test our model, we used a dataset of approximately 3.5 million articles from the EIOS platform from 01/01/2018 to 09/06/2022 (about four years and six months). This dataset contains all articles and information about user interactions with those articles in all the different languages captured by the platform. For this work, which constitutes the first iteration to create a recommendation solution for PHI systems, the features we focused on are the text of the article, the event labels generated through an event classifier, and the user activity for each article (i.e., relevance score). Figure 1 illustrates the high-level pipeline involving three input data streams in the recommender system.

2.1. Text

The dataset has the full text for each article. However, due to memory limitations and to keep the focus on the core information of the article, we decided to consider only the first few sentence(s), up to 1000 characters.

To preprocess this truncated-article text, we only removed stop words from English articles. In order to vectorise the articles, we used the `TfidfVectorizer` function from the `scikit-learn`³ using the maximum document frequency set to ignore terms that have a document frequency strictly higher than 1.

³<https://scikit-learn.org/>

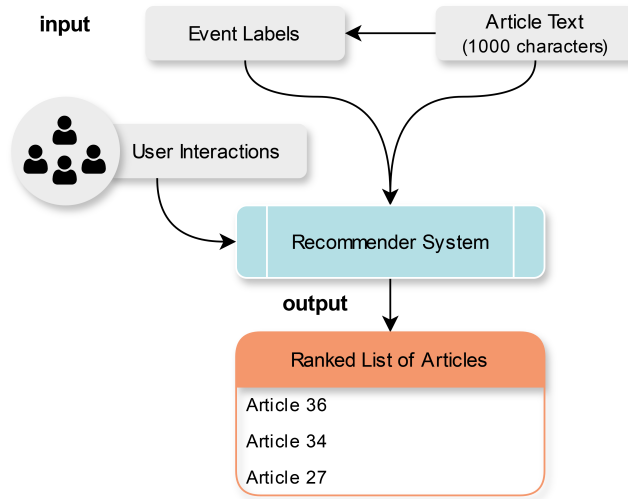


Figure 1: Recommender system high-level pipeline with the three data streams and expected output.

2.2. Event Labels

We assigned event labels to the articles to boost the system’s performance and better characterize and differentiate between articles. We ran an event classifier for each article within the dataset to classify them into one or more of 27 events following a taxonomy and pipeline created and developed by Piskorski et al. [9]. Some of the most frequent labels are (1) Reporting Cases (i.e., reporting on cases of infections, hospitalizations, deaths, recoveries of single persons and groups, provision of updates thereon, which covers a short time span and specific location), (2) Reporting Situation (i.e., provision of updates on the overall situation of the outbreak, current total figures, observed trends, forecast, which spans longer period of time, and also covers cross-regional and cross-country comparisons), (3) Measuring Vaccine/Medicine Roll-out (i.e., covers events revolving around the roll-out of vaccines, medicines, equipment to combat the disease or mitigate the consequences, and includes also events related to sharing experience, measure hesitancy, anti-vax movements, etc.). Other coarse-grain labels are Impact, Violation, Research & Development, Communication, Support, and Miscellaneous.

To preprocess these event labels, we applied the MultiLabelBinarizer function, given that each article can have more than one label wrapped to work with ColumnTransformer, both from the scikit-learn library.

2.3. User Activity

The user activity for each article is pre-determined by the weighted sum of user interactions, which we express as a relevance score. Different types of interactions yield different weights. The platform computes the user activity using the weights presented in Table 1.

Table 1

Types of user activities and their corresponding weights.

User Activity	Weight
Read Preview	0 or 1
Read Detail	0 or 2
Flag for Follow Up	3
Export to Report	5
Attach to Team Communication	5
Comment	5
Pin to Board	Variable

When it comes to the "Read Preview" interaction, the weight assigned will be zero if there are no other user interactions on the article, and one otherwise (excluding "Read Detail"). For the "Read Detail" interaction, the weight assigned will be zero if there are no other user interactions on the article, and two otherwise (excluding "Read Preview"). As for the "Pin to Board" activity, the weight assigned is five or ten, based on whether the board is private or public, respectively. The weights assigned to each activity are proportional to the complexity of the activity being performed.

One of the issues we had to address before the application of our system was the low proportion of articles with user interactions (2.03%). The news feeds presented to users are ordered by time and user preference settings (i.e., pre-determined keywords, languages, etc.). When a new story emerges, EIOS users often interact with the first article reporting on the story, with the article they deemed to be from the most reliable source, or even with the article that reports the story in their language, among other preferences.

This interaction pattern means that if we have a single story reported in multiple articles from multiple sources, the user activity will vary widely among almost identical articles, with only a few articles getting interacted with. Thus, raw user activity does not directly equate to user interests. In the following section, we will outline how we intend to tackle this issue using clustering to make the relevance score a reliable measure of user interest.

3. Cluster-based Harmonization

We considered that articles with no interaction are articles for which the relevance is unknown rather than zero, transforming the problem into a semi-supervised learning one. We corrected the relevance score of articles in clusters to deal with this and fall back on a supervised learning problem.

The harmonization of user activity/relevance scores happens at the level of clusters of related articles, some of which have an interaction score and others potentially none. We intended that the clusters captured reports on the same event; as such, they were computed considering both the time and semantic aspects. The clustered article data corresponds to the text described in the Data section. The entire dataset was split into five-day chunks, capturing a story's average duration, as represented in Figure 2. Inside a chunk, all the pairs of articles were compared using sentence embeddings, and the pairs whose similarity was above a given threshold were put into a graph. The semantic similarity model used was `distiluse-base-multilingual-cased-v2`, with a threshold of 0.90. Finally, the graphs of all clusters were merged, and the set of connected components yielded the global set of clusters. This approach is designed to be adaptable, allowing it to pick up news stories that last longer than five days and preventing the merging of similar stories from widely different time spans.

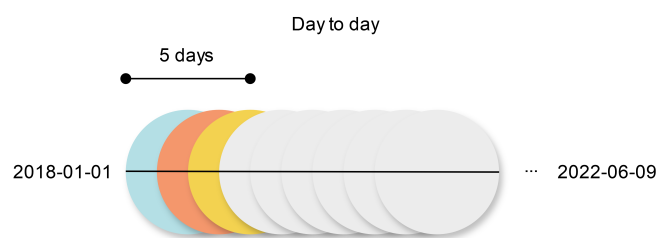


Figure 2: Representation of five day time-span local clusters in the timeframe considered.

Once the clusters were computed, the second step of our procedure was to harmonize the score of all the articles belonging to each cluster. To illustrate this, we will consider this example cluster of four identical articles and their corresponding user activities scores:

- **Cluster:** [Article 1, Article 2, Article 3, Article 4]
- **User Activities:** [0, 5, 17, 0]

Table 2

Counts of the number of articles per modality, their percentage with user activity, and the threshold used for predictions.

Modality	Number of Articles	User Activity	Threshold
Original	3 589 739	2.03	9
Sum	3 589 739	2.25	10
High	3 589 739	2.25	9
AVG	3 589 739	2.22	9
Low	3 589 739	2.20	9
Random	3 589 739	2.22	9
Discard	3 288 085	2.22	9
Null	3 287 754	2.20	9

Clusters containing articles with only zero relevance are left untouched, except for the Null configuration, detailed below. Clusters with mixed or only positive relevance were further processed to reassign the relevance score of every article within that cluster. We considered seven different modalities to perform the harmonization, which are illustrated in the following example:

- **Original:** Nothing changes $\rightarrow [0, 5, 17, 0]$.
- **Sum:** Application of the sum of all user activities in the cluster to all the articles in the cluster $\rightarrow [22, 22, 22, 22]$.
- **High:** Application of the highest user activity in the cluster to all the articles in the cluster $\rightarrow [17, 17, 17, 17]$.
- **Average:** Application of the average of all user activities computed by dividing the sum of all user activities by the number of articles in the cluster $\rightarrow [5.5, 5.5, 5.5, 5.5]$.
- **Low:** Application of the lowest user activity in the cluster to all the articles in the cluster $\rightarrow [5, 5, 5, 5]$.
- **Random:** To each cluster, application of a random configuration from the ones described above $\rightarrow [22, 22, 22, 22]$ or $[17, 17, 17, 17]$ or $[5.5, 5.5, 5.5, 5.5]$ or $[5, 5, 5, 5]$.
- **Discard:** Keep only articles in the cluster that have user activity $\rightarrow [5, 17]$.
- **Null:** Remove clusters where there is no article with user activity $\rightarrow [0, 5, 17, 0]$.

The Discard and Null modalities constitute filtering options, not modifying the relevance score but excluding articles with no score, using different approaches. For Discard, all non-relevant articles are removed from the cluster for the clusters with at least one relevant article. For Null, all clusters where all the articles have a zero relevance score are removed.

Table 2 showcases the augmentation in general percentage for each modality compared to Original, reflecting our extremely conservative clustering procedure. The Threshold column is the user activity value considered at the recommendation level to decide if an article should be recommended. We obtained this value by considering the average of the positive (> 0) user activities for each modality. Figure 3 reports the histogram of the user activity/relevance score of articles comparing the distribution of all the original data and the clustered articles' distribution of the sum modality, presenting similar profiles.

4. Recommender System

The data available does not specify which users interacted with the articles; it only shows the overall user activity for each article. Therefore, recommendations are not based on individual user behaviour but on global preferences towards specific topics and domains, making adopting a collaborative filtering approach unfeasible.

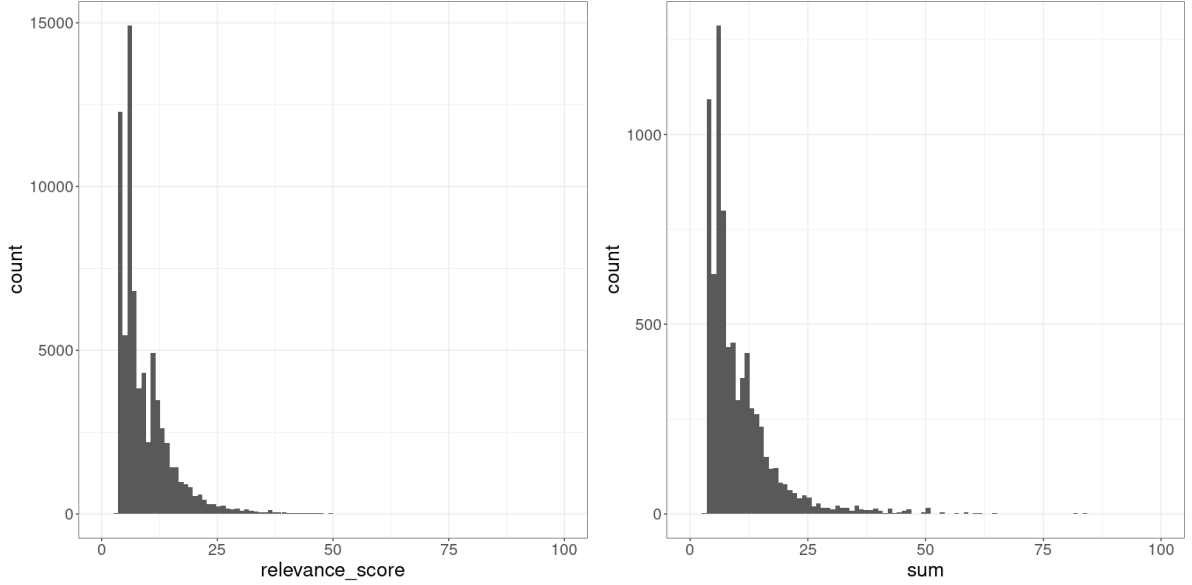


Figure 3: Histogram of relevance score distribution excluding non-relevant articles: original for all articles (left) and with sum harmonization (right).

4.1. Model

In this approach, each row of our data represented an article with a relevance score corresponding to the weighted sum of user interactions with the article. As stated in the previous sections, the features considered for training were the article attributes: a text section at the beginning of the article, the events labels that report on the article classification, and the relevance scores. Our goal was to recommend articles with higher engagement that are, therefore, more relevant.

We divided our data into training (80%) and testing (20%) with a 5-fold cross-validation. For the training data, we used an XGBoost regression model [8]. This model learns to predict each article’s user engagement by building a series of decision trees sequentially, using gradient descent to minimize the loss. We did not do hyperparameter tuning, leaving the default parameters stated in the package documentation⁴, to avoid overfitting the model to our data and maintain its generalizability to new data.

4.2. Evaluation Metrics

The evaluation metrics considered for the different settings were the following:

- **RMSE:** Root mean square error (RMSE) or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.
- **NDGC@K:** Normalized Discounted Cumulative Gain (NDCG) considers both the relevance and the position of items in the ranked list in the top K items.
- **Precision@K:** Precision at K measures the proportion of relevant items among the top K items.
- **Recall@K:** Recall at K measures the coverage of relevant items in the top K items.
- **F-measure@K:** Harmonizes precision and recall to provide a balanced metric in the top K items.

We considered 5, 10, 15, and 100 items for K. For Precision, Recall, and F-measure, since the values considered are binary, we present only the $K = 100$ configuration to reflect better the real user needs in our setting.

⁴<https://xgboost.readthedocs.io/en/stable/parameter.html>

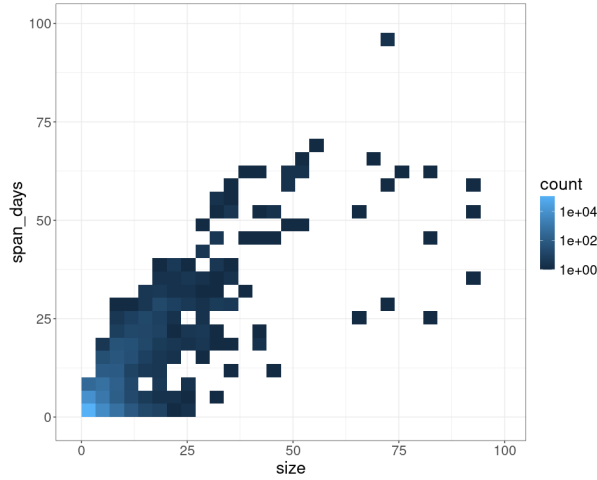


Figure 4: Heatmap of cluster size versus cluster span

Table 3

Statistics over clusters characteristics with different distributions.

	Only 0	Only <i>pos</i>	More 0	More <i>pos</i>	Eq. prop.
Count	124332	829	1188	182	3798
Max size	177	6	180	21	10
AVG size	2.3	2.1	5.3	3.3	2.0
Max span	133	7	244	28	12
AVG span	1.9	1.8	4.5	3.0	1.8
Max peak	25	6	11	4	4
% total rel.	0.00	0.24	0.18	0.06	0.52

5. Results and Discussion

This section presents the main results regarding all modalities and discusses the model’s successes and potential limitations given the simplified approach.

5.1. Local Clusters Distribution

The settings used for clustering were conservative as it was performed on relatively long text with a high threshold. In total 8.7% of the articles were clustered. The data revealed a predominant pattern of small clusters, with 81% having a size of 2 and 99% under size 7. These clusters also tend to be short-lived, with 49% lasting a single day and 99% up to 8 days. The manual review confirms that articles in these clusters are remarkably similar, often being near-perfect duplicates. Notably, the clusters with the longest lifespan appear to be populated by automatically generated reporting articles.

In Figure 4, we plotted the distribution of cluster size and the distribution of the span of the cluster in days; some outliers fall outside the limits of the figure and are not shown. A cluster’s median size was two articles, and the median span was two days. Table 3 reports several statistics over the clusters, grouping them based on whether the relevance of related articles contains only 0, only positive (*pos*), mostly 0, mostly *pos*, both 0 and *pos* in equal proportion. We report the mean and max cluster size and span, and the maximal peak article count, and the proportion of the total relevance. We can observe that clusters attracting most of the relevance tend to be relatively small and short.

5.2. Modality Performance

Table 4 presents the results of comparison of different clustering modalities for user data augmentation using the RMSE, NDGC@K, Precision@K, Recall@K and F-measure@K metrics, taking into account

Table 4

Comparison of different clustering modalities for user data augmentation using the RMSE, NDGC@K, Precision@K, Recall@K and F-measure@K metrics.

Modality	RMSE	NDGC@K				Precision@K 100	Recall@K 100	F-measure@K 100
		5	10	15	100			
Original	1.5739	0.1903	0.1807	0.1606	0.1749	0.3466	1.0000	0.5136
Sum	1.7362	0.4946	0.4382	0.4137	0.4108	0.5740	1.0000	0.7287
High	1.6591	0.1722	0.2283	0.2318	0.2537	0.4320	1.0000	0.6015
AVG	1.5313	0.1612	0.1575	0.1559	0.1734	0.3478	0.9946	0.5137
Low	1.6188	0.1767	0.1950	0.1903	0.2201	0.4060	1.0000	0.5762
Random	1.6380	0.1622	0.2015	0.2023	0.2516	0.4440	1.0000	0.6139
Discard	1.6381	0.1516	0.1518	0.1377	0.1802	0.3880	1.0000	0.5575
Null	1.6318	0.1861	0.1884	0.1855	0.1834	0.3720	1.0000	0.5420

Table 5

Comparison of different clustering modalities with user data augmentation performance on the original test set (non-augmented) using the RMSE, NDGC@K, Precision@K, Recall@K, and F-measure@K metrics.

Modality	RMSE	NDGC@K				Precision@K 100	Recall@K 100	F-measure@K 100
		5	10	15	100			
Sum	1.5794	0.2259	0.1878	0.1822	0.2014	0.3640	1.0000	0.5330
High	1.5718	0.1678	0.2052	0.1828	0.1864	0.3280	1.0000	0.4929
AVG	1.5732	0.1599	0.1542	0.1531	0.1660	0.3417	0.9944	0.5073
Low	1.5715	0.1535	0.1678	0.1627	0.1882	0.3520	1.0000	0.5195
Random	1.5719	0.1369	0.1627	0.1634	0.1966	0.3880	1.0000	0.5578
Discard	1.5752	0.1516	0.1518	0.1377	0.1548	0.3240	1.0000	0.4886
Null	1.5741	0.1861	0.1884	0.1788	0.1728	0.3460	1.0000	0.5139

5-fold cross validation.

Most modalities surpass the Original configuration. However, when considering NDGC@100, only Sum, High, Low, and Random perform distinctly better than the Original, with Sum being significantly better. The performance of Sum places the possibility that the actual user activity value represents the sum of all identical article interactions, performing twice as well as the Original.

Table 5 showcases the same procedure but using the Original modality test set. In this setting, the superior performance of the Sum modality is not as noticeable, but all modalities, except AVG, Discard, and Null, perform better than Original. A possible justification for this behaviour could be that our system performs better with more data regardless of how it is labelled, hindering the performance of Null and Discard modalities. Additionally, the AVG configuration could make stronger and weaker signals less noticeable, diluting their relative importance in a ranking setting.

5.3. Error Analysis

Table A1 (Appendix) showcases the false positives found across the five rounds of cross-validation for the different modalities at the top five ($K = 5$). All modalities introduce errors compared to the Original, with Sum and High introducing fewer wrong articles as also reflected in Table 4.

We analysed the articles for a fail rate of over or equal to 7/8 modalities to interpret what could have made most modalities assign relevance. We then analysed whether it was indeed a failure by our models or if it could have been a missed relevant article by the users and/or the clustering procedure for data augmentation. This selection resulted in six articles represented in Table 6 and marked with an asterisk (*) in Table A1 (Appendix). Table 7 reports on the details of these articles.

Even though Table 7 does not report on the sources for the articles, all of these are pieces that

Table 6

Most frequent articles and their scores across all modalities represented in the Top five ($K = 5$) false positives (fail rate $\geq 7/8$).

Article	Original	Sum	High	AVG	Low	Random	Discard	Null
1757749	51	46	34	41	46		51	41
210702	52	46	46	52	52	46	52	52
1177084	36		36	36	36	36	36	39
1642976	42		42	42	42	42	42	42
2083725	33		45	39	46	46	46	46
458168	68	64	64	68	68	64	68	68

Table 7

Most frequent articles across all modalities represented in the Top five ($K = 5$) false positives, their event labels, and the general topic they discuss (fail rate $\geq 7/8$).

Article	Event Labels	General Topic
1757749	REPORTING-CASES	First-case reporting on coronavirus in Africa
210702	MISCELLANEOUS-UNRELATED	Paediatric acute hepatitis reporting in the UK
1177084	IMPACT-OTHER	Coronavirus impact on the industry in India
1642976	MISCELLANEOUS-OTHER	Political landscape in Haiti
2083725	REPORTING-CASES	Coronavirus cases reporting in France
458168	REPORTING-SITUATION, REPORTING-CASES	Vaccination fears and chickenpox cases in Angola

primarily reflect the general opinion of an isolated expert of the respective fields and not official sources from health organisations, such as the WHO. So, even if the articles' domain and general topic might be relevant, analysts can avoid the article for not being factually about what is happening but more of a reflection on what has been happening throughout a specific outbreak, such as in article 458168. In this article, an expert demonstrates how vaccination fears are at fault for rising chickenpox cases in Angola. If other sources are already monitoring the number of cases, this piece can be overlooked because it is primarily about cause rather than consequence. Nevertheless, we believe this article and similar articles can indicate the worsening of ongoing outbreaks. As such, these shouldn't be ignored but used as indicators to flag future similar events pre-emptively.

6. Conclusion and Future Work

This article presented the first step in developing a recommendation system for a pre-existing platform, EIOS, developed for PHI. Therefore, the results and analysis still need to be completed. However, this work successfully showcases a pipeline for developing a content-based system recommending articles in real-world PHI scenarios. It introduces a clustering-based approach to tackle data sparsity and the use of event classifier labels to enrich the recommender algorithm. While more complex metadata and advanced models and approaches are available and will be used in the future, this first attempt successfully demonstrated a way of dealing with data sparsity for our case study, which in turn improved the model performance from an NDGC@K of 0.1749 to 0.4108, at $K = 100$, for the Sum cluster-based harmonization modality.

Looking ahead, we plan to further develop this approach by considering multiple users, article sources, other types of article metadata, and exploring the conjugation of clustering modalities and filters. Additionally, we aim to involve analysts in our approach to evaluate performance on actual end-users, thereby enhancing the robustness and applicability of our system.

References

- [1] C. C. Freifeld, K. D. Mandl, B. Y. Reis, J. S. Brownstein, Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports, *Journal of the American Medical Informatics Association* 15 (2008) 150–157. doi:10.1197/jamia.M2544.
- [2] J. S. Brownstein, C. C. Freifeld, L. C. Madoff, Digital disease detection—harnessing the web for public health surveillance, *The New England journal of medicine* 360 (2009) 2153. doi:10.1056/NEJMp0900702.
- [3] J. M. Balbus, R. Barouki, L. S. Birnbaum, R. A. Etzel, P. D. Gluckman, P. Grandjean, C. Hancock, M. A. Hanson, J. J. Heindel, K. Hoffman, et al., Early-life prevention of non-communicable diseases, *The Lancet* 381 (2013) 3–4. doi:10.1016/S0140-6736(12)61609-2.
- [4] H. Schäfer, S. Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, C. Trattner, Towards health (aware) recommender systems, in: *Proceedings of the 2017 International Conference on Digital Health, DH '17*, Association for Computing Machinery, New York, NY, USA, 2017, p. 157–161. doi:10.1145/3079452.3079499.
- [5] M. R. Pereira, Updated 2024 us vaccine recommendations from the advisory committee on immunization practices, *American Journal of Transplantation* 24 (2024) 514–516. doi:10.1016/j.ajt.2024.02.012.
- [6] W. T. Riley, D. E. Rivera, A. A. Atienza, W. Nilsen, S. M. Allison, R. Mermelstein, Health behavior models in the age of mobile interventions: are our theories up to the task?, *Translational behavioral medicine* 1 (2011) 53–71. doi:10.1007/s13142-011-0021-7.
- [7] H. Torkamaan, J. Ziegler, Recommendations as challenges: Estimating required effort and user ability for health behavior change recommendations, in: *Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 106–119. doi:10.1145/3490099.3511118.
- [8] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. doi:10.1145/2939672.2939785.
- [9] J. Piskorski, N. Stefanovitch, J. P. Linge, S. Kharazi, J. Mantero, G. Jacquet, A. Spadaro, G. Teodori, Multi-label infectious disease news event corpus, in: *Proceedings of the Text2Story'23 Workshop*, Elsevier, Dublin, Republic of Ireland, 2023, pp. 171–183.

A. Error Analysis with False Positives

Table A1

Top five ($K = 5$) false positives and their predicted scores across the five rounds of cross-validation for the different modalities.

Article	Original	Sum	High	AVG	Low	Random	Discard	Null
1757749*	51	46	34	41	46		51	41
2113787	42		37	50	38	37	45	
2355188	42		37	50	38	37	45	
1524234	34							
164429	55	49	55	42	42	41		
210702*	52	46	46	52	52	46	52	52
2713654	50			51		45		
1177084*	36		36	36	36	36	36	39
1642976*	42		42	42	42	42	42	42
1463715	36	48		32	48	36		
2083725*	33		45	39	46	46	46	46
462637	38			37	32			38
1388337	30			33				
458168*	68	64	64	68	68	64	68	68
2924180	40			40	40	40	40	40
1450025		44	32					
2671862		50	50		40	58		
455925		36						
857479		49						
1418055			32			32		
1152620			32			31		
185204			43	37	36	43		
3105707					30			
1255976						43		
471697							27	
390951							30	
2724292							30	
2274914							38	38
727629							42	
416240								66
2398504								33
979659								37
2163701								38
1133939								40

Improving the Prediction of Individual Engagement in Recommendations using Cognitive Models^{*}

Roderick Seow^{1,*}, Yunfan Zhao², Duncan Wood¹, Milind Tambe² and Cleotilde Gonzalez¹

¹Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, Pennsylvania, United States of America

²Harvard University, Massachusetts Hall, Cambridge, Massachusetts, United States of America

Abstract

For public health programs with limited resources, the ability to predict how behaviors change over time and in response to interventions is crucial for deciding when and to whom interventions should be allocated. Using data from a real-world maternal health program, we demonstrate how a cognitive model based on Instance-Based Learning (IBL) Theory can augment existing purely computational approaches. Our findings show that, compared to general time-series forecasters (e.g., LSTMs), IBL models, which reflect human decision-making processes, better predict how individuals' behaviors change over time (transition-consistency) and in response to receiving an intervention (intervention-sensitivity). We further show that IBL parameters capture the individual differences in transition-consistency and intervention-sensitivity and that other time series models can use these individual-level IBL parameters to improve their training efficiency.

Keywords

Limited resource allocation, Restless multi-armed bandit, Time-series forecasting, Cognitive modeling, Instance-based learning

1. Introduction

Public health programs play an essential role in improving the health outcomes of individuals and communities, often through education and subsequent behavioral change. Some health programs interact with their intended beneficiaries in a broad and infrequent manner. For example, a campaign about the health risks of smoking may address a general population of smokers through scattered advertisements in the media [1]. Others rely on repeated direct interactions with their intended beneficiaries—for example, a maternal health program that sends automated messages about exercise and nutrition to enrolled expectant mothers [2]. In this case, it is crucial that mothers remain engaged for the duration of the program or as long as possible to receive the maximum benefit. Unfortunately, many such programs face high levels of disengagement and dropout, which severely limit their effectiveness [3].

To reduce dropout, programs can provide interventions designed to increase beneficiary engagement. For example, program staff can make personalized service calls to beneficiaries at risk of dropping out to address concerns about declining participation. However, programs with limited staff and resources cannot provide interventions to all their beneficiaries. Instead, they must choose a subset of beneficiaries to receive the intervention.

Recent approaches have adopted the Restless Multi-Armed Bandit (RMAB) framework [4]. Classically, the RMAB framework models the engagement dynamics of each beneficiary (i.e., the transition between discrete engagement and disengagement states) as a Markov Decision Process (MDP). A centralized planner then recommends which beneficiaries to intervene on at each time point based on the dynamics learned from the MDPs. Importantly, beneficiaries can change their engagement state even without an intervention. The best-performing approach to this framework is the Whittle index [5], which ranks beneficiaries by their likelihood of becoming and remaining engaged after getting an intervention.

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ yseow@andrew.cmu.edu (R. Seow); yunfanzhao@fas.harvard.edu (Y. Zhao); djwood@andrew.cmu.edu (D. Wood); milind_tambe@harvard.edu (M. Tambe); coty@cmu.edu (C. Gonzalez)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This framework assumes that beneficiaries’ dynamics—how their level of engagement changes over time—can be modeled as a MDP. However, recent work has shown that these dynamics are often non-Markovian; that is, a beneficiary’s transition between being engaged and disengaged (their state) depends on their individual history [6]. As a substitute, Danassis and colleagues (2023) developed the Time-series Arm Ranking Index (TARI), which allows non-Markovian dynamics and continuous engagement levels within the RMAB framework. The TARI approach models beneficiaries using time-series forecasters such as LSTMs [7] or transformers [8] instead of MDPs. Using these learned time-series models, TARI computes an index that captures the relative engagement benefit of intervening on a beneficiary by comparing their expected future engagement with and without an intervention at the current time point and then ranks beneficiaries by the computed indices.

Within the TARI approach to the RMAB framework, we propose modeling individual beneficiaries with computational cognitive models instead of LSTMs. Specifically, we use models based on Instance-Based Learning Theory (IBLT) [9] to represent each beneficiary’s time series activities. We expect using an IBL model will convey two advantages over general time-series forecasters (TSFs).

First, IBL models are personalized to the individual instead of trained across the entire dataset, allowing the models to capture individual differences in behavioral dynamics (i.e., individual-level differences in the response to interventions and general engagement patterns). Because TSFs need large amounts of data to fine-tune their many parameters and there is relatively sparse data available per beneficiary, existing approaches train a single model using the combined data from all available beneficiaries. For instance, Danassis and colleagues (2023) create a set of fixed-length vectors from each beneficiary’s trajectory. Each vector represents the intervention history and engagement levels of seven consecutive timesteps. Then an LSTM trains on these vectors to predict the engagement level at the next timestep given the previous seven [6]. In contrast, an IBL model has enough cognitively grounded structure that it can be applied to an individual with sparse data [10]. This is because IBL theory makes strong assumptions about the cognition of human decision making derived from prior research. In particular, IBL theory accounts for memory effects (i.e., the more recent a data point is, the more it contributes to the prediction) and similarity effects (i.e., engagement levels that are observed under similar contexts should be similar to each other) that general TSFs do not.

Second, the parameters of IBL cognitive models are grounded in psychological constructs such as memory and attention, making them more interpretable than the parameters of general LSTMs. This advantage allows cognitive models to shed light on the individual differences that drive beneficiaries’ behaviors. We demonstrate how these individual differences can be used to augment existing approaches that rely primarily on general LSTMs.

1.1. Our Contributions

1. Using real-world data from a maternal health program, we show that cognitive models more accurately capture the temporal dynamics of individual behaviors and predict the behaviors of individual beneficiaries than LSTMs.
2. We show that personalized cognitive models reveal individual characteristics through weights fitting. We illustrate how to cluster individuals using learned characteristics.
3. We show clustering based on learned individual characteristics could guide and improve the performance of time series models. Specifically, we show that LSTMs trained in clusters better predict behaviors within those clusters compared to LSTMs trained on random samples of beneficiaries.

1.2. Background

1.2.1. Time-series Arm Ranking Index (TARI)

In the context of multiple potential beneficiaries treated as arms in an RMAB, each arm is considered an independent time series [6]. We train a model to predict the next state s_{t+1} based on three pieces of information:

- A historical record (length h) of the past states and actions of the arm (denoted by $\{(s_i, a_i)\}_{i < t}^h$).
- The current state (e.g., engagement level) of the arm s_t .
- A potential action a_t , where 1 represents an intervention and 0 represents no intervention.

This training takes place offline. During test time, the model's parameters are fixed, and we use the trained model with iterated multi-step forecasting to generate a long-term forecast of future states $s_{t+1}, s_{t+2}, \dots, s_{t+H}$ [11].

For each arm, we use the TSF model to estimate two values:

- Time to disengagement **with** intervention u_n : The predicted number of timesteps until the arm becomes non-engaged if we intervene at the current timestep (t) and never again.
- Time to disengagement **without** intervention v_n : The predicted number of timesteps until the arm becomes non-engaged with no interventions at any point.

The TARI index for an arm is the ratio of these two times. A higher TARI index suggests that intervening now would be more beneficial. Finally, similar to the Whittle index, we choose to act on the k arm with the highest TARI index at each timestep:

$$\text{TARI}(n) = \frac{u_n}{v_n}.$$

1.2.2. Instance-Based Learning Theory

Instance-Based Learning Theory (IBLT) is a cognitive theory of dynamic decision making, grounded in human learning and memory mechanisms. It has successfully accounted for human behavior in a variety of contexts, ranging from abstract repeated-choice tasks to more real-world search and choice [12] and continuous control tasks [9]. In most of these contexts, human participants are required to make a series of decisions while exposed to changing environments.

The essence of IBLT is storing past decisions in the form of an instance. An instance has three parts: the context, the choice made, and the associated utility. For example, these would correspond to the cloudiness of the sky, the decision to bring an umbrella, and the inconvenience of bringing an umbrella when it does not rain (supposing it did not rain in this instance).

IBL agents compute the expected utility of a choice by considering past instances where they made that choice, weighted by the contextual similarity to the present. For example, an IBL agent would compute the expected utility of bringing an umbrella by considering its past experiences with an umbrella, most strongly considering the experiences with the most similar cloudiness to the current moment.

More precisely, the expected value of a choice is the weighted average of the utilities from all the past instances where that choice was made weighted by the retrieval probability (i.e., the probability of remembering that instance) to generate estimates of the choice's expected utility called a blended value. The agent will deterministically make the available choice with the highest computed blended value (expected utility). The utilities of instances in memory are the actual outcome the agent observed when it chooses, regardless of what it predicted. After choosing, a new instance is added to the agent's memory, reflecting what happened. If the exact combination of context, choice, and utility has appeared before, that instance instead has its frequency and recency information updated.

The contribution of an instance's utility to a choice's expected value depends on the instance's memory activation (the salience of that instance given the current context). The activation of an instance (A_i) reflects how readily it comes to mind, which in turn depends on the frequency, recency, and similarity factors according to the following equation as proposed by the ACT-R cognitive architecture [13].

$$A_i = \ln\left(\sum_j (t - t'_{i,j})^{-d}\right) + \mu \sum_k (w_k(\text{Sim}(s_{i,k}, s_{t,k}) - 1) + \sigma\xi \quad (1)$$

Three expressions additively determine an instance's activation (salience in memory). The first expression is the contribution of frequency and recency where t is the current timestep, $t'_{i,j}$ is the

timestep of the j th appearance of instance i , and d is a decay parameter. The second expression is the contribution of similarity where μ is a scaling parameter on the overall impact of similarity to activation, w_k is the importance (or weight) of an attribute in the similarity calculations, Sim is a similarity function that returns the degree of similarity for a particular attribute k between instance i ($s_{i,k}$) and the current context ($s_{t,k}$). The third expression adds some randomness to the memory retrieval process where $\sigma\xi$ is a scaled noise distribution. The activation of an instance is then converted into a retrieval probability using the Boltzmann softmax function, where τ is the temperature parameter:

$$P_i = \frac{e^{A_i/\tau}}{\sum_j e^{A_j/\tau}} \quad (2)$$

The IBL model computes the expected utility of a choice (a) from the utility u_i of each retrieved instance i weighted by their respective retrieval probability P_i [9, 14]:

$$V(a) = \sum_i P_i u_i \quad (3)$$

The agent then makes the choice with the highest expected utility, $V(a)$.

2. Proposed Approach

We model each beneficiary with their own individual-level IBL model. We do not have the IBL model make a choice. Instead, we use the IBL’s expected utility calculation to predict the engagement level in the next period (see Figure 1a). Because we only use the model’s blended value, the choice part of the instance is always held constant. The utility of an instance is a beneficiary’s engagement level (thus, the expected utilities are weighted averages of past engagement levels then used to predict future engagement levels). The context has two components derived from a beneficiary’s history: (1) the engagement level of the previous period and (2) the number of timesteps since the last intervention. The IBL model determines the similarity between the current context and each instance in memory according to these two attributes, and these similarities are factored into an instance’s activation according to Equation 1.

The IBL models are trained through model tracing: recording each timestep of an individual beneficiary as an instance with the relevant context (previous engagement level and time since the last intervention) and utility (engagement level). Thus, the model “traces” the beneficiary it is modeling by giving itself memories as if it were the beneficiary that produced the data.

As shown in Equation 1, the similarity of two instances depends not only on the similarity between attribute values, but also on the weight or influence of the attribute (w in Equation 1). Psychologically, an attribute weight represents how relevant the agent considers an attribute to make similarity judgments between two contexts. To capture potential individual differences between beneficiaries, we personalized an IBL model for a specific beneficiary by finding the combination of attribute weight values that resulted in the smallest training prediction error.

Given the predictions of engagement level, we follow TARI and generate a ranked list of beneficiaries according to the predicted relative benefit in the improvement in engagement each beneficiary receives from an intervention versus not receiving one.

3. Simulation Setup

3.1. Data

We tested our approach on data from a maternal healthcare program operated by an NGO called ARMMAN [15]. This program sends automated messages about maternal and infant care to expectant and recent mothers in vulnerable communities. However, over the many weeks of pregnancy and postnatal care, many mothers stop engaging with the automated messages. ARMMAN has a limited

capacity to intervene: Program staff can call mothers individually, allowing them to hear from a real person and ask questions, hopefully increasing their future engagement.

ARMMAN collected the data in 2022 from 12,000 mothers over 40 weeks. The data have two recorded values per mother per week: the amount of time she spent listening to the automated health message (engagement level, recorded as a number in $[0, 1]$) and if she received a call from ARMMAN staff (if she received an intervention, recorded as a binary value). The data also contain mothers' demographic information; however, previous work found little or no prediction benefit in incorporating demographic variables as additional features for time-series forecasters [6]. Thus, we do not consider them in this approach.

For an IBL model to predict an intervention's effect on a mother's engagement level, it needs to be trained on a trajectory that includes at least one intervention. However, of the 12,000 mothers, only 5,400 mothers received at least one intervention early on in their program tenure. We also assume that mothers receive an intervention (or the equivalent of an intervention) upon their enrollment in the program. Because the time needed to train all the IBL models scales with the number of mothers, we explored our approach with a subset of 210 mothers out of the 5,400 mothers who received at least one intervention in their actual trajectories.

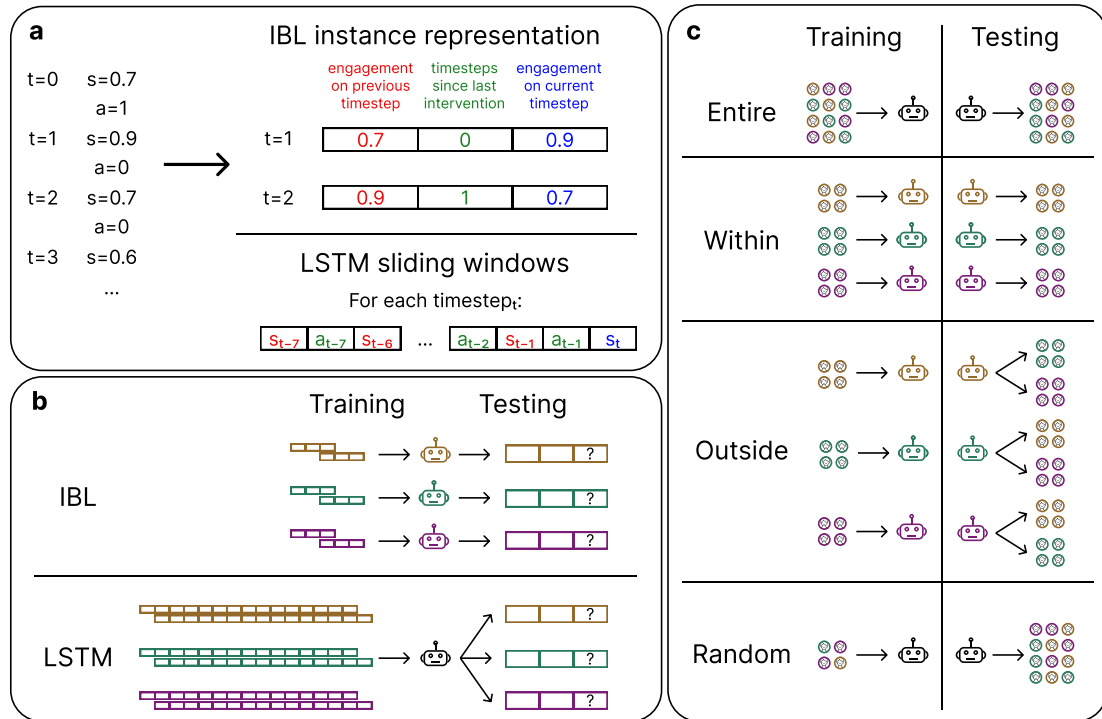


Figure 1: a) Transforming a beneficiary's trajectory into IBL instances and LSTM sliding windows where "s" is state (engagement level) and "a" is action (intervention status). Each IBL instance represents the engagement context, consisting of three parts: the engagement level in the immediately prior timestep, the number of timesteps since the last intervention, and the engagement level in the current timestep. The LSTM receives the state and action of the previous seven timesteps. **b)** Training / testing setup for next-step prediction. Different colors indicate data from individual beneficiaries. One IBL model is trained per beneficiary and the same model predicts for that beneficiary in testing. One LSTM is trained using data from all beneficiaries and that model predicts for all beneficiaries in testing. **c)** LSTM training / testing setups under various IBL-clusters-related conditions. Different colors indicate different clusters. The Random condition LSTM is trained on a random third of the training data to make it comparable to the Within and Outside methods.

3.2. Model Training

We designated the first 25 weeks of each mother’s trajectory as training data. Following the approach of previous studies, we reconstructed the training data into sliding windows of 7 consecutive timesteps to train the LSTM model. Then, we trained an LSTM model on the entire training dataset (see Figure 1b for a visual description of the differences between the IBL and LSTM training / testing procedure).

Each IBL model (one for each mother) traced a mother’s trajectory through each timestep up to week 25. To find the best-fitting profile of attribute weights per mother, a grid search looked for the combination that minimized the weighted sum of next-step prediction errors (loss). The grid iterates all possible combinations of parameter values in the range (0, 0) to (5, 5) in intervals of 0.5.

An IBL model should improve its predictive accuracy as it accumulates more instances and becomes more “familiar” with the individual it’s modeling. So, we weight earlier prediction errors (e.g., at weeks 1 and 2) less than later ones (e.g., at weeks 24 and 25) according to the following equations:

$$loss = \sum_t q_t * SE_t \quad (4)$$

$$q_t = \frac{e^{t/10}}{\sum_i e^{t_i/10}} \quad (5)$$

where SE_t is the squared prediction error and q_t is the normalized weight factor for the timestep t .

3.3. Next-step prediction task

For the next-step prediction task, the IBL and LSTM models generated predictions for the next 14 weeks, starting from week 26. The models iteratively generated predictions for the next timestep (i.e., for week 26, then for week 27, then for week 28, etc.). In our data, ARMANN did not provide interventions to mothers past 14 weeks, so there are no interventions during the testing period for the next-step prediction task. This allows us to compare the models’ engagement predictions with the ground truth data without the need to generate synthetic counterfactuals.

3.4. Model comparison with simulated counterfactuals

In a separate analysis, we compared the IBL-TARI and LSTM-TARI policies with three other baseline policies: (i) myopic uniformly random allocation (ii) round-robin allocation, which iteratively cycles through the population, and (iii) no interventions. Consistent with ARMMAN’s limited resources, all policies are allowed a 3% budget within a timestep (except the no intervention policy), i.e., 6 beneficiaries per timestep.

To simulate the effect of these policies’, we followed previous studies and trained a separate LSTM model on trajectories from the entire dataset (approximately 12000 beneficiaries)—far larger than the policy LSTM’s training set—to generate mothers’ counterfactual engagement. This counterfactual generator only operates when a mother’s simulated trajectory deviates from her ground-truth trajectory (i.e., received a simulated intervention when she did not receive one in reality or not receiving a simulated intervention when she did receive one in reality).

4. Results

4.1. Predicting next-step engagement levels

Figure 2 shows the prediction error of the two types of models during the 14 week the testing period of the next-step prediction task. The IBL models consistently achieve lower prediction errors than the LSTM—about 10% less mean error (0.23 on average) compared to the LSTM model (0.32 on average).

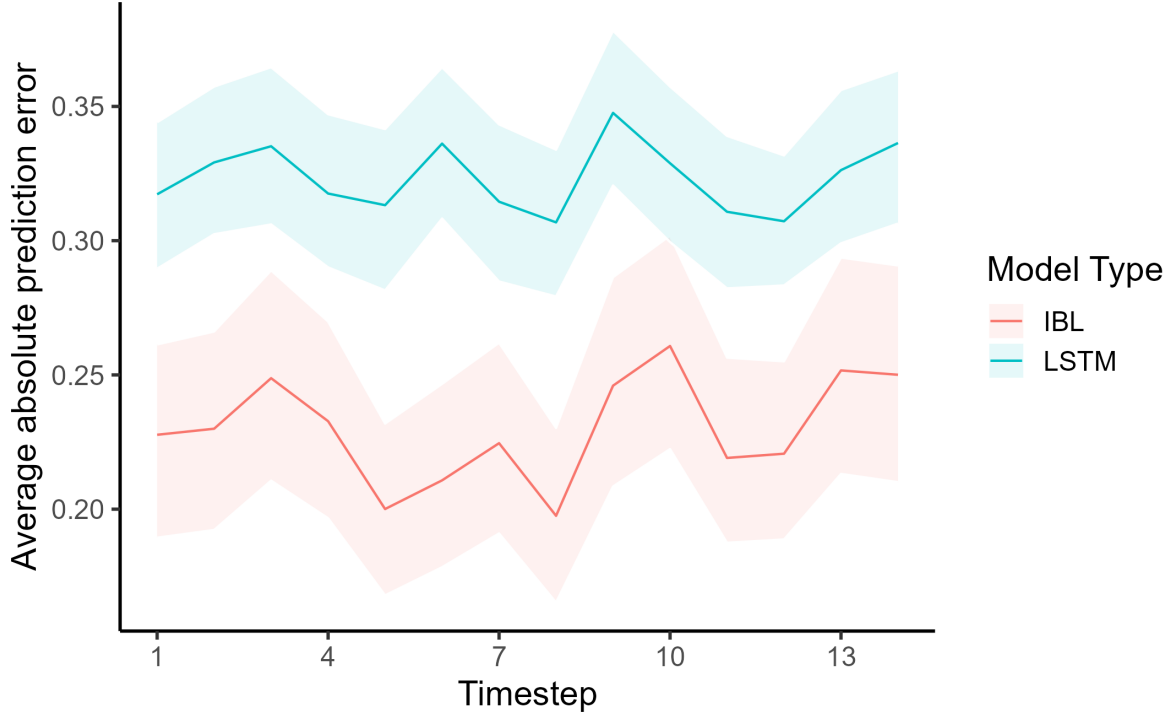


Figure 2: Next-step prediction error per testing timestep (smaller is better). $N = 210$. The personalized IBL models consistently achieve lower average prediction error (0.23 across all timesteps) than the LSTM model (0.32)—a 29% reduction in error.

4.2. Using predictions to inform intervention allocation

Having demonstrated that IBL models yield more accurate next-step predictions of engagement levels compared to the LSTM model, we employed simulations to examine if integrating IBL models with the TARI index would result in higher overall levels of engagement. Figure 7 (in Appendix) shows the percentage of mothers who have an engagement level of 0.25 and above during each test timestep following the various intervention allocation policies. The LSTM-TARI policy outperforms all other policies, particularly later on in the testing period. The average percentage of engaged mothers across testing timesteps for the various policies are as follows: IBL-TARI (55.58%); Round-robin (55.65%); Random (54.29%); LSTM-TARI (61.05%); Control (46.22%). Generating counterfactuals using an LSTM model may explain why the IBL’s advantage in next-step prediction does not translate to higher counterfactual engagement in the simulated task.

4.3. Describing individual differences with IBL parameters

Another advantage of the proposed IBL approach over more general TSFs (as exemplified by the LSTM approach) is identifying individual differences in beneficiaries’ behavioral characteristics. Because we model each beneficiary with a separate IBL model, the models’ parameter values capture how beneficiaries differ from each other in their engagement and intervention dynamics.

Specifically, the attribute weights of each mother’s IBL model trained on data from the first 25 weeks tracks (1) how reliably her engagement changes in response to an intervention (intervention-sensitivity) and (2) how consistently her engagement changes between timesteps (transition-consistency), i.e., how similarity between the engagement levels in different timesteps corresponds to similarity between the engagement levels in their following timesteps.

Figure 3 plots mothers according to their attributes weights. Visually, there appear to be three types of mothers. We confirmed that the optimal number of clusters is between 3 to 4 clusters using multiple measures of cluster validity indices (silhouette: 4, within-sum-of-squares: 4, gap statistic: 3).

For interpretability, subsequent analyses assume three clusters.

Each cluster was labeled according to its defining characteristics. Intervention-sensitive mothers ($N = 62$) have a high intervention lag weight; i.e., a mother’s engagement level t timesteps after her most recent intervention will be similar to other times she was t timesteps after the most recent intervention. Transition-consistent mothers ($N = 66$) have a high previous engagement weight; i.e., if engagement levels at timesteps $i - 1$ and $j - 1$ (x_{i-1} and x_{j-1}) are similar to each other, then we should expect engagement levels x_i and x_j to also be similar to each other. State-stable mothers ($N = 82$) have low values in both dimensions; i.e., these mothers will tend to engage at the level they engage most frequently and most recently.

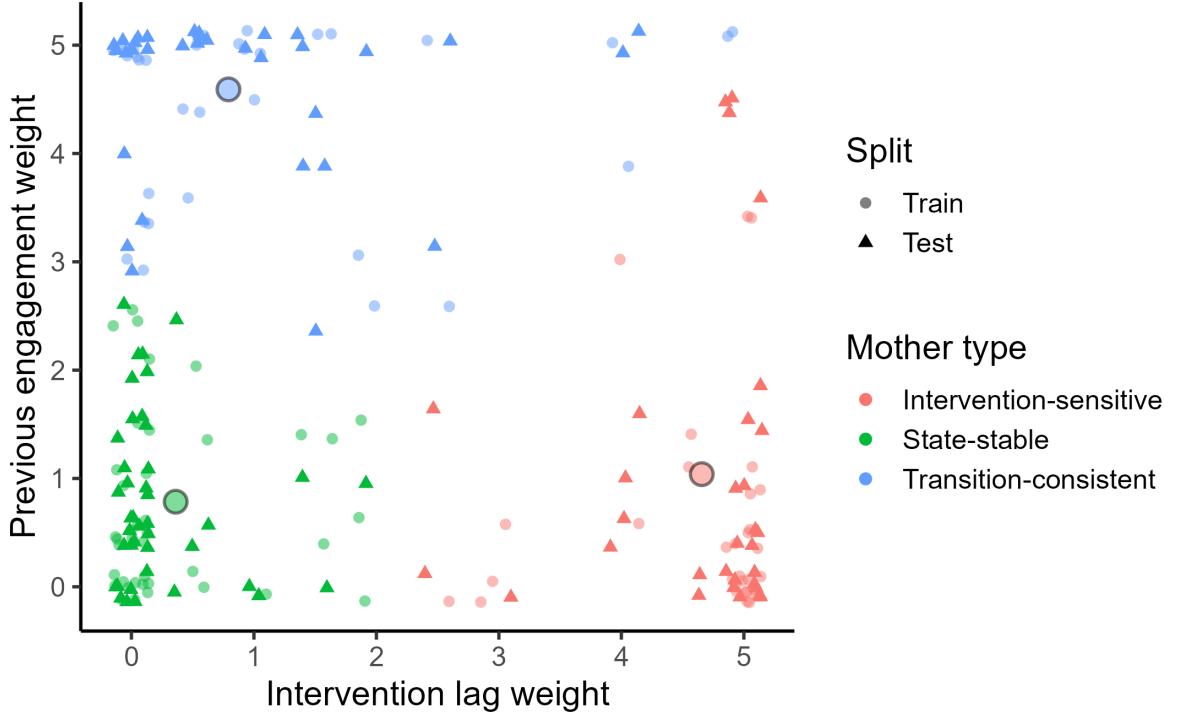


Figure 3: Distribution of attribute weight profiles. Each dot represents a specific mother’s IBL model. Points are slightly jittered to better visualize density. Lighter points indicate training set mothers while darker points indicate testing set mothers. Beneficiaries are clustered into 3 clusters according to a k-means algorithm. Larger circles outlined in black indicate cluster centroids.

4.3.1. Testing the generalizability of individual differences in IBL attribute weights

The previous results showed that individual mothers have different optimal weight profiles for the IBL instance similarity attributes. However, we still must show these weights capture individual differences that predict behavioral dynamics beyond an IBL modeling context. If these attribute weights reflect underlying individual differences in engagement and intervention dynamics, then we would expect other modeling approaches that account for these attribute weights to perform better than those that do not.

We tested this hypothesis by training and testing LSTM models with cluster-related samples. Using the cluster assignments described in the previous section, we randomly split each cluster in half to construct a training and a testing set. For all methods, the task is to predict the next-state engagement of the testing set during the testing period (i.e., the final 14 weeks). We examined the following methods (see 1c for a visual description):

- **Entire:** An LSTM trained on all training mothers and predicting for all test mothers.

- **Within-cluster:** An LSTM trained on each cluster and predicting for test mothers in their respective training cluster. We hypothesize that this method should yield the most accurate predictions given the informativeness of the identified IBL clusters.
- **Outside-cluster:** An LSTM trained on each cluster and predicting for test mothers from a different cluster. Because there are 3 clusters, we average prediction error from running the model on both out-of-cluster testing sets.
- **Random:** An LSTM trained on a random subset of one-third of the training set ($N = 35$) and predicting for all test mothers

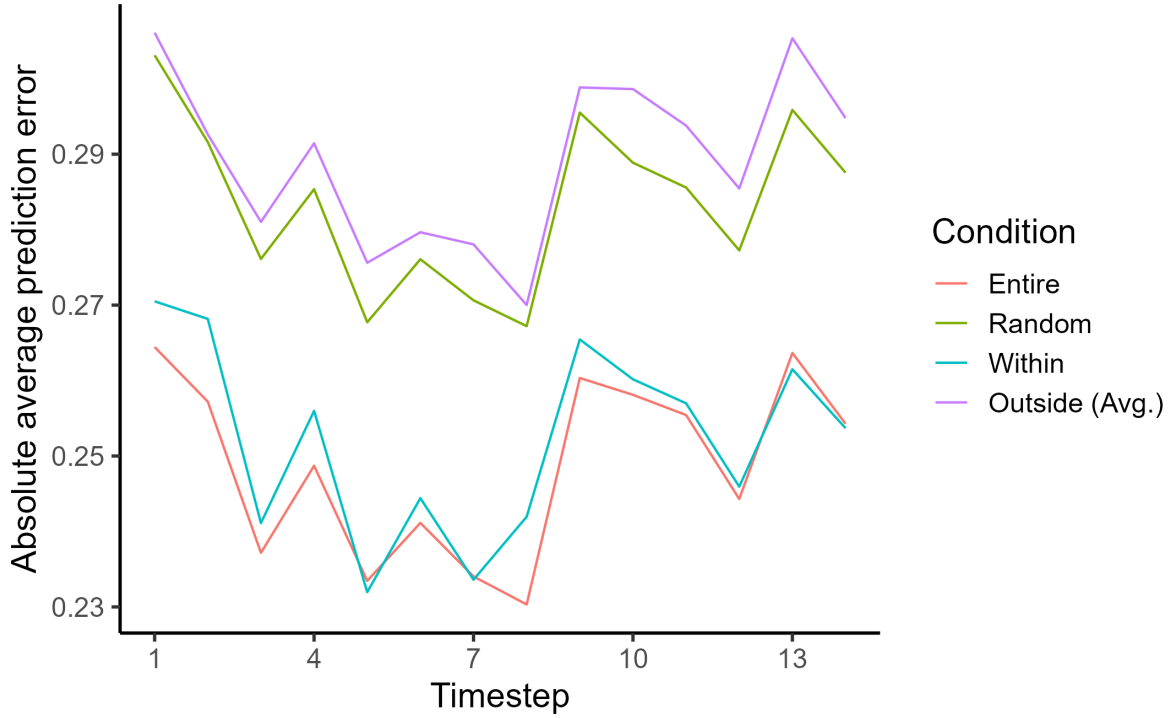


Figure 4: Next-step prediction error per testing timestep for the various LSTM training-testing methods. Outside-cluster method error is the average error of a model trained on each of the three testing sets tested on each of the two other clusters (six total combinations).

Figure 4 shows the average prediction error for the various methods. Supporting the generalizability of the IBL attribute weight clusters, LSTMs trained and tested on different clusters (Outside-cluster) and the LSTM trained on a random sample (Random) performed worse than LSTMs trained and tested on the same clusters (Within).

The Within-cluster and Entire methods yield similar average prediction accuracies (despite each Within-cluster LSTM having less training data). However, as shown in Figure 5, the Within-cluster method produces more accurate predictions than the Entire method for two-thirds of the mothers (Within 70 vs. Entire: 35), which suggests that the clusters effectively categorize the majority of the population.

Follow-up exploratory analyses suggest that mothers closer to their cluster’s centroid are better predicted by the Within-cluster method as compared to the Entire method. Using a distinct binomial regression model for each cluster, we used the Euclidean distance between a mother’s attribute weights and her centroid to predict the method (Within-cluster or Entire) with lower prediction error. The effects were statistically insignificant, plausibly due to the imbalanced and small classes. Still, for State-stable ($\beta_{dist} = -0.8637$) and Transition-consistent mothers ($\beta_{dist} = -1.235$), the closer they were to the cluster centroid, the more likely the Within-cluster method gave better predictions than the Entire method (see Figure 6).

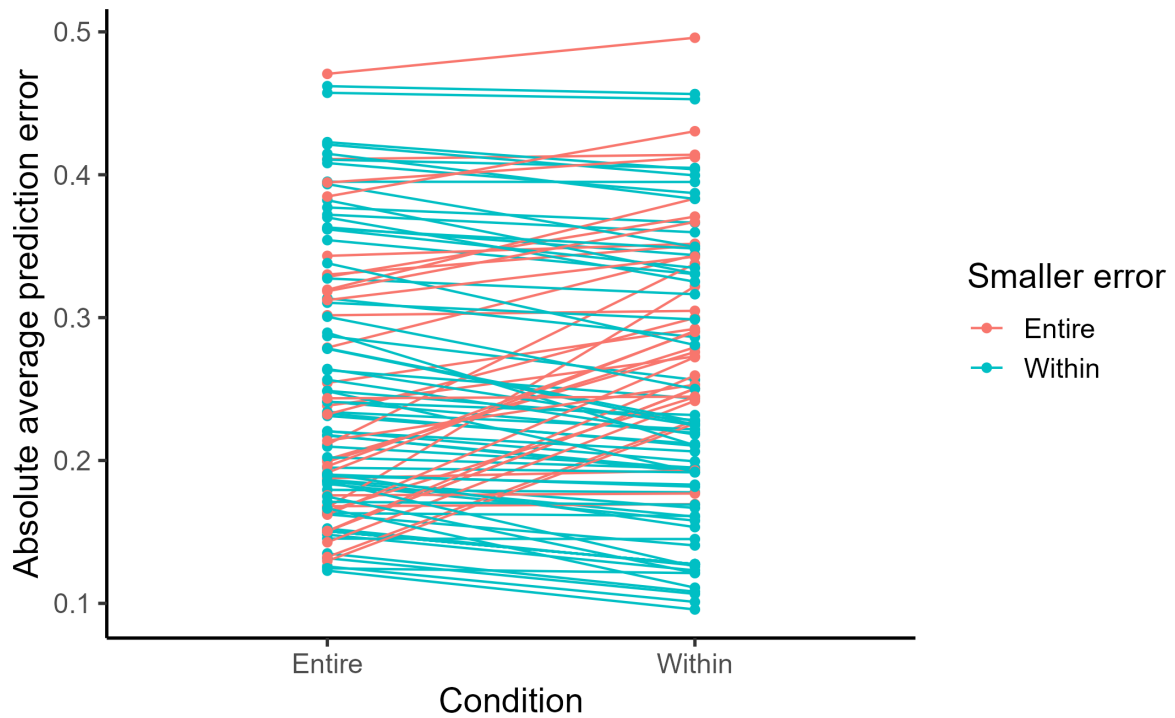


Figure 5: Comparison of prediction errors between Entire and Within-cluster methods. Points show average prediction error for individual mothers in the testing set and lines indicate the same mother across conditions. Blue lines indicate mothers best predicted by the Within-cluster method. Red lines indicate mothers best predicted by the Entire method.

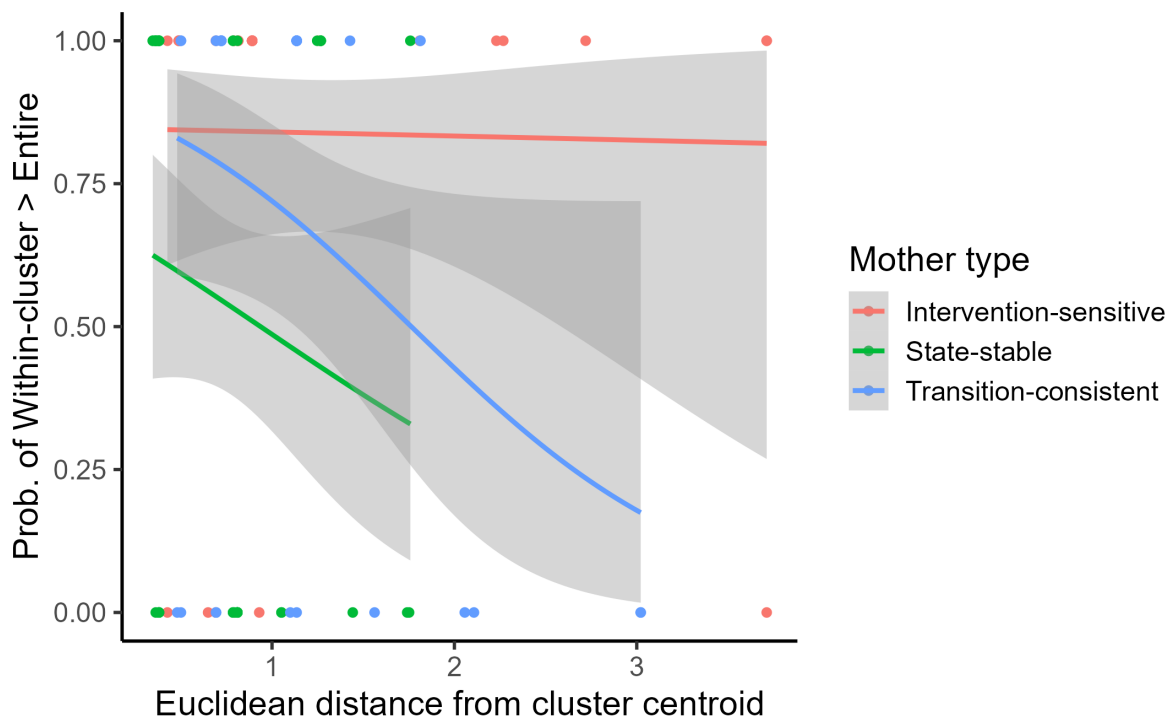


Figure 6: Probability the Within-cluster method had lower prediction error than the Entire method. Probability 1 means the Within-cluster method best predicts a mother's engagement. Lines show predictions by binomial regression models, and shaded areas reflect the 95% CIs.

5. Conclusions and Future work

Public health programs with limited resources to provide interventions must face the difficult problem of determining which beneficiaries to give interventions. Prior work demonstrated that the TARI algorithm—which relies on the predictions of time-series forecasters—outperforms existing recommendation algorithms [6]. To improve the effectiveness of TARI’s recommendations through more accurate predictions, we developed a novel approach that uses computational cognitive models derived from Instance-Based Learning Theory [9, 14] to model the human transition dynamics of program engagement. Given the non-Markovian nature of a beneficiary’s transitions between states, a cognitive algorithm that reflects the effects of time, attention, and context similarity on human memory and decisions provides more accurate predictions of beneficiaries’ engagement dynamics and more meaningfully describes relevant individual differences that other prediction methods can leverage.

We tested our method on real-world engagement data from mothers enrolled in a maternal health program, finding it results in higher next-step prediction accuracy compared to existing time-series methods using LSTM models. The personalized IBL models revealed that mothers clustered into three types based on their IBL attribute weights—specifically, how their engagement patterns could be attributed to intervention-sensitivity and transition-consistency. These clusters generalized to other time-series models: LSTMs trained and tested on within-cluster data outperformed all other LSTM training methods for most mothers.

Future work could explore recommending personalized interventions and intervention schedules based on these identified individual differences in behavioral characteristics. For example, intervention-sensitive beneficiaries could receive more frequent but shorter calls as the model predicts that each call would provide a substantial boost to their engagement levels. In comparison, state-stable beneficiaries who are likely to maintain their engagement levels over time could receive fewer but more intensive calls that maximize their engagement levels by the end of each call. These beneficiaries would then be able to sustain a high engagement level until the next time they receive a call.

One limitation of this work is the counterfactual generator used to simulate the effect of recommended interventions—an LSTM model trained on a much larger dataset. The higher next-step prediction accuracy of the IBL model suggests that the LSTM architecture may systematically fail to pick up certain cognitively-determined regularities in the human data. Thus, the LSTM-generated counterfactuals may also fail to produce simulated data with those regularities. This might explain why IBL’s next-step prediction advantage over the LSTM does not translate to better performance in the model comparison with simulated counterfactuals. Future work should prioritize testing our method with other datasets, particularly with a paradigm where the effects of recommended interventions on participants’ task engagement can be experimentally measured, rather than relying on plausibly biased counterfactuals.

Another limitation is that our method uses the first 25 weeks of the program for model training, only making predictions from the 26th week of a mother’s pregnancy onward. Possible extensions of this work could explore hybrid methods, such as initially relying on predictions from other trained IBL models and gradually transitioning to the fully personalized model.

Acknowledgments

This research was supported by the AI Research Institutes Program funded by the National Science Foundation under AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881

References

- [1] S. O. Nogueira, A. McNeill, M. Fu, C. N. Kyriakos, U. Mons, E. Fernández, W. A. Zatoński, A. C. Trofor, T. Demjen, Y. Tountas, et al., Impact of anti-smoking advertising on health-risk knowledge and quit attempts across 6 european countries from the eurest-plus itc europe survey, *Tobacco Induced Diseases* 16 (2018). doi:10.18332/tid/96251.

- [2] A. Lalan, S. Verma, P. R. Diaz, P. Danassis, A. Mahale, K. M. Sudan, A. Hegde, M. Tambe, A. Taneja, Improving health information access in the world's largest maternal mobile health program via bandit algorithms, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024, pp. 22913–22919. doi:<https://doi.org/10.1609/aaai.v38i21.30329>.
- [3] S. Amagai, S. Pila, A. J. Kaat, C. J. Nowinski, R. C. Gershon, Challenges in participant engagement and retention using mobile health apps: literature review, *Journal of medical Internet research* 24 (2022) e35120. doi:[10.2196/35120](https://doi.org/10.2196/35120).
- [4] J. A. Killian, A. Biswas, L. Xu, S. Verma, V. Nair, A. Taneja, A. Hegde, N. Madhiwalla, P. R. Diaz, S. Johnson-Yu, et al., Robust planning over restless groups: engagement interventions for a large-scale maternal telehealth program, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023, pp. 14295–14303. doi:<https://doi.org/10.1609/aaai.v37i12.26672>.
- [5] P. Whittle, Restless bandits: Activity allocation in a changing world, *Journal of applied probability* 25 (1988) 287–298. doi:[10.2307/3214163](https://doi.org/10.2307/3214163).
- [6] P. Danassis, S. Verma, J. A. Killian, A. Taneja, M. Tambe, Limited resource allocation in a non-markovian world: the case of maternal and child healthcare, *arXiv preprint arXiv:2305.12640* (2023). doi:<https://doi.org/10.48550/arXiv.2305.12640>.
- [7] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (1997) 1735–1780.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [9] C. Gonzalez, J. F. Lerch, C. Lebiere, Instance-based learning in dynamic decision making, *Cognitive Science* 27 (2003) 591–635. doi:https://doi.org/10.1207/s15516709cog2704_2.
- [10] E. H. Bugbee, C. Gonzalez, Making predictions without data: How an instance-based learning model predicts sequential decisions in the balloon analog risk task, in: *Proceedings of the annual meeting of the cognitive science society*, volume 44, 2022.
- [11] S. B. Taieb, R. J. Hyndman, et al., Recursive and direct multi-step forecasting: the best of both worlds, volume 19, Department of Econometrics and Business Statistics, Monash Univ., 2012.
- [12] T. N. Nguyen, C. Gonzalez, Theory of mind from observation in cognitive models and humans, *Topics in cognitive science* 14 (2022) 665–686. doi:<https://doi.org/10.1111/tops.12553>.
- [13] J. R. Anderson, C. J. Lebiere, The atomic components of thought, Psychology Press, 2014. doi:<https://doi.org/10.4324/9781315805696>.
- [14] C. Gonzalez, V. Dutt, Instance-based learning: integrating sampling and repeated decisions from experience., *Psychological review* 118 (2011) 523. doi:<https://doi.org/10.1037/a0024558>.
- [15] ARMMAN, mmitra - armman - helping mothers and children, 2024. URL: <https://armman.org/mmitra/>.
- [16] A. Mate, L. Madaan, A. Taneja, N. Madhiwalla, S. Verma, G. Singh, A. Hegde, P. Varakantham, M. Tambe, Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022, pp. 12017–12025. doi:<https://doi.org/10.1609/aaai.v36i11.21460>.
- [17] Y. Zhao, T. Wang, D. Nagaraj, A. Taneja, M. Tambe, The bandit whisperer: Communication learning for restless bandits, *arXiv preprint arXiv:2408.05686* (2024). doi:<https://doi.org/10.48550/arXiv.2408.05686>.
- [18] G. Xiong, J. Li, Finite-time analysis of whittle index based q-learning for restless multi-armed bandits with neural network function approximation, in: *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [19] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, N. De Freitas, Social influence as intrinsic motivation for multi-agent deep reinforcement learning, in: *International conference on machine learning*, PMLR, 2019, pp. 3040–3049.
- [20] N. Modi, P. Mary, C. Moy, Transfer restless multi-armed bandit policy for energy-efficient heterogeneous cellular network, *EURASIP Journal on Advances in Signal Processing* 2019 (2019) 1–19. doi:<https://doi.org/10.1186/s13634-019-0637-1>.
- [21] Y. Zhao, N. Behari, E. Hughes, E. Zhang, D. Nagaraj, K. Tuyls, A. Taneja, M. Tambe, Towards

- zero shot learning in restless multi-armed bandits, AAMAS (2023). doi:<https://doi.org/10.48550/arXiv.2310.14526>.
- [22] N. Behari, E. Zhang, Y. Zhao, A. Taneja, D. Nagaraj, M. Tambe, A decision-language model (dlm) for dynamic restless multi-armed bandit tasks in public health, arXiv preprint arXiv:2402.14807 (2024). doi:<https://doi.org/10.48550/arXiv.2402.14807>.
- [23] S. Verma, Y. Zhao, S. Shah, N. Boehmer, A. Taneja, M. Tambe, Group fairness in predict-then-optimize settings for restless bandits, 2024.
- [24] E. Lee, M. S. Lavieri, M. Volk, Optimal screening for hepatocellular carcinoma: A restless bandit model, *Manufacturing & Service Operations Management* 21 (2019) 198–212. doi:<https://doi.org/10.1287/msom.2017.0697>.
- [25] T. Ayer, C. Zhang, A. Bonifonte, A. C. Spaulding, J. Chhatwal, Prioritizing hepatitis c treatment in us prisons, *Operations Research* 67 (2019) 853–873. doi:<https://doi.org/10.1287/opre.2018.1812>.

A. Real-world ARMMAN Dataset

A.1. Secondary Analysis

Our experiment falls into the category of secondary analysis of the data shared by ARMMAN. This paper does not involve deploying the proposed algorithm or any other baselines to the service call program. As noted earlier, the experiments are secondary analyses with approval from the ARMMAN ethics board.

A.2. Consent and Data Usage

Consent is obtained from every beneficiary enrolling in the NGO’s mobile health program. The NGO owns the data collected through the program and only the NGO is allowed to share data. In our experiments, we use anonymized call listenership logs to calculate empirical transition probabilities. No personally identifiable information (PII) is available to us. The data exchange and usage were regulated by clearly defined exchange protocols, including anonymization, read-access only to researchers, restricted use of the data for research purposes only, and approval by ARMMAN’s ethics review committee.

A.3. Interventions in ARMMAN data

The interventions in ARMMAN data are chosen based on restless multi-arm bandit (RMAB) algorithms [16, 17]. RMABs are a model for sequentially distributing scarce resources to a set of agents [5, 18, 19]. Concretely, we have a set of arms and a limited budget and face the question of deciding which arms to pull in each round. The state of arms evolves according to a Markov Decision Process where transition probabilities depend on whether the arm is pulled in this step. RMABs have a broad range of applications, including resource allocation in anti-poaching, machine maintenance, cellular networks [20, 21]. RMABs have had extensive use in healthcare settings such as call scheduling in a maternal and child care program [22, 23], screening patients at risk of cancer [24], and allocating hepatitis C treatment [25].

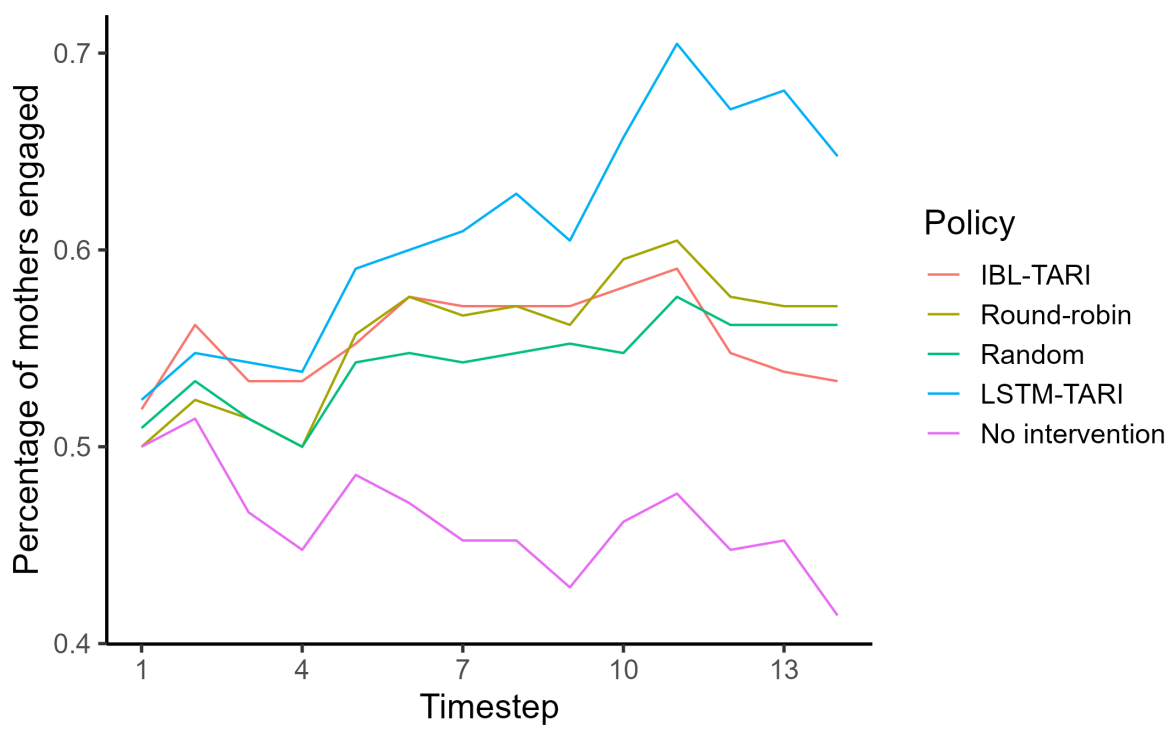


Figure 7: Percentage of mothers engaged in the program per testing timestep using simulated counterfactuals. $N = 210$, budget $k = 6$.

Advancing Visual Food Attractiveness Predictions for Healthy Food Recommender System

Ayoub El Majjodi^{1,*}, Sohail Ahmed Khan¹, Alain D. Starke^{1,2}, Mehdi Elahi¹ and Christoph Trattner¹

¹MediaFutures, University of Bergen, Lars Hilles Gate 30, Bergen, Norway

²Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, Netherlands

Abstract

The visual representation of food on digital platforms affects the foods chosen by users, including in the context of recommender systems. Previous studies show that small changes in visual features can influence human decision-making, regardless of whether the food is healthy. This paper reports on a study aimed at better understanding how users perceive the attractiveness of food recipe images in the digital world. In an online mixed-methods survey ($N = 192$), users provided visual attractiveness ratings of food images on a 7-point scale, along with textual assessments. We found robust correlations between fundamental visual features (e.g., contrast, colorfulness) and perceived image attractiveness. The analysis also revealed that, among other user factors, cooking skills positively affected perceived image attractiveness. Regarding food image dimensions, appearance and perceived healthiness were significantly correlated with user ratings of food image attractiveness.

Keywords

Food recommender systems, User modeling, Image attractiveness, Health, Personalization, Digital nudges

1. Introduction

Visual cues and attractiveness play a crucial role in everyday food choices [1]. Even when only presented with a food image, humans tend to instantly assess a food's energy density, expected taste and other characteristics [2]. As such, images are one of the key determinants of food preferences [2, 3], tapping into emotional and hedonic processes of an individual [4].

The importance of visual attractiveness also applies to digital choice context, including food recommender systems [5]. Our previous research has shown the capability of recommender systems to influence food behaviors via visual features, including the promotion of either high-fat or low-fat food choices [6], as well as encouraging the search for healthier options [3]. Additionally, our earlier work has established that visual attractiveness significantly contributes to predicting the online popularity of food items [7], and these visual features can also be leveraged to infer cultural backgrounds [8].

What is currently missing is an in-depth examination of image feature modeling. Although previous studies have extracted image features and examined the relation between those features, visual attractiveness and user preferences [6, 3], these models have not been optimized. Moreover, to date, image features have not been related to user characteristics (e.g., demographics, food knowledge), which are also important determinants of food preferences [9].

We present the results of a mixed-method study that explores the determinants of visual attractiveness in digital recipe images more comprehensively. Our approach builds upon previous work by modeling perceived visual attractiveness based on low-level image features [10, 11, 3]. Additionally, we seek to

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ ayoub.majjodi@gmail.com (A. E. Majjodi); sohail.khan@uib.no (S. A. Khan); a.d.starke@uva.nl (A. D. Starke); mehdi.elahi@uib.no (M. Elahi); christoph.trattner@uib.no (C. Trattner)

🌐 <http://ayoubmajjodi.info/> (A. E. Majjodi); <https://mediafutures.no/2021/09/13/sohail-ahmed-khan/> (S. A. Khan); <https://mediafutures.no/2021/06/21/alain-starke/> (A. D. Starke); <https://mediafutures.no/2020/11/25/mehdi-elahi/> (M. Elahi); <https://christophtrattner.info/> (C. Trattner)

🆔 0000-0002-7478-5811 (A. E. Majjodi); 0000-0001-5351-2278 (S. A. Khan); 0000-0002-9873-8016 (A. D. Starke); 0000-0003-2203-9195 (M. Elahi); 0000-0002-1193-0508 (C. Trattner)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

optimize this model by integrating user characteristics that have been employed in knowledge-based food recommender systems to promote healthier recipe choices [12, 13, 14].

Finally, we inquire more qualitatively on user justifications for provided visual attractiveness ratings, asking to motivate their quantitative judgment. We formulate the following research questions:

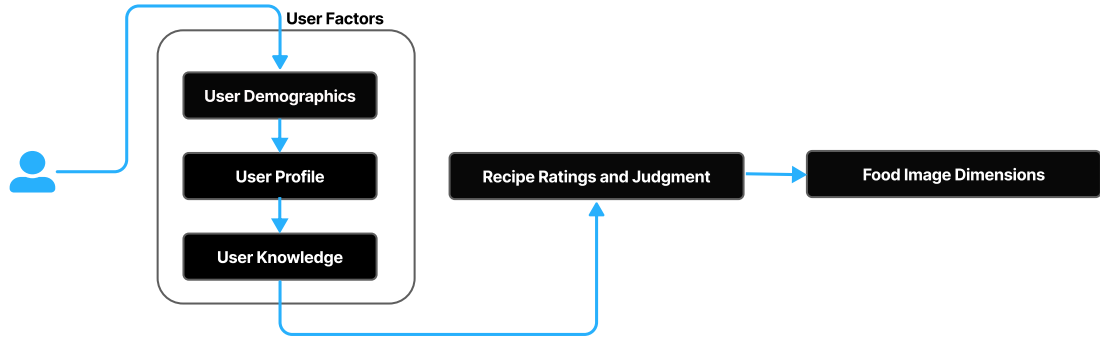


Figure 1: Steps of the user flow designed for the online survey.

- **RQ1:** To what extent do the latest deep learning methods predict visual attractiveness compared to state-of-the-art low-level features?
- **RQ2:** To what extent do user characteristics, including demographics, food knowledge, and eating goals, predict food image attractiveness?
- **RQ3:** What dimensions determine the attractiveness of food image?

1.1. Contributions

Compared to our extensive previous work in the field of visual attractiveness and food recommender systems [6, 7, 8, 13], this study offers novel insights into several key aspects:

- Previous work mostly relied on low-level image attractiveness features, while this study shows how new deep-learning models compare to these old features.
- This work, compared to any before, also shows as to what extent demographic features play a role in predicting visual food attractiveness. To our knowledge, no other work has shown this before.
- Finally, this study tries to go beyond traditional quantitative black box approaches and reveals why images are rated less or more attractive.

2. Study Design

To perform our study, we employed a dataset sourced from the well-known recipes website All-Recipes.com, with the addition of new recipe photos [3, 14]. The dataset comprised various recipe features, including image URL, ingredients, amount of fats and sugar, and instructions and ingredients. To generate a diverse set of images, we randomly selected 200 recipes with relatively from the dataset of 58,000. As most images in this dataset were relatively unattractive [3], we used the recipe’s title in search engines and image websites (e.g., Unsplash) to look for more attractive images for 100 of these recipes. To validate this process, three computational food researchers, including a co-author, voted on which of the two photos was the most attractive to ensure a diverse set of recipe images in terms of expected attractiveness.

The study involved a survey design, as depicted in Figure 1. Participants first provided demographic information, as well as responded to items that measured their subjective food knowledge (4 items) and cooking skills (6 items), using 5-point Likert scales based on earlier work [15, 16, 17]. We also used questions from earlier work on a knowledge-based food recommender [14], to inquire on other

user characteristics, including recipe website usage and home cooking frequency, cooking experience and dietary goals. Afterwards, users were invited to rate the visual attractiveness of 12 semi-randomly selected recipe images, on 7-point attractiveness scales. In addition, to address [RQ3], they were asked to write at least one sentence about why they had given this rating. Finally, to support our examination of [RQ3], we used 5-point Likert scales on food image dimensions [18], to ask to what extent a recipe’s appearance, expected taste, healthiness, and familiarity affected their attractiveness ratings.

We employed the Prolific crowdsourcing platform to recruit 192 users (65% male; $M_{age} = 33.54$) to participate in our study. The study took approximately 11 min to complete and participants were reimbursed with GBP 1.65¹.

Table 1

Linear regression models predicting visual attractiveness ratings for recipe images: (A) with low-level image visual features, (B) with deep learning-based visual features. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

(A)		(B)			
Low-level Image Features		Image features Extractor			
	β (S.E)	VGG16	ResNet	CLIP	
Colourfulness	6.725 (1.521)***				
Brightness	2.136 (0.155)***				
Naturalness	1.925 (0.530)***				
Entropy	1.026 (0.154)***				
Saturation	−3.976 (1.020)***	R^2	0.351***	0.349***	0.357***
Sharpness	−1.182 (1.187)*	RMSE	1.500	1.491	1.501
RGBContrast	−1.782 (3.808)				
Contrast	7.401 (11.101)				
Constant	−6.884 (1.243)***				
R^2	0.110***				
RMSE	1.753				

3. Results

To address our research questions, we primarily employed linear regression models. This helped to understand the principal impacts of image attributes and user characteristics on image attractiveness, the latter derived from user ratings. For our thematic analysis, the images were split into attractive and unattractive based on the mid-point of the rating scale (4) ($M = 4.33$, $SD = 1.80$). Details of used materials and conducted analyses can be accessed through the following URL [19].

3.1. RQ1: Predicting Visual Attractiveness

We first modeled perceived visual attractiveness based on the underlying image features. We extracted diverse low-level visual features using the OpenIMAJ Java Framework (cf. [7]). Subsequently, we conducted a linear regression analysis to predict attractiveness based on these extracted visual features. The results are outlined in Table (1.A), revealing that several image features significantly affected the attractiveness of a recipe image: $F(8, 2100) = 32.66$, $p < 0.001$. Specifically, Colourfulness, Brightness, Naturalness, and Entropy demonstrated a positive association with image attractiveness. In contrast, Saturation, Sharpness, and RgbContrast negatively affected image attractiveness. In line with [3], these results suggested that users perceived colorful, bright, and naturalistic food images as more attractive.

¹Our study complied with the ethical guidelines of the Research Council of Norway and the guidelines of University of Bergen for scientific research. It was judged to pass without further extensive review.

Going beyond low-level visual image features, we used deep learning architecture models. Our toolkit included established models, such as VGG16 [20] and ResNet [21], along with the well-known transformer [22] architecture for visual feature extraction, CLIP² [23]. Table (1.B) outlines the performance of these different models, outperforming our regression model in terms of R^2 and RMSE. This aligns with previous research where deep learning embeddings also outperformed low-level visual features within the context of food [7, 24].

3.2. RQ2: User characteristics and Image Attractiveness

We further examined whether user factors affected the perceived visual attractiveness of images. Accordingly, we divided user characteristics into different categories: User demographics, User profile, which represented the backbone of a food knowledge-based recommender system, and User knowledge, which measures the user’s food knowledge and cooking skills. A confirmatory factor analysis, reported in Table 2, showed that both subjective food knowledge and cooking skills adhered to internal consistency guidelines ($\alpha > 0.70$) while they also met the guidelines for convergent validity ($AVE > 0.5$).

Table 2

Results of the principal component factor analysis across different subjective food knowledge and cooking skills. Items were measured on 5-point Likert scales. Cronbach’s Alpha is denoted by α , AVE is the average variance explained. Items in grey and without loading were omitted.

Aspect	Item	Loading
Subjective Food Knowledge $\alpha = 0.866$ $AVE = 0.858$	Compared with an average person, I know a lot about healthy eating.	0.777
	I think I know enough about healthy eating to feel pretty confident when choosing a recipe.	0.885
	I know a lot about how to evaluate the healthiness of a recipe.	0.773
	I do not feel very knowledgeable about healthy eating.	0.932
	I can confidently cook recipes with basic ingredients.	0.751
Cooking skills $\alpha = 0.783$ $AVE = 0.591$	I can confidently follow all the steps of simple recipes.	
	I can confidently taste new foods.	0.737
	I can confidently cook new foods and try new recipes.	0.869
	I enjoy cooking food.	0.655
	I am satisfied with my cooking skills.	0.816

Table (3.A) presents the outcomes of the linear regression model aimed at forecasting the attractiveness of image recipes: $F(9, 2090) = 3.60$. Among the various user factors examined, only two significantly affected recipe attractiveness: cooking skills ($\beta = 0.34$, p-value= 0.00021) and recipe website usage ($\beta = 0.18$, p-value= 0.020). However, none of the other user aspects affected user ratings for a given image recipe. Additionally, we also analyzed a combined model of image features and user factors, but this lead to results similar to the separate models reported in Tables (1 and 3.A). This suggested that low-level visual features had a more significant impact on food image attractiveness than user features, largely in line with preliminary findings in previous research [3, 18].

3.3. RQ3: Justifications for Visual Attractiveness

To assess the influence of different food image dimensions on user ratings for food images, we modeled visual attractiveness based on the reported importance of food image dimensions. Table 3 outlines the results of the regression model: $F(4, 21) = 2.41$.

Two factors significantly impacted attractiveness. First, appearance significantly impacted user ratings ($\beta = 0.12$, $p = 0.03$). Second, the expected healthiness from the images also demonstrated a

²Contrastive Language-Image Pre-training (CLIP).

Table 3

Linear regression models predicting user rating for recipe image attractiveness: (A): with user factors, (B): with food image dimensions. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

(A)		(B)	
User Factors		Food Image Dimension	
		β (S.E)	
User Demographic			
Age	−0.047 (0.116)	Appearance	0.129 (0.061)*
Education	−0.424 (0.320)	Healthiness	0.077 (0.035)*
Gender	−0.077 (0.088)	Taste	−0.005 (0.050)
User Profile		Familiarity	0.0231 (0.038)
Recipe Website Usage	0.201 (0.086)*	Constant	3.487 (0.365)***
Home Cooking	−0.009 (0.078)	R ²	0.011***
Cooking Experience	−0.052 (0.079)	RMSE	1.855
Eating Goals	0.019 (0.063)		
User Knowledge			
Subjective Food Knowledge	−0.213 (0.138)		
Cooking Skills	0.315 (0.086)***		
Constant	4.001 (0.570)***		
R ²	0.015***		
RMSE	1.845		

significant impact ($\beta = 0.07$, $p = 0.03$). However, perceived taste and familiarity did not impact user ratings.

To gain insights into the reasons behind the visual attractiveness ratings, we collected qualitative justifications from users. Natural Language Processing (NLP) techniques, such as punctuation removal, repeated character elimination, and stop word filtering, were applied to analyze 2,019 user justifications for both attractive and unattractive images. From these responses, we generated two word clouds to highlight the most prominent terms.

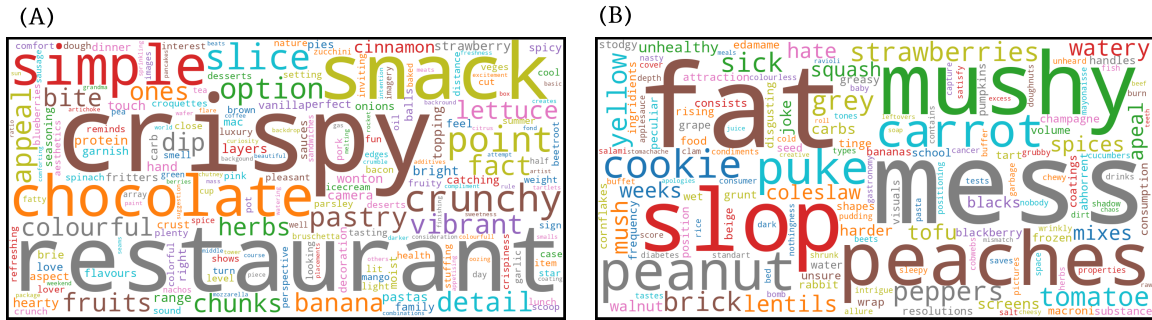


Figure 2: Word cloud for terms in the user judgment: (A) : judgments for attractive images, (B): judgments for unattractive images.

Figure 2 presents the most frequent responses associated with both attractive and unattractive images. These findings are discussed in relation to the themes of ‘appearance’ and ‘health’ (cf. Table (3.B)).

3.3.1. Appearance-based justifications

Figure 4 shows a few examples of user textual justifications. Several participants, including user (U_a), expressed the term ‘crispy’ in their assessments of attractive images, mainly referring to appearance. The word ‘simple’ is frequently used by users, such as user (U_b), to convey the simplicity of recipe

content. In contrast, ‘mess’ was more commonly associated with judgments of unattractive food images, indicating their unappealing appearance. Moreover, the repeated use of the term ‘fat’ suggested that fatty foods were generally perceived as unattractive, as in judgments by users ($U_c - d$).



(U_a): “looks juicy with nice crispy bits, which is nice and clear in the picture”



(U_b): “Interesting, slightly unusual, and does look visually appealing with simple ingredients presented well”



(U_c): “It looks messy and unappealing”



(U_d): “Too much carbs/fat”

Figure 3: Examples of images used in the study, associated with users’ textual judgments related to the appearance. (U_a) and (U_b) are textual justifications for attractive images, while (U_c) and (U_d) are justifications for unattractive images.

3.3.2. Healthiness-based justifications

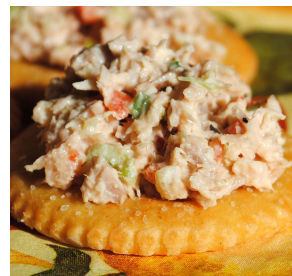
Judgments related to health frequently appeared in connection with the food’s appearance, such as by user (U_e) in Figure ?? . The term ‘restaurant’ was employed in various user judgments, often associated with presentation and healthiness, as described by the user (U_f). Conversely, the concept of unhealthiness was linked to fatty foods and messy representation, as evident in the judgments of users (U_{g-h}) in Figure ??.



(U_a): “Healthy salad option with balanced nutrients. It’s also quite colorful”



(U_b): “The dish looks very nice, like in a restaurant. It is colorful and looks very healthy”



(U_c): “It looks a bit mushy and brown and I don’t like Turkey”



(U_d): “Chicken is unhealthy and gross”

Figure 4: Examples of images used in the study, associated with users’ textual judgments related to the appearance. (U_a) and (U_b) are textual justifications for attractive images, while (U_c) and (U_d) are justifications for unattractive images.

4. Conclusion & Future work

This work has explored different aspects of the relationship between the user and food images. Through an online user study, we have found that various visual features can predict the attractiveness of a given image (i.e. colorfulness, brightness, naturalness). This prediction accuracy could be slightly improved using image features extracted using deep learning techniques (RQ1). In line with earlier work

[11, 3, 18], this suggests that the visual attractiveness of food images can be enhanced by increasing their colorfulness, brightness, and naturalness, while decreasing other features, such as saturating and sharpness. Obviously, there may be tradeoffs between these features when altering them.

Regarding user characteristics, none of the user demographics are related to food image attractiveness. In contrast, using online recipe websites and cooking skills are positively associated with the attractiveness of food images (RQ2). More novel is our contribution on the user justifications, for which we have found image appearance and perceived healthiness to be important dimensions of visual attractiveness ratings (RQ3). It seems that attractiveness are related to the expect taste or hedonic food goals (e.g., 'crispy'), while unattractive images focused on poor presentation and disliked ingredients.

Our study offers valuable insights into techniques for image attractiveness selection for various goals and domains. In particular, these techniques can be leveraged to persuade or nudge users towards specific eating goals, such as health [3, 25]. We believe that leveraging the visual appeal of attractive images can address this issue. Our future studies will focus on designing image selection pipelines for the application of food recommender systems tailored to guide people toward healthy food choices without compromising the benefits of personalization. We aim to analyze and categorize the collected textual judgment through thematic analysis to build word dictionaries related to image dimensions. These dictionaries can then be used to train learning models, enabling the evaluation of food image attractiveness based on user textual inputs.

Acknowledgments

This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the centers for Research-based Innovation scheme, project number 309339.

The authors acknowledge the use of ChatGPT [26] for checking and correcting the grammar of this article. No new content was generated this way; only existing text was checked and, if needed, corrected.

References

- [1] I. Vermeir, G. Roose, Visual design cues impacting food choice: A review and future research agenda, *Foods* 9 (2020) 1495.
- [2] C. Spence, K. Motoki, O. Petit, Factors influencing the visual deliciousness/eye-appeal of food, *Food Quality and Preference* 102 (2022) 104672.
- [3] A. D. Starke, M. C. Willemsen, C. Trattner, Nudging healthy choices in food search through visual attractiveness, *Frontiers in Artificial Intelligence* 4 (2021) 20.
- [4] R. Cadario, P. Chandon, Which healthy eating nudges work best? a meta-analysis of field experiments, *Marketing Science* 39 (2020) 465–486.
- [5] D. Elswailer, H. Hauptmann, C. Trattner, *Food Recommender Systems*, Springer US, New York, NY, 2022, pp. 871–925. URL: https://doi.org/10.1007/978-1-0716-2197-4_23. doi:10.1007/978-1-0716-2197-4_23.
- [6] D. Elswailer, C. Trattner, M. Harvey, Exploiting food choice biases for healthier recipe recommendation, in: *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 2017, pp. 575–584.
- [7] C. Trattner, D. Moesslang, D. Elswailer, On the predictability of the popularity of online recipes, *EPJ Data Science* 7 (2018) 1–39.
- [8] Q. Zhang, D. Elswailer, C. Trattner, Visual cultural biases in food classification, *Foods* 9 (2020) 823.
- [9] B. Scheibehenne, L. Miesler, P. M. Todd, Fast and frugal food choices: Uncovering individual decision heuristics, *Appetite* 49 (2007) 578–589.

- [10] A. Khosla, A. Das Sarma, R. Hamid, What makes an image popular?, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 867–876.
- [11] J. San Pedro, S. Siersdorfer, Ranking and classifying attractiveness of photos in folksonomies, in: Proceedings of the 18th international conference on World wide web, 2009, pp. 771–780.
- [12] C. Musto, C. Trattner, A. Starke, G. Semeraro, Towards a knowledge-aware food recommender system exploiting holistic user models, in: Proceedings of the 28th ACM conference on user modeling, adaptation and personalization, ACM, New York, NY, USA, 2020, pp. 333–337.
- [13] A. D. Starke, C. Musto, A. Rapp, G. Semeraro, C. Trattner, “tell me why”: using natural language justifications in a recipe recommender system to support healthier food choices, *User Modeling and User-Adapted Interaction* (2023) 1–34.
- [14] A. El Majjodi, A. D. Starke, M. Elahi, C. Trattner, et al., The interplay between food knowledge, nudges, and preference elicitation methods determines the evaluation of a recipe recommender system, in: Proceedings of the 10th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2023), 2023, pp. 1–18.
- [15] L. R. Flynn, R. E. Goldsmith, A short, reliable measure of subjective knowledge, *Journal of business research* 46 (1999) 57–66.
- [16] Z. Pieniak, J. Aertsens, W. Verbeke, Subjective and objective knowledge as determinants of organic vegetables consumption, *Food quality and preference* 21 (2010) 581–588.
- [17] N. Frans, Development of cooking skills questionnaire for EFNEP participants in Kansas, Ph.D. thesis, Kansas State University, 2017.
- [18] Q. Zhang, D. Elswailer, C. Trattner, Understanding and predicting cross-cultural food preferences with online recipe images, *Information Processing & Management* 60 (2023) 103443.
- [19] A. El Majjodi, S. A. Khan, A. D. Starke, M. Elahi, C. Trattner, Examining the visual attractiveness of digital recipe images: Material, 2024. URL: <https://github.com/ayoubGL/Health-RecSys-2024>.
- [20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
- [21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [22] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need, in: *Neural Information Processing Systems*, 2017. URL: <https://api.semanticscholar.org/CorpusID:13756489>.
- [23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.
- [24] J.-j. Chen, C.-W. Ngo, T.-S. Chua, Cross-modal recipe retrieval with rich food attributes, in: *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1771–1779.
- [25] L. Yang, C.-K. Hsieh, H. Yang, J. P. Pollak, N. Dell, S. Belongie, C. Cole, D. Estrin, Yum-me: a personalized nutrient-based meal recommender system, *ACM Transactions on Information Systems (TOIS)* 36 (2017) 1–31.
- [26] OpenAI, Chatgpt, 2024. URL: <https://openai.com/chatgpt/>, accessed: 2024-09-16.

Position Paper: Towards Recommender System Supported Contact Tracing for Cost-Efficient and Risk Aware Infection Suppression*

Vladimir Marbukh^{1,*}

¹ National Institute of Standards & Technology, Information Technology Laboratory, 100 Bureau Dr., Gaithersburg, Maryland, USA

Abstract

In public health, contact tracing is the process of identifying people who may have been exposed to an infected person. Contact tracing performance criteria, which include infection suppression, protection of high-risk individuals, and cost-efficiency, are not necessarily aligned with each other. Pareto optimization of the corresponding inherent trade-offs, especially at the early stages of infection, is typically unrealistic due to insufficient information on infection propagation, risk factors, prevention and treatment options, etc. We suggest that contact tracing performance can be significantly improved with the support of a specialized Recommender System (RS). Based on the combination of up-to-date contact tracing and medical data, RS can identify and test through Exposure Notification System (ENS) not only high-risk individuals but also potential superspreaders to suppress infection propagation. Due to incomplete information, the dynamic nature of the problem, and a large state and action spaces, the RS should be supported by Deep Reinforcement Learning (DRL) for solving the corresponding Partially Observable Markov Decision Process (POMDP).

Keywords

contact tracing, exposure notifications, recommender system, deep reinforcement learning, partially observable Markov decision process

1. Introduction & Motivation

In public health, contact tracing is the process of identifying people who may have been exposed to an infected person, subsequent testing them for infection, and isolating or treating the infected [1]. Contact tracing performance criteria include infection suppression, protection of high-risk individuals, and cost-efficiency. These criteria are not necessarily aligned with each other, e.g., given testing capacity, infection suppression requires high priority testing for the potential super spreaders, while protection of high-risk individuals requires testing them with higher priority. Given the testing priorities, the existing Google/Apple Exposure Notification (GAEN) technology [2] can support an Exposure Notification System (ENS) by allowing public health authorities to quickly notify people for subsequent testing. GAEN is a framework and protocol specification developed by Apple Inc. and Google to facilitate digital contact tracing during the COVID-19 pandemic to augment more traditional contact tracing techniques using Android or iOS smartphones.

Extensive research on COVID-19 has revealed that while risk-aware, multi-criteria optimization of contact tracing has significant potential, realization of this potential requires deep knowledge of the infection propagation mechanisms, medical prognoses and treatment options for infected individuals with different risk profiles [3]. Even though COVID-19 originated almost five years ago, such knowledge is still lacking [4, 5], which suggests that a contact tracing system should have the ability to collect and make sense of all up-to-date available information on infection. This can be achieved with the

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

✉ marbukh@nist.gov (V. Marbukh)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

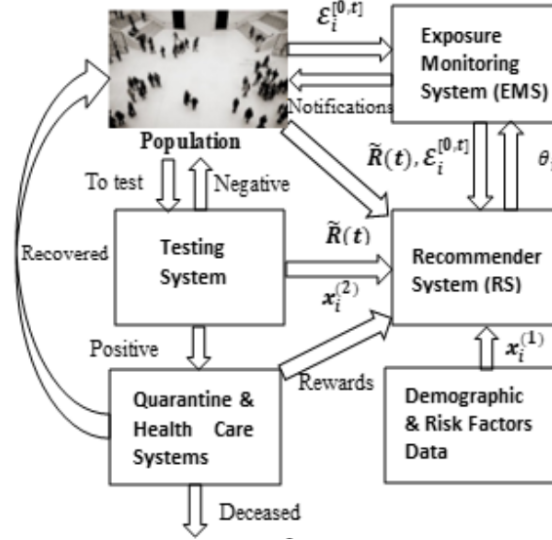


Figure 1: Recommender System supported Contact Tracing.

support of a specialized Recommender System (RS). Given testing capacity, the RS should utilize the up-to-date contact tracing, medical, and all other available relevant data to identify and through Exposure Notification System (ENS) notify individuals to be tested [2, 6]. Due to incomplete information, the dynamic nature of the problem, and a large state and action space the RS should be supported by Deep Reinforcement Learning (DRL) for solving the corresponding Partially Observable Markov Decision Process (POMDP) [7, 8]. POMDP describes the evolution of health status of each participating individual, where infectious status may not be observable and testing decisions are constrained by available testing capacity. Since a positive test result for some individual may reveal increased accumulated exposure for other individuals due to proximity to the newly discovered infection spreaders, the problem cannot be decoupled. These interdependencies significantly complicate the problem. The paper is organized as follows. Section 2 outlines operations and flow of information in the proposed RS supported contact tracing, and section 3 provides some technical details on accumulated exposure evaluation.

2. RS Supported Contact Tracing

Figure 2 presents a highly aggregated view of a Recommender System supported Contact Tracing System.

The Exposure Monitoring System (EMS) monitors “accumulated exposure to infection” for each participating individual i , $\mathcal{E}_i^{[0,t]}$ (defined in the next section) in near real time t , and feeds this information into the RS. RS also gets available information on demographic and risk factors of participating individuals $x_i^{(1)}$ as well as health status of both participating and not participating individuals who went through Health Care System $x_i^{(2)}$. Note that participating individuals are likely to consent to revealing their health information since they would benefit from accounting for their risk factors, e.g., advanced age, preexisting conditions, etc. For not participating individuals, some relevant information, which does not require revealing individual identity, can be obtained without violating their privacy.

RS is also fed the estimate of the infection reproduction number $R(t)$, i.e., the average number of new infections produced by one infected individual during his/her lifetime: $\tilde{R}(t) \approx R(t)$. Estimate $\tilde{R}(t)$ may combine information from EMS, the Health Care System, and possibly from other tracing mechanisms not shown in Figure 1, e.g., from manual tracing. Infection suppression requires keeping the infection reproduction number less than one: $R(t) < 1$. Due to numerous uncertainties in the $R(t)$ estimation: $\tilde{R}(t) \approx R(t)$, the infection suppression condition is $\tilde{R}(t) \leq 1 - \varepsilon$, where “safety margin” $\varepsilon < 1$ depends on the confidence level of the corresponding estimate. The reward of the RS supported Contact Tracing

is quantified by the negative loss $-L(t)$, where $L(t) = L_{ec}(t) + L_{sc}(t)$. Here economic loss due to lost productivity and cost of testing/treatment is $L_{ec}(t)$, and “social cost” quantifying suffering and, most importantly, deaths due to the infection $L_{sc}(t)$. The cost estimates are provided to RS by the Health Care System and Agencies collecting and processing economic data.

System evolution is described by POMDP $\delta(t) = (\delta_i(t))$, where component $\delta_i(t)$ describes the health status of participating individual i , i.e., “non-infected,” “infected,” “deceased.” “Non-infected” and “infected” states may not be observable which makes process $\delta(t)$ partially observable. The decision to test a participating individual reveals his/her infected or not-infected status at a certain cost due to limited testing capacity. RL employs DRL to make testing decisions on the basis of $\mathcal{E}_i^{[0,t]}$, $x_i^{(1)}$, $x_i^{(2)}$. Constraints on the infection reproduction number can be incorporated through penalty function $h(\tilde{R}(t))$ which is flat for $\tilde{R}(t) \leq 1 - \varepsilon$ and sharply increases for $\tilde{R}(t) > 1 - \varepsilon$.

Our conjecture is that (near) optimal notification strategy is threshold-based: individual i should be notified at the first moment $t = \theta_i > 0$ when this individual’s accumulated exposure reaches threshold $\hat{\mathcal{E}}_i$:

$$\theta_i = \inf_{t \geq 0} \{t : \mathcal{E}_i^{[0,t]} \geq \hat{\mathcal{E}}_i\}, \quad (1)$$

where threshold $\hat{\mathcal{E}}_i = \Delta(\mathcal{E}, x)$ depends on the history of former testing decisions/results combined with medical and demographic data of these individuals. The function $\Delta(\mathcal{E}, x)$ can be evaluated by employing a Deep Supervised Learning (DSL) algorithm. Note that in practice, notification strategy may operate on the basis of a small number of risk groups [3], which may be defined and then redefined by an on-line clustering algorithm. Assumptions of homogeneity and large number of individuals within each group, simplifies optimization of group-specific thresholds in (1).

3. Accumulated Exposure

For each participating individual i , the contact tracing system identifies “accumulated exposure” to another participating individual j during time interval $[0, t]$ as follows:

$$\mathcal{E}_{ij}^{[0,t]} = \int_0^t \phi[\tilde{d}/d_{ij}(\tau)]^\alpha d\tau, \quad (2)$$

where the corresponding instantaneous exposure rate $\phi(z)$ is an increasing function of $z > 0$, the distance between individuals i and j at moment τ is $d_{ij}(\tau)$, and $\tilde{d} > 0$, $\alpha \geq 1$ are some parameters. Individual i accumulated exposure to infection during time interval $[0, t]$ is defined as the aggregated exposure to all known spreaders during this time interval:

$$\mathcal{E}_i^{[0,t]} = \sum_j \int_0^t \pi_j(\tau) \phi[\tilde{d}/d_{ij}(\tau)]^\alpha d\tau, \quad (3)$$

where $\pi_j(\tau) = 1$ if individual j is infected at moment τ and $\pi_j(\tau) = 0$ otherwise.

Consider some examples. As currently defined by the CDC [1], a high-risk COVID-19 exposure is a contact with a person who tests/tested positive for SARS-CoV-2 which takes place at a distance of less than two meters for a total of 15 minutes or more over a 24-hour period. In this case, $\tilde{d} = 2$ m, $\alpha \rightarrow \infty$, $\phi(x) \equiv \min(x, 1)$, and thus an individual is assumed exposed if $\mathcal{E}^{[0,T]} = \int_0^{24} \mathbb{1}(d(t) - 2 \text{ m}) dt > 15 \text{ min}$, where $\mathbb{1}(x) = 0$ if $x \leq 0$ and $\mathbb{1}(x) = 1$ otherwise. In another example [9], $\tilde{d} = 2$ m, $\phi(x) \equiv x$, and thus an individual is assumed exposed if $\mathcal{E}^{[0,T]} = \int_0^{24} (2/d(t))^\alpha dt > 15 \text{ min}$.

Finally note that available information on accumulated exposure to specific individuals can be used to identify “infection superspreaders” who otherwise could be unidentified, e.g., due to being asymptomatic or for any other reason. This can be done with known algorithms [10] on undirected exposure graph G where nodes i and j are connected if $\mathcal{E}_{ij}^{[0,t]} \geq \tilde{\mathcal{E}}^{[0,t]}$, and $\tilde{\mathcal{E}}^{[0,t]} > 0$ is a properly defined threshold.

References

- [1] CDC, Contact tracing, 2024. URL: <https://www.cdc.gov/museum/pdf/cdm-pha-stem-lesson-contact-tracing-lesson.pdf>.
- [2] C. Nebeker, D. Kareem, A. Yong, R. Kunowski, M. Malekinejad, E. Aronoff-Spencer, Digital exposure notification tools: A global landscape analysis, *PLOS Digital Health* 2 (2023) 1–16. doi:10.1371/journal.pdig.0000287.
- [3] D. Acemoglu, V. Chernozhukov, I. Werning, M. D. Whinston, Optimal Targeted Lockdowns in a Multigroup SIR Model, *American Economic Review: Insights* 3 (2021) 487–502. doi:10.1257/aeri.20200590.
- [4] C. F. Manski, F. Molinari, Estimating the COVID-19 infection rate: Anatomy of an inference problem, *Journal of Econometrics* 220 (2021) 181–192. doi:10.1016/j.jeconom.2020.04.041.
- [5] W. Rauch, H. Schenk, N. Rauch, M. Harders, H. Oberacher, H. Insam, R. Markt, N. Kreuzinger, Estimating actual SARS-CoV-2 infections from secondary data, *Scientific Reports* 14 (2024) 6732. doi:10.1038/s41598-024-57238-0.
- [6] X. Chen, L. Yao, J. McAuley, G. Zhou, X. Wang, Deep reinforcement learning in recommender systems: A survey and new perspectives, *Knowledge-Based Systems* 264 (2023) 110335. doi:10.1016/j.knosys.2023.110335.
- [7] X. Xiang, S. Foo, Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (POMDP) problems: Part 1—fundamentals and applications in games, robotics and natural language processing, *Machine Learning and Knowledge Extraction* 3 (2021) 554–581. doi:10.3390/make3030029.
- [8] X. Xiang, S. Foo, H. Zang, Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (POMDP) problems part 2—applications in transportation, industries, communications and networking and more topics, *Machine Learning and Knowledge Extraction* 3 (2021) 863–878. doi:10.3390/make3040043.
- [9] K. Sayrafian, B. Cloteaux, V. Marbukh, Impact of using soft exposure thresholds in automatic contact tracing, in: *2022 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, 2022, pp. 1–6. doi:10.1109/HealthCom54947.2022.9982790.
- [10] D. Zhang, Y. Wang, Z. Zhang, Identifying and quantifying potential super-spreaders in social networks, *Scientific Reports* 9 (2019) 14811. doi:10.1038/s41598-019-51153-5.

Enriching Clinical Sample Analysis with Biological Knowledge Graphs: A Preliminary Study

Fatemeh Shad Bakhsh¹, Juan Manuel Rodriguez¹, Alessandro Ranieri²,
Kenneth Kastaniegaard² and Daniele Dell’Aglio¹

¹Department of Computer Science, Aalborg University, Aalborg, Denmark

²Biogenity, Aalborg, Denmark

Abstract

Biological researchers often face challenges in analyzing clinical samples due to the limited amount of samples they can collect. This issue hinders the use of traditional statistical methods; instead, they often rely on their domain knowledge to guide the exploration of the data. To ease the task, we aim to develop a system to support the researchers by integrating data from biological knowledge graphs (KGs), such as Reactome and UniProt, which can drive data exploration through recommendation system techniques. In this article, we present the first step towards such a system by studying whether the data from biological KGs can be represented through embeddings so that proteins of interest are organized and categorized according to their shared functionalities. We propose Cluster-GAE, a graph autoencoder method inspired by Cluster-GCN. Cluster-GAE combines graph sampling techniques with Graph Neural Networks (GNNs) to learn embedding representations from large-scale biological networks. Our experiments suggest that Cluster-GAE produces embeddings with properties similar to the embeddings of traditional KG embedding methods without the need to process the whole graph at once. Moreover, the experiments show that the embeddings produced by Cluster-GAE are of a higher quality than the embeddings produced by a KG obtained through traditional sampling techniques, in particular Random Walk and Forest Fire. Finally, through t-SNE visualization and functional enrichment analysis, we showcase the ability of Cluster-GAE to identify protein clusters that are related to different biological processes, molecular functions, and cellular components.

Keywords

Knowledge graphs, Graph Neural Networks, Pathway analysis

1. Introduction

Biological researchers collect clinical samples to study living organisms, quantifying the amounts of thousands of proteins of interest and looking for relations among them. However, collecting samples is expensive and time-consuming. Therefore, trials often include only a limited number of samples [1]. Moreover, the number of features tends to be much greater than the number of samples, and considering small perturbation, finding the features related to a particular phenomenon is a complex task. Currently, state-of-the-art methods, such as Stabl [2], rely on multiple sampling and data augmentation to find relevant features.

Biological databases, such as Reactome [3], KEGG [4], UniProt [5], are invaluable tools for enhancing the analysis of biological data, as they act as encyclopedic knowledge about proteins, reactions, and biological pathways. These databases are often structured through graph-based data structures or knowledge graphs. Hence, to ease the analysis of datasets with a few data points (the samples) and many features (the proteins), we argue for leveraging this information to create an item-item recommender system for discovering relevant features leveraging the information encompassed by Reactome or UniProt. In particular, we propose to rely on graph representation learning to capture the relevant

HealthRecSys’24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

✉ fbakhs22@student.aau.dk (F. S. Bakhsh); jmro@cs.aau.dk (J. M. Rodriguez); ara@biogenity.com (A. Ranieri); kkas@biogenity.com (K. Kastaniegaard); dade@cs.aau.dk (D. Dell’Aglio)

ORCID: 0000-0002-1130-8065 (J. M. Rodriguez); 0000-0003-3806-5638 (K. Kastaniegaard); 0000-0003-4904-2511 (D. Dell’Aglio)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

structural and semantic information from the graph by using Graph Auto-Encoders (GAE) [6] to extract node embeddings to discover relevant features within the original dataset.

In this context, this work presents an exploratory analysis of different graph encoder capabilities to capture relevant information on the node embeddings. Specifically, we analyze the effectiveness of GAE and graph sampling techniques to leverage biological knowledge graphs, such as **Reactome** and **UniProt**, to create embeddings. Such embeddings can later be used to discover protein-protein associations in the context of analysis of limited numbers of clinical samples. We propose *Cluster-GAE*, a method that combines GAE and Cluster-GCN [7], a technique for applying Graph Convolutional Network (GCN) to large graphs. We experimentally study the embeddings generated by Cluster-GAE and other GAE solutions. We observe that Cluster-GAE organizes protein embeddings so that they can be clustered according to their biological features. This is a promising result towards the creation of a recommender system to support biological researchers.

In the following, we first introduce the background and related work in Section 2. Next, we illustrate the overall solution we envision in Section 3, and we describe the part we implemented so far, which is later experimentally analysed in Section 4. We conclude with some remarks and the next steps in Section 5.

2. Background and related research

Let P be a set of proteins of interest and y a phenomenon, e.g. being sick. Let S be a clinical sample data set. Each clinical sample can be represented as a vector of size $|P| + 1$, which contains the amount of proteins of interest and the presence or absence of the phenomenon. Therefore, S can be represented as a matrix of size $|P| + 1 \times n$, where n is the number of clinical samples.

Biologist researchers are interested in discovering relations between groups of proteins (i.e. subsets of P) and the phenomenon through the analysis of S . The main challenge for the analysis is the number of proteins $|P|$, which is much larger than the number of clinical samples n . This is because the collection of clinical samples is costly and time-consuming, so trials often involve a number of subjects [8].

This combination of few data instances and numerous features presents substantial challenges for conventional statistical analysis [9, 2]. Even well-known sparsity-promoting methods, such as LASSO [10], are hindered as small errors in the samplings might result in completely different results.

A natural way to cope with the problem is to adopt feature selection techniques like Stabl [2]. Stabl adds noise by generating artificial features and repeatedly subsamples the dataset, looking for features that are frequently selected in the different runs. However, Stabl is over-conservative as it tries to minimize the false discovery ratio that depends on the quality of the generated features and subsampling of the existing data.

Alternatively, one can consider exploiting pre-existing knowledge about the proteins P , coming from literature, to drive the analysis of S . Such knowledge is, for example, stored in biological knowledge graphs (KGs) such as UniProt [5] or Reactome [3]. They store information about proteins, genes, genomes, and pathways, along with their relations and functions. Moreover, such KGs are usually well curated and connected to each other through links between the same entity in different KGs. As a result, there is a massive amount of data that can be useful to drive the analysis process of biological researchers.

While the traditional manner of exploiting these knowledge graphs is to manually query the proteins and navigate through the relations [11], representation learning emerged in the recent years as a means to feed KG data to machine learning algorithms, to solve tasks like classification and recommendation. For example, Burkhart et al. [12] propose to enrich biological prediction models using a Graph Convolutional Network (GCN) to enrich the samples in the dataset. As a result, the data is not processed as tabular data, but information about the relations of the different features is added to the predictive model. Although this model increments the information in the model, it is not a feature reduction technique, and the results are hard to interpret.

Pershad et al. [13] show that the use of node embeddings can be more effective than traditional drug

recommendation methods. Unlike Burkhardt et al. [12], the base graph is not Reactome but a graph computed using a probabilistic algorithm over the samples that return protein-protein-interaction networks. In particular, this work presents evidence that the Node2Vec algorithm can derive meaningful embeddings using these PIP networks.

However, biological researchers are rarely equally interested in all proteins. They have a number of proteins they are interested in, and they aim to discover what are the connected proteins, and how they vary w.r.t. each other. Therefore, we argue that this can be modeled as an item-item recommendation task, where given one or more proteins, the goal is to retrieve a list of related proteins which can be later analyzed by the researchers.

3. A recommender system for clinical sample analysis

Figure 1 shows a high-level schema of the solution we envision to support researchers in analyzing clinical samples. There are three main steps, represented by the dark boxes. The first is the sample-based *selector*, which takes the clinical samples S and a biological KG as inputs, and outputs a KG with information specific to S . The second is the embedding *extractor*, which learns embeddings from the KG produced by the selector. Finally, the protein *recommender* uses the embeddings to compute the list of recommended proteins given the input ones. In the following, we provide information on the current implementation of the components.

The sample-based selector

Instead of working on the whole graph, we filter a subgraph with the proteins of interest. To do it, we query Reactome to extract the part of the graph that contains the protein/features present in the clinical samples to analyze. Specifically, for each protein in P , we extract the nodes of the features and nodes up to two hops of distance.

This is implemented as Cypher queries, which are evaluated over Reactome using Neo4J. Listing 1 shows an example query, which is parametrized. At Line 1, the query reads the list of protein names (Line 1). At Lines 2 and 3, the query retrieves the Reacome nodes associated with the input proteins. The block at Lines 4–7 retrieves the neighbor nodes (at a maximum distance of two) of the input protein nodes. The block at Lines 10–11 adds constraints on the nodes to be retrieved, such as being related to mice (*mus musculus*), a possible type of organism studied by the clinical samples. Finally, the operation at Line 12 returns the edges.

After the construction of the out graph, we verified its connectivity, to avoid having isolated proteins which could not be processed in the next steps.

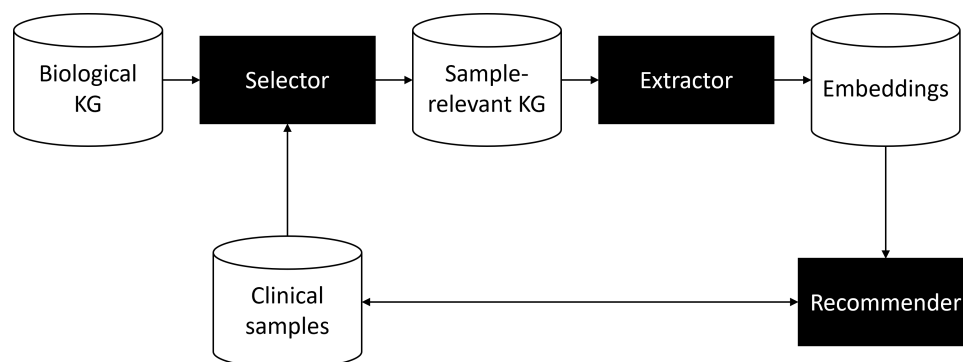


Figure 1: The recommender system architecture

Listing 1: The query executed by the sample-based selector

```

1 UNWIND $proteinNames AS proteinName
2 MATCH (p)
3 WHERE ("EntityWithAccessionedSequence" IN labels(p) OR
4        "GenomeEncodedEntity" IN labels(p)) AND
5        ANY(name IN p.name WHERE name = proteinName)
6 CALL apoc.path.subgraphAll(p, {
7     maxLevel: 2,
8     minLevel: 1
9 })
10 YIELD nodes, relationships
11 WITH p,
12     [node in nodes WHERE node.speciesName = $specie] AS filteredNodes,
13     relationships
14 RETURN p AS protein, filteredNodes AS nodes, relationships

```

The embedding extractor

Despite the graph induced by the clinical samples is smaller than the original biological knowledge graph, its size is still large. Therefore, the learning algorithm that extracts the embeddings needs to exploit an adequate strategy to cope with the graph volume. We considered two alternatives: the GAE extractor and the Cluster-GAE extractor. The former uses graph sampling to reduce the size of the knowledge graph and generates the embeddings using a GAE. The latter does not perform the graph sampling. Instead, it uses the approach initially proposed by Cluster-GCN based on a divide-and-conquer strategy.

The *GAE extractor* exploits graph sampling to reduce the size of the graph. The goal is to improve computation performance while retaining the essential structural properties of the original graph. As sampling algorithms, we consider Random Walk (RW) with Restart [14] and Forest Fire (FF) [14]. We set as a constraint the necessity of having a connected graph as an output of the process.

After that, a GAE model learns low-dimensional, yet informative representations of the graph's nodes. The GAE model comprises an encoder and a decoder, with the encoder being a GCN that embeds nodes into a latent space. The decoder then attempts to reconstruct the graph's adjacency matrix from these embeddings. This process is facilitated by a reconstruction loss function that guides the model to capture the essential topological similarities based on features between nodes [6].

The *Cluster-GAE extractor* divides the data into partitions and then samples from these partitions during the GAE training phase. This extractor omits the sampling step described above. Instead, it adapts the Cluster-GCN algorithm [7], which enhances computational efficiency by dividing the graph into multiple clusters. We replace GCN with a GAE, as the former solves a node classification task, while we aim at learning embeddings to build a recommendation task. The GAE was built using the architecture and loss function and task as defined in [6].

This modification allows us to generate meaningful embeddings and understand the underlying structure of the graph. Basically, this method divides the data into partitions and then samples from these partitions during the training phase.

The protein recommender

The last step consists of using the embeddings to group features using clustering or detecting similar proteins/features using embedding distance, recommending how to group the features to analyze to the researcher.

Table 1

Comparison of Clustering Metrics for Different Embedding Dimensions and Sampling Methods for K-means with $K = 2$ and $K = 3$. Metrics include Silhouette Score (Sil), Calinski-Harabasz Index (CH), and Davies-Bouldin Index (DB), evaluated for embedding dimensions of 64 and 128.

Sampling Method	K-Means $K = 2$						K-Means $K = 3$					
	d = 64			d = 128			d = 64			d = 128		
	Sil	CH	DB	Sil	CH	DB	Sil	CH	DB	Sil	CH	DB
RW	0.672	791.043	0.866	0.417	1130.384	1.083	0.502	1141.947	0.777	0.420	904.776	0.943
FF	0.432	1068.283	0.972	0.460	1261.628	0.969	0.513	1433.257	0.707	0.497	1129.500	0.805
Cluster-GAE	0.551	2138.868	0.736	0.534	2026.599	0.576	0.775	1794.486	0.770	0.556	1615.266	0.761
No Sampling	0.369	784.169	1.281	0.226	394.942	1.922	0.438	864.572	0.918	0.270	454.419	1.493

4. Preliminary Analysis of the Solution

This section presents our preliminary analysis of the system, specifically on the embedding extractor component. We analyze the behavior of the GAE and Cluster-GAE extractors by analyzing the quality of the generated embeddings through cluster analysis. The code for the experiments is available at <https://github.com/dkw-aau/clinical-sample-enrichment>.

4.1. Experimental setting

Dataset

We use a dataset consisting of clinical samples from mice characterized by 3 825 proteins, one boolean variable associated with the presence of the phenomenon of study, and 48 data instances.

Using the features of the dataset, we query Ractome with a query similar to the one in Listing 1. The query builds a graph with 50, 164 nodes and 1, 667, 138 edges. Compared to Reactome, which includes 2, 427, 555 nodes and 10, 102, 445 edges, we observe that the generated graph is denser, as it preserves 2.06% of the nodes and 16.5% of the edges.

Evaluation Metrics

To evaluate the extractors, we consider two metrics.

Firstly, we use the *Earth Mover’s Distance (EMD)*. EMD evaluates the dissimilarity between two probability distributions, offering a quantitative assessment of the differences in protein embedding matrices across different sampling methods [15].

Secondly, we use *clustering metrics*. These metrics evaluate the quality of clustering results by assessing how well the identified clusters adhere to desirable properties like compactness, separation, and connectedness. The clustering metrics we consider are:

- **Silhouette Score:** Measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters [16].
- **Calinski-Harabasz Index:** Measures the ratio of the between-cluster dispersion to the within-cluster dispersion. A higher Calinski-Harabasz index indicates denser and more well-separated clusters [17].
- **Davies-Bouldin Index:** Measures the average similarity between each cluster and its most similar cluster. A lower Davies-Bouldin index indicates better cluster separation [18].

4.2. Effect of sampling

We study how the sampling method affects the generated embeddings. We create graph embeddings using the sampling pipeline but omit the sampling step. As a direct comparison of the embeddings is not feasible due to the randomness in the learning process, we compare the distances between embeddings of various proteins in the dataset. We calculated the cosine distance for all protein embeddings derived from

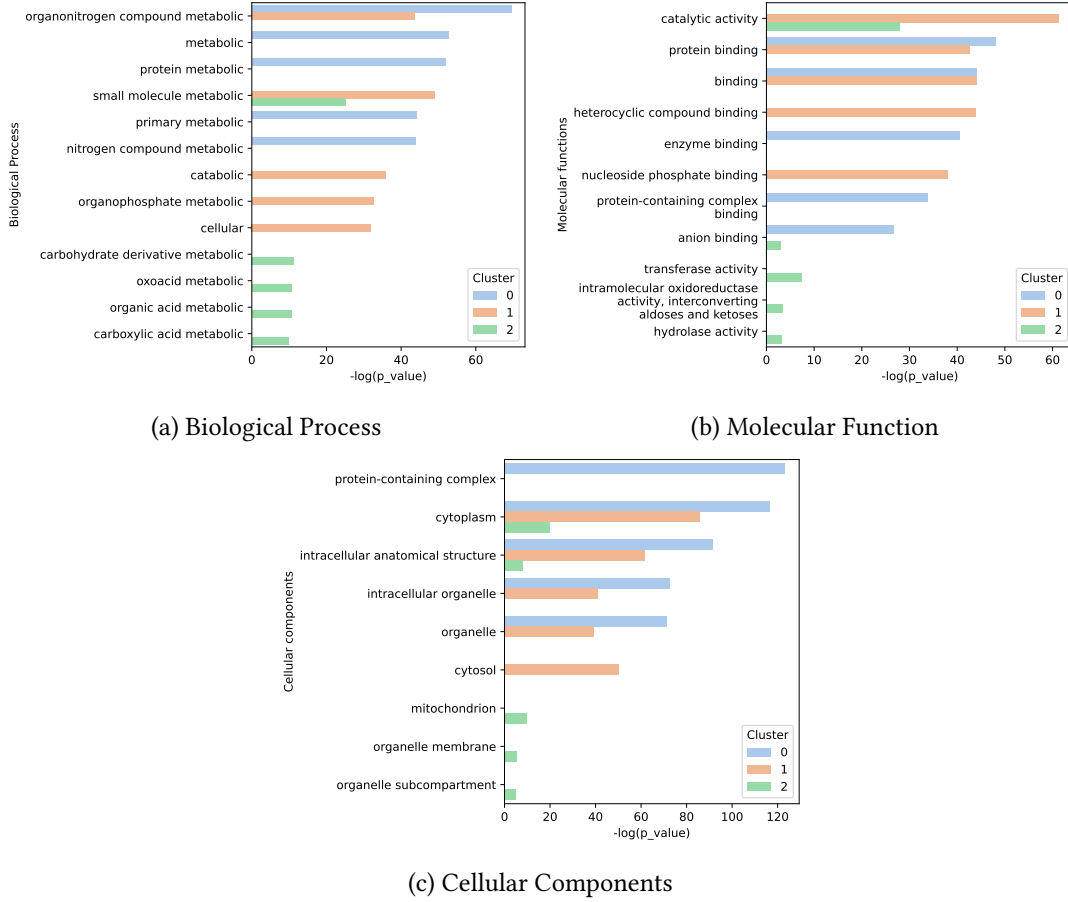


Figure 2: Top 5 over-represented pathways in the clusters after doing the functional enrichment analysis.

different extractions: the GAE extractor with RW and, FF and the Cluster-GAE extractor. Subsequently, we compute the EMD between the extractors and a baseline, named *No sampling*. This baseline consists of a GAE extractor without any sampling procedure.

Table 2

The Earth Mover’s Distances between the extractors and the baseline

Sampling Method	$d = 64$	$d = 128$
RW	0.481	0.495
FF	0.277	0.321
Cluster-GAE	0.081	0.308

Table 2 presents the EMD values for embeddings of dimensions 64 and 128. We observe that the behaviour of the Cluster-GAE embeddings is the most similar to the one of the baseline. This is evidence that preserving the input graph allows the exploitation of more information in the learning process.

Looking into the performance of the two GAE extractors, we observe that the one using FF as a sampling method produces closer embeddings to the baseline. When the embedding dimension is 128, the EMD value of the GAE extractor with FF is closer to the one of the Cluster-GAE extractor than the GAE extractor with RW. The results suggest that FF is effective in preserving the input graph structure.

4.3. Cluster analysis

While the previous analysis offers some insights into the effect of sampling, it does not provide useful information about the quality of the embeddings. Therefore, we perform a cluster analysis to obtain

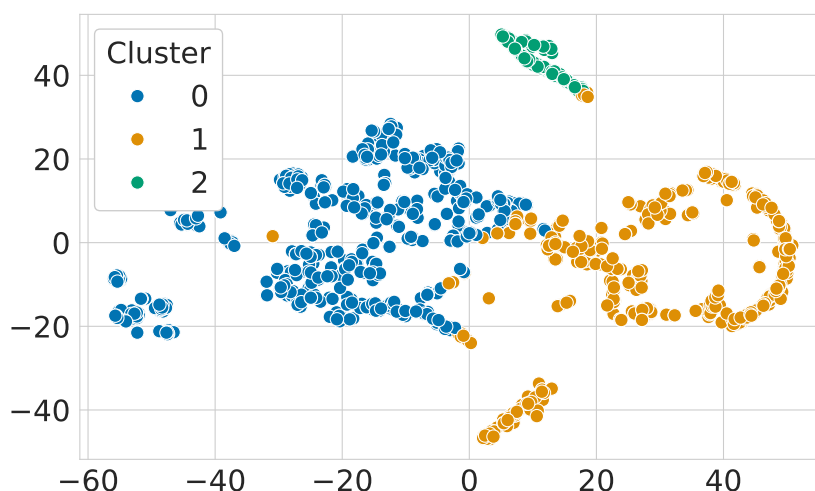


Figure 3: t-SNE visualization of protein embeddings clustered using K-means ($K = 3$) with Cluster-GAE.

qualitative and quantitative insights on the learned space.

As the clustering algorithm, we use K-means, as this algorithm is well-recognized for its effectiveness with embeddings across various domains [19, 20]. Specifically, we consider K-means with $K = 2$ and $K = 3$.

Table 1 displays the cluster metrics for the embeddings produced by the various extractors. In general, Cluster-GAE produces embeddings that are distinctly clustered and well-separated. This observation holds true regardless of the vector dimensions.

To visually investigate the space created by the Cluster-GAE extractor, we used t-SNE. Figure 3 shows the embeddings with dimensions created by Cluster-GAE and clustered through K-means when $K = 3$. The figure shows that there are two prominent clusters and a third, smaller yet more distinct cluster.

4.4. Functional enrichment analysis

We further investigate the quality of the embedding generated by Cluster-GAE through a functional enrichment analysis [21].

We use g:Profiler¹ to run such an analysis. This tool takes a set of proteins and a target organism as inputs and identifies overrepresented biological processes, molecular functions, and cellular components to which those proteins might be related. To analyze the properties of the clusters, we run the g:profile analysis using the proteins of clusters. We expect each cluster to be related to different results in a large way. This means that the quality of the clustering is reflected in the degree to which these functions are distinct between clusters.

We divided the analysis by the three top-level categories of the Gene Ontology, namely “Biological Process” (BP), “Molecular Function” (MF), and “Cellular Components” (CC). We consider $p_{value} < 0.05$ when performing the functional enrichment analysis. Figure 2 presents the top-5 results for each category divided into clusters. This Top-5 is based on $-\log_{10}(p_{value})$, i.e., these are the results with the lowest p_{value} . We expect that an effective clustering algorithm would identify proteins that are related to specific BPs, MFs, and CCs in the same group. In other words, if different clusters share few functional results, the clustering is effective in capturing the distinct BPs, MFs, and CCs; therefore providing evidence of the biological relevance of the clusters.

The results show that for BPs, there is a clear separation of the top-5 retrieved processes for each

¹

<https://biit.cs.ut.ee/gprofiler/gost>

cluster. However, there is an overlap between Clusters 0 and 1 for “organonitrogen compound metabolic” and between proteins in Clusters 1 and 2 for the “small molecule metabolic.” All in all, there is a clear difference. If we consider all the results, Cluster 0 might be involved in 682, and 67.01% of them are unique to this cluster, Cluster 1 might contribute to 462 processes, where 47.15% of them are only related to this Cluster. Cluster 2 contributes to 72 processes and uniquely contributes to 25.0% of them. Moreover, 43.06% of Cluster 2 processes are shared with both Cluster 0 and 1. In this sense, Cluster 2 mainly comprises proteins that can participate in the BPs of both Clusters.

When considering MFs, there is a more extensive overlap between Clusters 0 and 1 on the potential MFs. However, there are some unique identifiable MFs associated with each cluster. Regarding all MFs retrieved for each Cluster, Clusters 0, 1, and 2 might be associated with 122, 93, and 23 MFs, respectively. 59.02% of Cluster 1 MFs are unique to that cluster, 44.09% of the MFs of Cluster 2 are not shared, and 43.48% of the MFs of Cluster 2 did not appear in other Clusters. Therefore, for molecular MFs, we can see that clustering separates the molecules according to their functions.

Something similar happens with MFs when we consider CCs. In the top 5, we observe mostly shared results for Clusters 0 and 1, with some unique components. When analyzing the full results of the functional enrichment, we observed that Clusters 0, 1, and 2 are related to 220, 151, and 20 CCs, respectively, and 64.09% are unique to Cluster 0, 45.03% are unique to Cluster 1, and 15.0% are unique to Cluster 2. Like in the case of molecular function, Cluster 2 encompasses proteins that are related to common CCs, as 65.0% of the 20 CCs are shared by both Clusters 0 and 1.

In summary, the results suggest that the embeddings convey information about the protein’s role in BPs, MFs, and CCs. Therefore, applying k-means using these embeddings results in clusters with a large number of unique roles in the type of results of functional enrichment analysis.

5. Conclusion and Future Works

This article presented our ongoing research in building support tools for biological researchers. The solution we envision exploits biological KGs to suggest new proteins for researchers to study in the context of clinical sample analysis.

We presented some possible strategies to learn embeddings from a biological KG. The experimental analysis suggested that Cluster-GAE is effective in processing biological KGs and creating embedding spaces, when compared to embedding techniques that exploit sampling to reduce the size of the graph.

Moreover, the embeddings of the proteins learned through Cluster-GAE can effectively be clustered according to the biological processes they contribute to. The functional enrichment analysis using g:Profiler highlights that proteins grouped within the same cluster tend to be associated with similar biological pathways and processes.

One of the current limitations we face is that we used one dataset. It is necessary to repeat the analysis on different datasets, ideally associated to various organisms and with diverse biological conditions, to ensure that our findings generalize.

The next natural step is to build the recommendations on top of the learned embeddings. The first approach is to rely exclusively on the embeddings and compute recommendations from them (e.g., through nearest neighbor techniques). However, the recommendation process should also take advantage of the information coming from clinical samples. Despite limited in the size of the samples, the contained values are a valuable source of information that should contribute to the recommendation process.

The natural focus to build a recommender system is on non-personalized recommender systems. However, the presence of various researchers analyzing the same samples opens opportunities for collaborative filtering approaches. Furthermore, incorporating user-interaction data, such as previous protein selections or research interests, could further enhance the personalization and relevance of recommendations.

Currently, the major issue for creating a recommender system for protein selection in biological research is the lack of datasets to evaluate its performance. State-of-the-art methods in protein selection

in a dataset, such as Stabl [2], follow a stochastic approach. Since these methods aim to optimize the performance of predicting models by selecting a subset of proteins, and the number of proteins is much greater than the number of samples, there is a great risk of overfitting. However, the results of these methods can be used for a first validation.

With time, our goal is to create a real-life dataset by providing researchers with recommendations and storing their feedback. After collecting a dataset, we plan to extend the recommender system to consider not only the biological knowledge graph as input but also details of the research being carried out to provide personalized recommendations.

Acknowledgments

This research has been partially supported by AI Denmark.

References

- [1] P. Feist, A. B. Hummon, Proteomic challenges: Sample preparation techniques for microgram-quantity protein analysis from biological samples, *International Journal of Molecular Sciences* 16 (2015) 3537–3563. URL: <https://www.mdpi.com/1422-0067/16/2/3537>. doi:10.3390/ijms16023537.
- [2] J. Hédou, I. Marić, G. Bellan, J. Einhaus, D. K. Gaudillière, F.-X. Ladant, F. Verdonk, I. A. Stelzer, D. Feyaerts, A. S. Tsai, E. A. Ganio, M. Sabayev, J. Gillard, J. Amar, A. Cambriel, T. T. Oskotsky, A. Roldan, J. L. Golob, M. Sirota, T. A. Bonham, M. Sato, M. Diop, X. Durand, M. S. Angst, D. K. Stevenson, N. Aghaeepour, A. Montanari, B. Gaudillière, Discovery of sparse, reliable omic biomarkers with stabl, *Nature Biotechnology* (2024). doi:10.1038/s41587-023-02033-x.
- [3] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, et al., The reactome pathway knowledgebase, *Nucleic acids research* 48 (2020) D498–D503.
- [4] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, M. Tanabe, Kegg: integrating viruses and cellular organisms, *Nucleic acids research* 49 (2021) D545–D551.
- [5] T. U. Consortium, Uniprot: the universal protein knowledgebase in 2023, *Nucleic acids research* 51 (2023) D523–D531.
- [6] T. N. Kipf, M. Welling, Variational graph auto-encoders, *arXiv preprint arXiv:1611.07308* (2016).
- [7] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, C.-J. Hsieh, Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks, in: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 257–266.
- [8] I. Subramanian, S. Verma, S. Kumar, A. Jere, K. Anamika, Multi-omics data integration, interpretation, and its application, *Bioinformatics and Biology Insights* 14 (2020) 1177932219899051. doi:10.1177/1177932219899051.
- [9] E. Candès, Y. Fan, L. Janson, J. Lv, Panning for Gold: ‘Model-X’ Knockoffs for High Dimensional Controlled Variable Selection, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80 (2018) 551–577. doi:10.1111/rssb.12265.
- [10] T. H. Noah Simon, Jerome Friedman, R. Tibshirani, A sparse-group lasso, *Journal of Computational and Graphical Statistics* 22 (2013) 231–245. doi:10.1080/10618600.2012.681250.
- [11] R. Haw, H. Hermjakob, P. D’Eustachio, L. Stein, Reactome pathway analysis to enrich biological discovery in proteomics data sets, *Proteomics* 11 (2011) 3598–3613. doi:10.1002/pmic.201100066.
- [12] J. G. Burkhart, G. Wu, X. Song, F. Raimondi, S. McWeeney, M. H. Wong, Y. Deng, Biology-inspired graph neural network encodes reactome and reveals biochemical reactions of disease, *Patterns* 4 (2023) 100758. doi:10.1016/j.patter.2023.100758.
- [13] Y. Pershad, M. Guo, R. B. Altman, Pathway and network embedding methods for prioritizing psychiatric drugs, *Pac Symp Biocomput* 25 (2020) 671–682.

- [14] J. Leskovec, C. Faloutsos, Sampling from large graphs, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 631–636. doi:10.1145/1150402.1150479.
- [15] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance as a metric for image retrieval, *International journal of computer vision* 40 (2000) 99–121.
- [16] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* 20 (1987) 53–65.
- [17] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods* 3 (1974) 1–27.
- [18] D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence PAMI-1* (1979) 224–227.
- [19] R. Mussabayev, N. Mladenovic, B. Jarboui, R. Mussabayev, How to use k-means for big data clustering?, *Pattern Recognition* 137 (2023) 109269. doi:10.1016/j.patcog.2022.109269.
- [20] Y. Yu, Q. Liu, L. Wu, R. Yu, S. L. Yu, Z. Zhang, Untargeted attack against federated recommendation systems via poisonous item embeddings and the defense, *Proceedings of the AAAI Conference on Artificial Intelligence* (2023) 4854–4863. doi:10.1609/aaai.v37i4.25611.
- [21] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, J. Vilo, g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update), *Nucleic Acids Research* 47 (2019) W191–W198. doi:10.1093/nar/gkz369.

Prompting Large Language Models for Tailored Exercise Recommendations in Office Spaces

Gaetano Dibenedetto^{1,*}, Marco Polignano¹, Pasquale Lops¹ and Giovanni Semeraro¹

¹University of Bari Aldo Moro, Via Orabona 4, Bari, 70125, Italy

Abstract

In the digital era, *Recommender Systems* are a crucial component, commonly used in services such as music and movie streaming. Despite their widespread adoption, surprisingly little attention has been devoted to developing systems that can positively impact users' *well-being* and *health*. In an effort to combat the negative effects of a sedentary way of life on people's health and the subsequent rise in healthcare expenses, we introduce an encouraging approach, i.e. a recommender system that, through webcam-based monitoring of subject postures, suggests *personalized exercise breaks* to do directly near users' desks.

Our system captures users' postures during work hours and employs 3D pose estimation to calculate key angles between shoulders, hips, and head. By identifying postural imbalances, we generate exercise recommendations using a *Large Language Model* (LLM). The system flags potential postural issues when angle thresholds are exceeded and prompts the LLM to provide tailored exercise suggestions. Our method's effectiveness is assessed by experts in the field. While the results are still preliminary, our approach deserves further investigation, with future developments likely to focus on enriching the data and refining the detection methods. The full-reproducible code is available at the following link: <https://github.com/GaetanoDibenedetto/healthrecsys24>

Keywords

Health Recommender Systems, 3D Pose Estimation, Explainability, Large Language Model

1. Introduction

Extended periods of sitting have become unremarkable in modern lifestyles, but recent research highlights the detrimental effects it can have on our health. Studies have shown that prolonged sedentary behavior is linked to a higher risk of obesity and metabolic syndrome, characterized by elevated blood pressure, high blood sugar, excessive abdominal fat, and unhealthy cholesterol levels [1]. Furthermore, investigations by Marras et al. [2] suggest that prolonged static sitting postures may compromise the nutrition of intervertebral discs. More severe outcomes have also emerged from further investigation. Dunstan et al. [3] observed a strong relation between sedentary and premature mortality. Hamilton et al. [4] found relations with chronic illnesses, and Inoue et al. [5] with obesity. The implications of maintaining inadequate sitting postures over prolonged periods are indeed alarming. Conditions such as cervical spondylosis, lumbar diseases, and other ailments commonly known as "chair diseases" have been linked to poor sitting habits [6, 7, 8]. These musculoskeletal disorders can have a significant impact on an individual's quality of life, leading to chronic pain, reduced mobility, and decreased productivity.

Building on our previous work focused on posture correction for office workers [9], this research aims to extend that approach by developing a recommender system that suggests targeted physical activities to improve posture. We recognize that simply offering posture correction and, for this work, physical activity recommendations, does not guarantee that users will follow the system feedback aimed to help themselves. Hence, our goal, starting with this preliminary work, is to provide expert-level advice in a convenient and accessible format. By delivering daily, personalized recommendations, we

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ gaetano.dibenedetto@uniba.it (G. Dibenedetto); marco.polignano@uniba.it (M. Polignano); pasquale.lops@uniba.it (P. Lops); giovanni.semeraro@uniba.it (G. Semeraro)

🆔 0000-0001-6083-3600 (G. Dibenedetto); 0000-0002-3939-0136 (M. Polignano); 0000-0002-6866-9451 (P. Lops); 0000-0001-6883-1853 (G. Semeraro)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

hope to enhance user engagement and motivation, ultimately leading to gradual improvements in posture and overall wellness. In this context, the role of technology in promoting physical activity and exercise becomes even more critical, particularly in office environments where sedentary behavior is prevalent. To address these issues, we propose an approach that leverages webcam-based posture analysis and 3D pose estimation to monitor users' postures during working hours. By analyzing key angles between the shoulders, hips, and head, our system detects postural imbalances and provides personalized exercise recommendations, generated through a Large Language Model (LLM).

The main contributions of this work are summarized below:

- **3D Keypoint Extraction Model:** we propose a 3D Human Pose Estimation (HPE) keypoint extraction module able to detect the human pose from a single laptop camera;
- **Pose Classification Model:** we develop a system able to detect wrong poses while sitting at a desk efficiently;
- **Angles Computation:** we develop a vector-based angles computation module that calculates the angles between key body parts (shoulders, hips, and head) using 3D pose data. This helps detect postural imbalances by identifying deviations from normal alignment;
- **Recommendation Generation Module:** a module that utilizes a LLM to generate personalized exercise recommendations based on detected posture issues. When key angle thresholds are exceeded, the system prompts the LLM to suggest corrective exercises tailored to the user's specific posture;
- **Dataset publicly released:** the dataset with the 3D keypoints extracted from frames, used for training and evaluating our model will be publicly released for replicability purposes on Zenodo¹.

2. RELATED WORK

2.1. Posture Classification

As highlighted in our previous work [9], several studies share our goal of posture classification [10, 11, 12]. However, many of these approaches rely on data collected under strict constraints, e.g., users are often required to position themselves directly in front of the camera or assume predefined sitting postures; are required specialized accessories that are impractical and expensive such as multiple cameras or sensors installed in chairs.

These limitations make it difficult to apply such systems in a typical daily routine. In our previous work, we introduced a posture classification approach tailored for office workers, which avoids the constraints imposed by related works. Specifically, we allowed users to record themselves while working, without imposing restrictions. In that study, we collected our own dataset and developed a system capable of posture classification using a multi-layer perceptron. The system offered explanations based on data statistics and featured a personalized feedback module designed to correct user posture. Building on this, our current research aims to **enhance the system by leveraging 3D HPE** while maintaining the same dataset and classification technique. This approach shifts the focus towards creating a recommendation system to provide personalized exercise suggestions.

2.2. Physical Activity Recommendation

In today's information-overloaded world, recommender systems have emerged as essential tools for navigating countless options. Whether we're shopping, streaming, or socializing, these intelligent systems are constantly working behind the scenes to personalize our experiences. Generally, these systems work on a daily basis based on a user's previous behavior in a system.

¹<https://zenodo.org/records/13498794>

As discussed in [13, 14], mostly health recommender systems aim to improve the general well-being of users, such as recommending diets and exercise plans. However, developing recommender systems specifically for physical activity has proven challenging due to practical limitations. In particular, the scarcity of ratings for exercise activities, a cornerstone of traditional recommender systems. This is largely attributed to the difficulty in quantifying individual interest in specific exercises. Therefore, in our search for closely related works, we came across a limited number of works similar to the one we have proposed. We will discuss them below.

A Physical Activity Recommender aims to recommend a daily routine of physical activities and workouts to the user, this could be based on data as user's characteristics, his health status or other demographic information such as age and gender. One of the first work analyzed by us is RecFit [15], a context-aware recommender system, which systematically suggests physical activities based on the user's context, e.g. risk tolerance, budget, location, weather, but not on user's physical characteristics. However, it does not consider the user's physical characteristics and is not specifically designed to address health concerns. Advancements in the field have incorporated more physical details into the recommendation process. For instance, in 2020, Ferretto et al. [16] developed a system for patients with arterial hypertension, generating personalized recommendations based on the patient's age, gender, and physical condition. In 2021, Sengan et al. [17] proposed a system to prevent respiratory diseases, utilizing data on physical activity levels, heart rate, and respiratory rate. The most recent work we reviewed, DEEP-CARDIO [18], proposed in 2024, is a content-based system offering physical and dietary recommendations for cardiac patients. Using Density-Based Clustering (DBSCAN), it categorizes patient data, such as class predictions for cardiovascular diseases, alongside age and gender. However, the recommendations are presented in a categorical format, which may be challenging for non-expert users to interpret, e.g., instructions like "Follow Type 1 diet plan. Exercise Tip: Walking stationary, cycling, rowing, or water aerobics".

We observed that while these systems produce positive outcomes for patients, **they often present categorical outputs**, which can be **difficult for non-expert users to interpret**. In contrast, our goal is to develop a system that not only delivers specific recommendations but also includes brief descriptions. This approach will help non-expert users easily understand the recommendations and the rationale behind them.

2.3. Recommendation through LLMs

In recent years, following the introduction of Transformers [19], Large Language Models (LLMs) based on this architecture have been widely adopted across various machine-learning applications. Transformers have gained popularity primarily due to their attention mechanism, which improves the model's ability to represent text by focusing on the semantics of surrounding words.

Similar works have been found with our aim in the area of recommendations generated by LLMs, but only based on movies or books, which are the most common applications in the recommendation area, thanks to the large availability of datasets. W.-C. Kang et al. [20] propose one of the studies that we explored, which analyzed multiple LLMs across different settings, i.e., zero-shot, few-shot, and fine-tuning, by feeding the LLM with a prompt representing the user profile based on their past item ratings. Another work proposed by Sanner et al. [21], focuses on sequential recommendations by incorporating item descriptions and user preferences. While these approaches, like ours, leverage LLMs for generating recommendations through prompting, **these differs from our approach, which does not rely on an item-based dataset**, due to the lack of data in our scenario. Instead, our system's core lies in the *Angles Computation module* (Sec. 3.4), which detects anomalies in the user's posture. These anomalies are prompt to the system to generate personalized physical exercise recommendations aimed at improving the user's well-being.

3. PROPOSED APPROACH

3.1. Data Collection/Dataset

Up to our knowledge, there is a noticeable gap in the literature related to the availability of a specific public dataset within the domain of sitting poses. In order to bridge this gap, in our previous research, we created our own dataset, where workers in office spaces have been recorded while sitting. No constraints have been fixed in terms of camera angles or perspectives, minimum or maximum video lengths, or number of videos. We gathered videos from 10 subjects, processed them to extract frames and finally, they were annotated by humans. Some statistics of the dataset are reported in Tab. 1, all the details about, collection, filtering and processing of data are already provided in [9].

Table 1
Dataset Data statistic

Subject	Frames	INcorrect	Correct	correct%
ap	500	411	89	18%
gc	35	30	5	14%
gd	174	111	63	36%
gs	145	139	6	4%
k	49	49	0	0%
mc	31	30	1	3%
ms	47	42	5	11%
nl	44	41	3	7%
rt	23	23	0	0%
vr	99	99	0	0%
TOTAL	1147	975	172	15%

3.2. Keypoint Extraction Module

Building on our previous 2D-based approach [9], we shift to a 3D environment. This transition is made possible using MotionBERT [22], a 3D HPE model, implemented via the MMPose framework [23]. MotionBERT operates through 2D-to-3D lifting, where, during the pretraining phase, a motion encoder learns to reconstruct the underlying 3D motion from incomplete 2D observations by integrating geometric, kinematic, and physical insights about human movement. We selected MotionBERT over other HPE models due to its performance, since it achieves the lowest 3D pose estimation error to date on the Human3.6M dataset [24]. An example of the Keypoint Extraction Module is shown in Fig. 2b. The keypoints extracted on our Dataset are available on Zenodo¹ to guarantee reproducibility.

3.3. Pose Classification Model

To guarantee consistency, we used the same architecture with same hyperparameters proposed with our 2D-based approach. We used a Multilayer Perceptron architecture (Fig. 1), Adam optimiser and the BCE (Binary Cross Entropy) loss function, weighted with the class distribution in the training fraction. Training was conducted with a batch size of 20 on 1000 epochs with a learning rate of 0.001. The high number of training epochs is due to a slow learning process due to the high variability of the keypoints position. The best performing model based on the training loss was selected and saved to be used in the recommendation step.

The training phase of this architecture uses the keypoints extracted with MotionBERT previously described (Sec. 3.2). A data augmentation step is performed, considering the imbalance of the dataset, inserting the vertical flip of the correct postures. It is worth to note that only two subjects exhibited a quite balanced amount of correct posture data, leading to the consideration of a binary split: training and testing. Among these two subjects, the one with a more balance volume of data was designated as

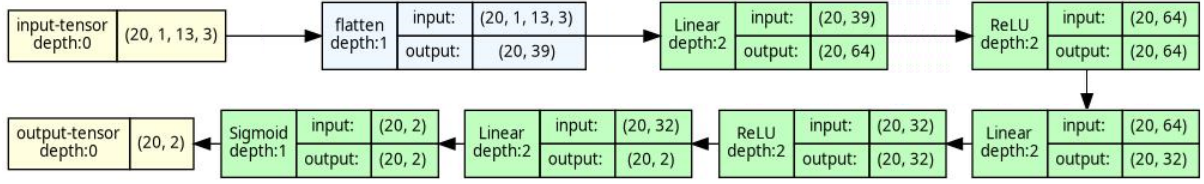


Figure 1: MLP architecture

the sole subject for the test split, i.e. *gd*, while the remaining subjects were allocated to the training split. The results of the classification are shown in the Tab. 2.

Table 2
Classification Report

	precision	recall	f1-score	support
INcorrect Posture	0.9135	0.8559	0.8837	111
Correct Posture	0.8797	0.9286	0.9035	126
accuracy			0.8945	237
macro avg	0.8966	0.8922	0.8936	237
weighted avg	0.8955	0.8945	0.8942	237

3.4. Angles Computation

We focus on computing specific angles related to the user’s posture, which are crucial for detecting imbalances and misalignments. The primary angles of interest are the inclination of the shoulders and hips, as well as the tilt of the head in both sideways and forward directions.

To compute these angles, we employ vector-based calculations. Using the 3D skeleton representation provided by MotionBERT, the position of each body part is defined as a point in 3D coordinate space. The vectors formed between these points allow us to determine their inclination relative to the coordinate axes. The angles are then derived using the dot product between these vectors and a reference axis, as detailed below:

Given two points representing different body parts, $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$, we compute the vector between these points as $\mathbf{v} = P_2 - P_1$. The general formula for computing the angle θ between two vectors \mathbf{A} and \mathbf{B} is:

$$\theta = \cos^{-1} \left(\frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} \right)$$

where $\mathbf{A} \cdot \mathbf{B}$ is the dot product, and $|\mathbf{A}|$ and $|\mathbf{B}|$ represent the magnitudes of the respective vectors. The direction of rotation can be inferred using the cross product of the two vectors.

In our case, vector \mathbf{A} corresponds to \mathbf{v} , the vector connecting two body parts, and vector \mathbf{B} is either the x -axis $[1, 0, 0]$ or the z -axis $[0, 0, 1]$, depending on the body part and the direction of interest.

- **Shoulder Inclination:** The shoulder inclination is computed relative to the x -axis. The shoulders are represented by a vector connecting the left and right shoulder points. Using the dot product, we calculate the angle between this vector and the horizontal axis to assess any unevenness in shoulder alignment.
- **Hip Inclination:** Similarly, hip inclination is measured with respect to the x -axis. A vector formed between the left and right hip points is analyzed to detect imbalances in the lower body, providing insight into hip misalignment.
- **Head Tilt:** To evaluate head posture, two distinct angles are computed:

- Sideways Tilt: This angle is measured relative to the x -axis and provides information on lateral imbalances in head and neck posture.
- Forward Tilt: The angle of forward head tilt is computed with respect to the z -axis, allowing us to assess how far the head is leaning forward compared to a neutral, upright position.

An example of angle detection is shown in Fig. 2.

By analyzing these angles using vector-based methods and dot products, we aim to detect deviations from normal alignment that could indicate issues with posture. Unfortunately, we were unable to establish specific angle thresholds to definitively define "anomaly" for various body parts, also with the support of experts in the field. To address this, we propose a data-driven approach. After calculating all angles in our dataset, we will determine the average angle for each specific body part. This average will serve as a threshold to flag any angles that deviate significantly from the norm. This threshold-based approach will serve as the foundation for generating personalized recommendations in the subsequent stage.

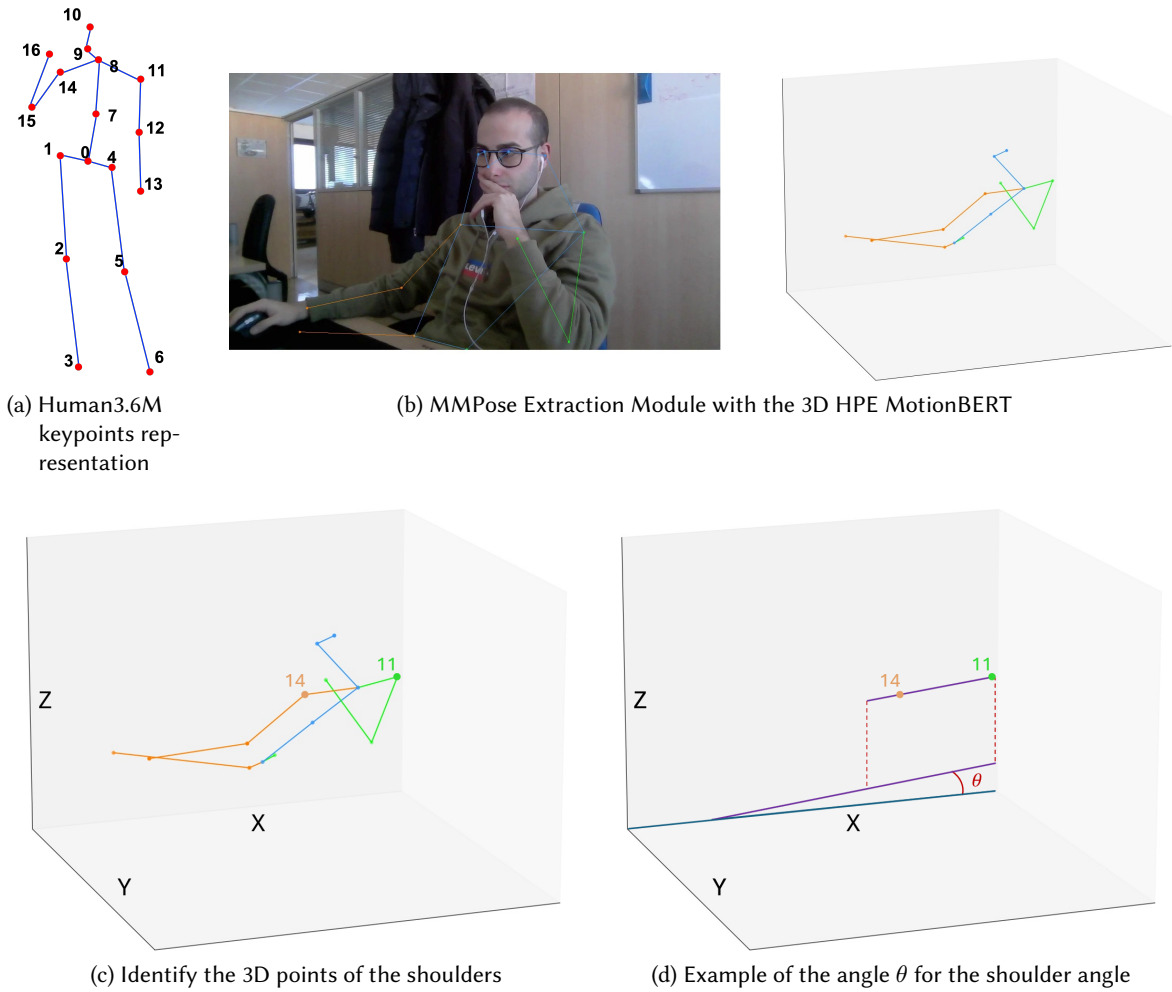


Figure 2: The figure illustrates an example of keypoints extracted from a frame, showcasing the computation of the shoulder inclination angle θ relative to the x -axis. Subfigure (a) shows the representation of the keypoints in the HPE model used; Subfigure (b) presents the output of the Keypoint Extraction Module (Sec. 3.2); Subfigure (c) highlights the 3D pose with a focus on the shoulder keypoints; Subfigure (d) details the computation of the shoulder inclination angle θ (Sec. 3.4)

3.5. Recommendation Generation

The angles, obtained as described previously, are used to detect anomalies in the user's posture. These detected anomalies are inserted in a textual prompt to be used as input of an LLM. The prompt starts with the sentence *"I've an incorrect posture caused by"*, then each anomaly is joint to the prompt with an "and" (except the first), and it ends with the sentence *", what exercise or stretching should I do?"*. If no anomalies are detected the prompt will be *"If I had an incorrect posture what exercise or stretching should I do?"*. The system prompt provided to the model is *"You are a medical assistant able to suggest a specific physical exercise for users with uneven body parts to prevent an increase in their unease"*. An example of the complete prompt provided to the LLM is shown in Listing 1.

Listing 1: LLM prompt example

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a medical assistant able to suggest a specific phisical exercise for users
with uneven body parts to prevent an increase in their unease.<|eot_id|><|
start_header_id|>user<|end_header_id|>

I've an incorrect posture caused by Shoulder Inclination , what exercise or
stretching should I do?<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Our recommendation generation process is powered by *Llama-3.1-8B-Instruct* [25]. We deploy the model on a machine using 2 NVIDIA A16 16GB VRAM GPUs, which required from 50 to 70 seconds for the text generation. Using the transformers library², we utilize the *"text-generation"* pipeline³ to generate the recommendations. The default hyperparameters from the pipeline are used, i.e. temperature=0.6, top-p=0.9, instead we set for the text generation length the *"max_new_tokens=1024"*. An output example of the recommendation is shown in the Fig. 3.

4. EXPERIMENTAL RESULTS

Due to the limited availability of data, we were unable to conduct a quantitative evaluation. Instead, we conducted a preliminary evaluation by administering questionnaires to domain experts, with at least 5 years of background experience in the field, specifically two physiotherapists. While the number of enrolled experts was small, their insights were crucial for understanding the pipeline's potential. They were asked to assess 10 different scenarios in which the user is seated in an incorrect posture. These scenarios were either manually selected by the evaluators from provided example images or uploaded by them using specific frames they found relevant. To facilitate this evaluation, we developed a website using a GRADIO⁴ interface, allowing users to interact directly with our system. The interface presents a chat-based interaction where users can either upload their own image or select one from a preloaded list of examples. Upon submission of an image, the system processes it through the keypoint extraction module and displays the results in the chat. The system also identifies any posture anomalies, providing the respective inclination angles for each anomaly detected. Following this, the system generates personalized exercise recommendations based on the detected anomalies. After reviewing the recommendations, users are prompted to answer a set of five questions on a 1 to 5 scale, followed by an open-ended feedback question. An example of interaction with the interface is shown in Fig. 3. The expert feedback collected through this process forms the basis of our evaluation. The questionnaire provided to the experts consists of the following questions:

- **Q1** - Do you agree with the recommended exercises based on the user's posture analysis? Rating: 1 (Totally Disagree) to 5 (Totally Agree)

²<https://huggingface.co/docs/transformers/index>

³<https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>

⁴www.gradio.app

- **Q2** - Do the recommended exercises adequately address the user's posture issue and imbalances? Rating: 1 (Totally Inadequate) to 5 (Totally Adequate)
- **Q3** - Are the exercises clearly explained and easy to understand? Rating: 1 (Very Unclear) to 5 (Very Clear)
- **Q4** - In your opinion, how effective would these exercises be in improving the user's posture? Rating: 1 (Not Effective at All) to 5 (Highly Effective)
- **Q5** - Are the recommended exercises simple enough for the user to perform without expert supervision? Rating: 1 (Not Simple at All) to 5 (Very Simple)
- **Q6** - What improvements would you suggest for this recommendation?

With these questions, we evaluate the correctness (Q1, Q2, Q4), accuracy(Q1), adequacy(Q2), clarity(Q3), effectiveness(Q4) and the simplicity of the execution (Q5) of the recommendation. The average rating for each question, on a scale from 1 to 5, is presented in Tab. 3. Overall, while none of the average results were negative (i.e., below 3), they were also not exceptionally high, except for Q3, which exceeded a rating of 4. This indicates that the clarity of the exercise recommendations was appreciated, but there is significant room for improvement in other areas. In particular, the average score for the "correctness" criterion was 3.4. After discussions with the evaluators, it became clear that, also if they appreciated the concept behind the system, they noted that it was difficult to determine whether a recommended exercise is entirely appropriate or not, starting with limited user characteristics, such as a single image. They emphasized that many more variables need to be considered for an accurate recommendation.

From the expert open-feedback, particularly the insights provided by physiotherapists, we identified several areas for improvement; these are discussed below to better illustrate limitations and future work.

Table 3
Evaluation Report

Q	Evaluation Criteria	Avg Rating
1	Accuracy	3.5
2	Adequacy	3.2
3	Clarity	4.2
4	Effectiveness	3.6
5	Simplicity	3.8
1,2,4	Correctness	3.4

5. Conclusion, Limitation and Future Works

This study has introduced a system for physical activity recommendation based on assessment of sitting postures utilizing accessible cameras, vector-based angles computation, and a LLM. Despite the promising results, several limitations were identified through open feedback evaluations:

- When the system detects uneven shoulders, it lacks to recommend exercises targeting the cervical region, which plays a significant role in connecting shoulder alignment and head posture;
- When the subject is sliding down in their chair (example in Fig. 3), the system may only detect misalignment in the head's inclination relative to the body;
- The angles computation module does not recognize shoulder lifting and excessive shoulder closure;

- The system should suggest to perform exercises in front of a mirror, which could help users maintain correct form.

As future work, based on these limitations, several ways for improvement have been identified. First of all, we need to improve the angle calculation module to detect a wider range of postural anomalies, enabling more accurate and varied exercise recommendations. Additionally, there is a need to explore the use of different LLMs to evaluate their performance in this recommendation task, and last but not least, a change of module that suggests the exercises, which could be based on a collected historical user data. These data are taken from the interactions of the current system, specifically combining the recommendations and the expert feedback: by storing the data and ratings provided by experts, the system could offer better recommendations to users with similar postural issues, leveraging the best-rated advice from prior interactions. Although the experimental results presented in this paper are preliminary, we can see a great potential to improve them. The exercise descriptions generated by the LLMs were particularly appreciated for their clarity, highlighting the possibility of using LLMs to improve users' understanding. In the future, by treating past recommendations and detected physical anomalies as items of a recommender system, we can refine the system in terms of correctness, adequacy, accuracy, and effectiveness.

Acknowledgments

The research is partially funded by PNRR - Mission 4 ("Education and research") – Component 2 ("From research to business"), Investment 3.3 ("Introduction of innovative doctorates that respond to the innovation needs of companies and promote the hiring of researchers by companies") D.M.n. 117/2023 - CUP: H91I23000170007 and is supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 – Partnerships extended to universities, research centers, companies, and research D.D. MUR n. 341 del 15.03.2022 – Next Generation EU (PE0000013 – “Future Artificial Intelligence Research – FAIR” - CUP: H97G22000210007).

We extend my sincere gratitude to Naps Lab S.r.l.s.⁵ for their support and collaboration in the realisation of this research.

We would also like to thank Ilaria Vianello⁶ and Michele Dilillo⁷, who provided invaluable assistance in the evaluation phase of this research.

⁵www.napslab.it

⁶www.linkedin.com/in/ilaria-vianello

⁷www.linkedin.com/in/michele-dilillo

Chatbot

MotionBert:

Detection:

Incorrect Posture

Uneven shoulders: -24.291795245316507

Forward Head Posture (Head leaning forward): -62.657037460799614

Recommendation:

Given your condition of uneven shoulders and forward head posture, I recommend the following exercises to help alleviate discomfort and improve your posture:

Stretching Exercises:

- Chest Stretch:** This stretch will help counteract the forward head posture by opening up your chest.
 - * Stand in a doorway with your hands on the doorframe at shoulder height.
 - * Lean forward until you feel a stretch in your chest and shoulders.
 - * Hold for 30 seconds.
- Shoulder Rolls:** This exercise will help loosen up your tight shoulder muscles and improve your posture.
 - * Roll your shoulders forward and backward in a circular motion.
 - * Repeat for 10-15 repetitions.
- Chin Tucks:** This exercise will help strengthen your neck muscles and improve your posture.
 - * Stand or sit with good posture.
 - * Tuck your chin in towards your chest, keeping your head level.
 - * Hold for 10 seconds.
 - * Repeat for 10-15 repetitions.

Strengthening Exercises:

- Scapular Squeezes:** This exercise will help strengthen your shoulder blades and improve your posture.
 - * Sit or stand with good posture.
 - * Squeeze your shoulder blades together and hold for 5 seconds.
 - * Release and repeat for 10-15 repetitions.
- Trapezius Strengthening:** This exercise will help strengthen the muscles in your upper back that help improve your posture.
 - * Hold a light weight (less than 1 pound) or a resistance band in your hand.
 - * Squeeze your shoulder blades together and lift the weight or band up towards the ceiling.
 - * Lower the weight or band back down and repeat for 10-15 repetitions.

Additional Tips:

- Practice good posture by keeping your shoulders back, chest open, and head level.
- Take regular breaks to stretch and move around, especially if you have a desk job.
- Consider working with a physical therapist or healthcare professional to develop a personalized exercise program.

Remember to consult with a healthcare professional before starting any new exercise program, especially if you have any underlying medical conditions or concerns.

1. Do you agree with the recommended exercises based on the user's posture analysis? Rating: 1 (Totally Disagree) to 5 (Totally Agree)

☐ 1
☐ 2
☐ 3
☐ 4
☐ 5

Use via API

Built with Gradio

Figure 3: Our system with Gradio Interface

References

- [1] M. Edward R. Laskowski, What are the risks of sitting too much?, <https://www.mayoclinic.org/healthy-lifestyle/adult-health/expert-answers/sitting/faq-20058005>, 2022. Accessed: 19-04-2024.
- [2] W. S. Marras, S. A. Lavender, S. E. Leurgans, F. A. Fathallah, S. A. Ferguson, W. Gary Allread, S. L. Rajulu, Biomechanical risk factors for occupationally related low back disorders, *Ergonomics* 38 (1995) 377–410. doi:<https://doi.org/10.1080/00140139508925111>.
- [3] D. Dunstan, E. Barr, G. Healy, J. Salmon, J. Shaw, B. Balkau, D. Magliano, A. Cameron, P. Zimmet, N. Owen, Television viewing time and mortality, *Circulation* 121 (2010) 384–391. doi:10.1161/CIRCULATIONAHA.109.894824.
- [4] M. T. Hamilton, D. G. Hamilton, T. W. Zderic, Role of Low Energy Expenditure and Sitting in Obesity, Metabolic Syndrome, Type 2 Diabetes, and Cardiovascular Disease, *Diabetes* 56 (2007) 2655–2667. URL: <https://doi.org/10.2337/db07-0882>. doi:10.2337/db07-0882.
- [5] M. Inoue, S. Yamamoto, N. Kurahashi, M. Iwasaki, S. Sasazuki, S. Tsugane, for the Japan Public Health Center-based Prospective Study Group, Daily Total Physical Activity Level and Total Cancer Risk in Men and Women: Results from a Large-scale Population-based Cohort Study in Japan, *American Journal of Epidemiology* 168 (2008) 391–403. URL: <https://doi.org/10.1093/aje/kwn146>. doi:10.1093/aje/kwn146.
- [6] I. Daian, A. van Ruiten, A. Visser, S. Zubic, Sensitive chair: a force sensing chair with multimodal real-time feedback via agent, in: W. Brinkman, D. Ham, B. L. W. Wong (Eds.), *Proceedings of the 14th European Conference on Cognitive Ergonomics: invent! explore!*, ECCE 2007, London, UK, August 28–31, 2007, volume 250 of *ACM International Conference Proceeding Series*, ACM, 2007, pp. 163–166. URL: <https://doi.org/10.1145/1362550.1362583>. doi:10.1145/1362550.1362583.
- [7] A. M. Lis, K. M. Black, H. Korn, M. Nordin, Association between sitting and occupational lbp, *European spine journal* 16 (2007) 283–298. doi:<https://doi.org/10.1007/s00586-006-0143-7>.
- [8] O. Okunribido, M. Magnusson, M. Pope, Low back pain in drivers: The relative role of whole-body vibration, posture and manual materials handling, *Journal of Sound and Vibration* 298 (2006) 540–555. URL: <https://www.sciencedirect.com/science/article/pii/S0022460X06004779>. doi:<https://doi.org/10.1016/j.jsv.2006.06.007>, special Issue on the Third International Conference on Whole-body Vibration Injuries.
- [9] G. Dibenedetto, M. Polignano, P. Lops, G. Semeraro, Human pose estimation for explainable corrective feedbacks in office spaces, in: *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2024*, Cagliari, Italy, July 1–4, 2024, ACM, 2024. URL: <https://doi.org/10.1145/3631700.3665184>. doi:10.1145/3631700.3665184.
- [10] S. Han, H. Kim, H. Choi, Rehabilitation posture correction using deep neural network, in: *2017 IEEE International Conference on Big Data and Smart Computing, BigComp 2017*, Jeju Island, South Korea, February 13–16, 2017, IEEE, 2017, pp. 400–402. URL: <https://doi.org/10.1109/BIGCOMP.2017.7881743>. doi:10.1109/BIGCOMP.2017.7881743.
- [11] J. C. T. Mallare, D. F. G. Pineda, G. M. Trinidad, R. D. Serafica, J. B. K. Villanueva, A. R. Dela Cruz, R. R. P. Vicerra, K. K. D. Serrano, E. A. Roxas, Sitting posture assessment using computer vision, in: *2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, 2017, pp. 1–5. doi:10.1109/HNICEM.2017.8269473.
- [12] Y. M. Kim, Y. Son, W. Kim, B. Jin, M. H. Yun, Classification of children’s sitting postures using machine learning algorithms, *Applied Sciences* 8 (2018). URL: <https://www.mdpi.com/2076-3417/8/8/1280>. doi:10.3390/app8081280.
- [13] Z. Zhao, A. Arya, R. Orji, G. Chan, Physical activity recommendation for exergame player modeling using machine learning approach, in: *8th IEEE International Conference on Serious Games and Applications for Health, SeGAH 2020*, Vancouver, BC, Canada, August 12–14, 2020, IEEE, 2020, pp. 1–9. URL: <https://doi.org/10.1109/SeGAH49190.2020.9201820>. doi:10.1109/SeGAH49190.2020.9201820.
- [14] J. Pincay, L. Terán, E. Portmann, Health recommender systems: A state-of-the-art review, 2019

- Sixth International Conference on eDemocracy & eGovernment (ICEDEG) (2019) 47–55. URL: <https://api.semanticscholar.org/CorpusID:189824028>. doi:10.1109/ICEDEG.2019.8734362.
- [15] Q. He, E. Agu, D. M. Strong, B. Tulu, Recfit: a context-aware system for recommending physical activities, in: S. K. S. Gupta, A. Banerjee (Eds.), *Proceedings of the 1st Workshop on Mobile Medical Applications, MMA '14*, Memphis, Tennessee, USA, November 3–6, 2014, ACM, 2014, pp. 34–39. URL: <https://doi.org/10.1145/2676431.2676439>. doi:10.1145/2676431.2676439.
- [16] L. R. Ferretto, E. A. Bellei, D. Biduski, L. C. P. Bin, M. M. Moro, C. R. Cervi, A. C. B. D. Marchi, A physical activity recommender system for patients with arterial hypertension, *IEEE Access* 8 (2020) 61656–61664. URL: <https://doi.org/10.1109/ACCESS.2020.2983564>. doi:10.1109/ACCESS.2020.2983564.
- [17] S. Sengan, S. V. R. H. Jhaveri, V. Varadarajan, R. Setiawan, L. Ravi, A secure recommendation system for providing context-aware physical activity classification for users, *Security and Communication Networks* 2021 (2021) 1–15. doi:<https://doi.org/10.1155/2021/4136909>.
- [18] A. Yashudas, D. Gupta, G. C. Prashant, A. Dua, D. AlQahtani, A. S. K. Reddy, Deep-cardio: Recommendation system for cardiovascular disease prediction using iot network, *IEEE Sensors Journal* 24 (2024) 14539–14547. doi:10.1109/JSEN.2024.3373429.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, December 4–9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [20] W. Kang, J. Ni, N. Mehta, M. Sathiamoorthy, L. Hong, E. H. Chi, D. Z. Cheng, Do llms understand user preferences? evaluating llms on user rating prediction, *CoRR abs/2305.06474* (2023). URL: <https://doi.org/10.48550/arXiv.2305.06474>. doi:10.48550/ARXIV.2305.06474. arXiv:2305.06474.
- [21] S. Sanner, K. Balog, F. Radlinski, B. Wedin, L. Dixon, Large language models are competitive near cold-start recommenders for language- and item-based preferences, in: J. Zhang, L. Chen, S. Berkovsky, M. Zhang, T. D. Noia, J. Basilico, L. Pizzato, Y. Song (Eds.), *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023*, Singapore, Singapore, September 18–22, 2023, ACM, 2023, pp. 890–896. URL: <https://doi.org/10.1145/3604915.3608845>. doi:10.1145/3604915.3608845.
- [22] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, Y. Wang, Motionbert: A unified perspective on learning human motion representations, in: *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, Paris, France, October 1–6, 2023, IEEE, 2023, pp. 15039–15053. URL: <https://doi.org/10.1109/ICCV51070.2023.01385>. doi:10.1109/ICCV51070.2023.01385.
- [23] M. Contributors, Openmmlab pose estimation toolbox and benchmark, <https://github.com/open-mmlab/mmpose>, 2020.
- [24] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, *IEEE Trans. Pattern Anal. Mach. Intell.* 36 (2014) 1325–1339. URL: <https://doi.org/10.1109/TPAMI.2013.248>. doi:10.1109/TPAMI.2013.248.
- [25] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, *CoRR abs/2407.21783* (2024). URL: <https://doi.org/10.48550/arXiv.2407.21783>. doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

Enhancing Health Recommendations through Patient Metadata Integration: A Persona-Based Evaluation Approach

Khushboo Thaker^{1,*}, Peter Brusilovsky¹, Daqing He¹, Youjia Wang², Mohammad Hassany¹, Behnam Rahdari¹, Heidi Donovan² and Young Ji Lee²

¹*School of Computing and Information, University of Pittsburgh, Pittsburgh, USA*

²*School of Nursing, University of Pittsburgh, Pittsburgh, USA*

Abstract

Patients and caregivers managing chronic diseases frequently seek information to navigate symptoms and make informed decisions as their conditions evolve. Health recommender systems are crucial in delivering personalized and relevant content tailored to individual needs. This paper presents a novel approach to enhancing content-based health recommendations by integrating patient metadata and trajectories. We developed a system that leverages these elements and evaluated its effectiveness with the assistance of healthcare professionals, specifically nurses. To address diverse needs and varied patient scenarios, we employed personas that simulate different patient journeys, ensuring the system's robustness across multiple contexts. Our findings indicate that incorporating user profiles with trajectory-based features significantly enhances the recommender system's performance. Our model enhanced with both profile and trajectory, achieved a Mean Average Precision (MAP) of 0.7318, demonstrating a 17 point increase. Our future research aims to utilize personas during the cold-start recommendation phase, where patient metadata is limited and persona can work as digital twin supporting patients journey.

Keywords

Health Recommender Systems, Personalized Health Information, Persona based Evaluation, Patient Metadata Integration

1. Introduction

Individuals with chronic illnesses such as diabetes, cancer, and heart disease are often deeply involved in managing their health. They actively seek out information to make informed decisions and effectively manage their conditions [1]. Traditionally, reliable health information has been disseminated through patient education materials and health literacy workshops [2, 3]. However, these resources are typically designed for a general audience and often fail to meet the specific needs of individual patients [3, 4].

In pursuit of more tailored health information, many patients turn to online search platforms. Unfortunately, these platforms frequently fall short in delivering the specialized, personalized content that patients require [5, 6]. As a result, patients increasingly seek information in disease-specific online health communities (OHCs), where they can connect with others who share similar experiences [7]. While beneficial, OHCs also pose significant risks of spreading misinformation, lacking the necessary infrastructure to ensure patients access the most relevant and verified information [8].

Consumer health search engines and recommender systems offer a promising solution to these challenges by providing personalized and reliable health information. Recent research has shown that enhancing the user experience on consumer health search engines, for example, through query reformulation for specialized health terms, can significantly improve search outcomes [9, 10]. Similarly,

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ k.thaker@pitt.edu (K. Thaker); peterb@pitt.edu (P. Brusilovsky); dah44@pitt.edu (D. He); yow14@pitt.edu (Y. Wang); moh70@pitt.edu (M. Hassany); ber58@pitt.edu (B. Rahdari); donovanh@pitt.edu (H. Donovan); leeyoung@pitt.edu (Y.J. Lee)

🌐 <http://kthaker.com> (K. Thaker)

🆔 0000-0003-3619-9376 (K. Thaker)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).


	Name	Maria -
	Profession	Administrative Assistant – stopped working after cancer
	Marital Status	Married
	Children	2 Children
	Maria's Age	58 (55 when diagnosed)
	Current Disease Trajectory State	Taxol plus Avastin for 4th recurrence; life threatening reaction to carboplatin (survivorship for progression/recurrence)
	Current Disease State	disease is progressing (CA-125 rising)
Maria's all Needs drug side effect management, fatigue		
<p>Maria is a 58 y.o. married woman with 2 adult daughters. She was diagnosed 3 years ago at age 55 and has had 4 recurrences since then. She has been on taxol plus avastin, but yesterday was taken off the Avastin for "too much protein in her urine". She is worried because the CA125 yesterday had taken a jump from 284 to 379. She also is desperate for something to help her fatigue. It is the one symptom she can't control. She is too tired to do normal house cleaning, crafts, going out to shop, much of anything she used to participated in. She continues to need more sleep and sometimes frequent naps. She gets depressed when she can't do what she wants to do. She does feel that she has a good quality of life....and still living after 3 1/2 years is a big deal to her!! She just wishes she could control and improve this fatigue.</p>		

Figure 1: Persona depicting a hypothetical patient Maria at disease progression stage

consumer health recommender systems can enhance user satisfaction by suggesting personalized content that incorporates user feedback [11, 12].

Despite these advancements, existing health recommender systems often focus on addressing users' immediate needs without sufficiently integrating user meta data like user profile and trajectory information. This limitation is particularly problematic in cold-start situations, where the system lacks adequate data to generate accurate recommendations. Additionally, current evaluation methods for these systems frequently overlook the diversity of user experiences, failing to account for the range of patient scenarios encountered in practice.

The primary motivation for this research is to address these limitations by developing a content-based recommendation system that integrates user profiles and trajectory information to enhance recommendation accuracy. Specifically, we aim to investigate how the inclusion of trajectory-based features can better address the diverse information needs of patients at different stages of their health journey. The key contributions of this paper are as follows:

- The design and implementation of a content-based health recommender system that effectively integrates user profiles and trajectory data, significantly improving recommendation accuracy.
- The introduction of a novel evaluation framework that employs personas, generated through clustering online health community data and refined via expert feedback, to simulate diverse patient scenarios.

The remainder of this paper is organized as follows: Section 2 details the system architecture and the methodology employed in this study. Section 3 presents the experimental results and discusses the findings. Section 4 offers implications, and potential areas for future research. Finally, Section 6 concludes the paper by summarizing the key findings and contributions.

2. Methodology

2.1. Creating Personas for Comprehensive System Evaluation

In designing a human-centered recommender system, it is critical to understand its target users, examine their information needs, and understand in which aspects the users might be different from each other. Specifically, in our case, where it is hard to evaluate the ongoing research on large patient population, which makes is essential to create a comprehensive user base to cover the broad spectrum of patient needs. For this we created ten Ovarian Cancer (OvCa) Patient and Caregiver personas by triangulating multiple sources, including online cancer forums, direct interviews with patients and caregivers, domain expert input, and clinical notes. The detailed process of persona creation is discussed in [13, 14]. To

Table 1

Clustering Features. All features were categorical.

Feature Name	Possible Values	Feature Name	Possible Values
User Type	Patient or Caregiver	Age	Above 50 or Below 50
Has Kids	Yes or No	Marital Status	Married, Separated, Divorced, Widow
Has Caregiver	Yes or No	Employed	Yes or No
Insurance	Yes or No	Cancer Stage	Early or Advanced
Disease Trajectory	Initial Treatment after Diagnosis, Survivorship after Initial Treatment (Remission), Survivorship during Treatment for Progression, Survivorship during Treatment for Recurrence, Survivorship after Recurrence/Progression, Advance Care Planning (End of Life)	Symptom Needs	Yes or No
Emotional Needs	Yes or No	Self-management Needs	Yes or No
Financial Needs	Yes or No	Treatment Needs	Yes or No
Sexuality Needs	Yes or No		

depict the appearance and details of a persona, we have included a comprehensive representation in Figure 1. The process include clustering of user base and then each cluster was enhanced by *OvCa* health professional to develop a general persona representing these cluster. Table 1 lists all the features utilized for clustering.

While creating personas it was revealed that the disease trajectory and information needs play key roles in differentiating groups of similar users. In contrast, demographics-related metadata, which serve as a key factor in grouping users in many existing recommender systems, were scarce and did not play a major role in forming the personas. This study provided us with important design implications for the development of personalized recommendation algorithms. First, the algorithms have to focus on adapting to both patients' disease trajectory and expressed information needs to recommend the most relevant documents. Since both factors could change over time, the system has to constantly monitor the users and keep their profiles up-to-date. Second, the presence of stable user clusters enables the system to perform efficient "cold-start" personalization at the start of the user's interaction with the system by matching the user to one of the clusters and offering a group-level adaptation. These implications affected the design of the system algorithms discussed in Section 2.3.

2.2. Data Source

The primary source of documents for this study is the HELPeR library ¹, a specialized repository meticulously organized into a topic hierarchy known as *OvCa TH* (Ovarian Cancer Topic Hierarchy) [15]. The *OvCa TH*, developed by a dedicated team of oncology nurses, categorizes documents into 17 broad categories and 90 specific sub-categories. This hierarchical structure ensures that documents are well-distributed collection across different categories. Such distribution is crucial for covering the diverse informational needs of *OvCa* patients and caregivers, ensuring that the library provides comprehensive and balanced content. The library consists of 2808 documents, a detail statistic about the library is provided in Table 3 and top 10 domains are listed in Table 2.

2.3. Recommendation Methods:

Our approach employs a content-based recommender system, which focuses on analyzing the content of documents to generate personalized recommendations. The input to the system includes explicit

¹<https://sites.pitt.edu/~dah44/helper/>

Table 2

Top 10 domains in HELPeR Library

Domain	Number of Articles
www.cancer.org	216
www.cancer.net	172
www.cancer.gov	155
www.mdanderson.org	61
www.mskcc.org	58
www.curetoday.com	56
www.ncbi.nlm.nih.gov	49
www.oncolink.org	38
clinicaltrials.gov	35
www.mayoclinic.org	31

Table 3

Summary Statistics of the HELPeR Library

Statistic	Value
Total Number of Documents	6,497
Manually Curated Documents	2,808
Crawler Collected Documents	3,689
Average Curated Documents per Broad Category	147
Median Curated Documents per Broad Category	109
Number of Unique Domains Crawled	141

user information, such as the patients’s profile (in our case Persona).

The output is a ranked list of documents that are tailored to the user’s specific needs.

2.3.1. Candidate Retrieval

In the initial phase of our recommendation pipeline, we focused on identifying a suitable candidate retrieval method. This step is crucial as it determines the pool of documents that will be further ranked and refined. The input query to this retrieval method is the need topics mentioned in the Patient Personas like nephropathy, financial assistance, cancer pain and hormonal therapy.

Our primary goal was not to find the absolute best model, but to select a reliable candidate ranking method that would allow us to conduct further experiments effectively. Based on previous benchmarks [16] on biomedical retrieval tasks we evaluated five underlying retrieval models: BM25, a well-established probabilistic model, served as our baseline [17]. We also explored BERT-based models, including BERT [18] and BioBERT [19] and MedCPT [16], which are designed to capture contextual nuances in text, particularly in the biomedical domain [16] and ColBERT [20, 17, 16], which has shown to be a strong baseline and generalized to bio-medical domain.

To conduct all these experiments, we utilized Pyserini [21], which integrates PyTorch and FAISS-based indexing for dense retrieval [22] and Lucene engine for BM25 retrieval. For FAISS indexing we utilized default parameters with number of clusters to be 1 as our dataset is small and we are not interesting in optimizing the performance. Further investigation is required to identify the best-performing models, we leave these enhancements for future as discussed in Section 4.1. In our dataset patient articles can be smaller less than 100 words to huge more than 11000 words with mean $2069(\pm 4091)$ we have indexed the documents at the section level. In this study, we employed several re-ranking methods to refine the initial retrieval results based on both patient profiles and their trajectories.

2.4. Profile and Trajectory-Based Re-Ranking

In this study, we implemented two primary re-ranking approaches to refine the initial retrieval results by enhancing the relevance to the patient’s specific profile and trajectory:

Table 4
Disease Trajectories for Ovarian Cancer Patients

Trajectory	Description
End of Life	The patient is in an advanced stage of disease where treatment is no longer curative but palliative. Focus may shift to symptom management, including pain relief, managing symptom, and emotional support. Advanced care planning, hospice care, and addressing emotional needs such as anxiety, grief, and fear of death become priorities.
Survivorship After Treatment	The patient has completed initial treatment and is transitioning to post-treatment care. This stage involves managing long-term side effects. Emotional challenges may include fear of recurrence and adjusting to a new normal. Patients focus on recovery, rehabilitation, and monitoring for recurrence.
Survivorship	The patient is currently receiving initial treatment for a new diagnosis. Information needs center around understanding treatment options, managing side effects, and coping with the emotional stress of a new cancer diagnosis. Patients may need support in navigating treatment decisions and maintaining quality of life during active treatment.
Survivorship During Progression/Recurrence	The patient is dealing with disease recurrence or progression, requiring new treatment strategies. Focus shifts to managing new or worsening symptoms, exploring second-line treatments, and addressing the emotional impact of disease recurrence. Information on advanced therapies, clinical trials, and symptom management becomes critical.
Previvorship	The patient is at high risk for developing ovarian cancer due to genetic predispositions but has not yet been diagnosed. Information needs include preventive strategies, such as risk-reducing surgeries, regular screening, and lifestyle modifications. Emotional challenges may include anxiety over potential diagnosis and decision-making regarding preventive measures.
Survivorship After Treatment (Recurrence/Progression)	The patient has completed treatment for recurrent or progressive disease and is now managing life post-treatment. There is a focus on managing long-term side effects of the second treatment, ongoing surveillance, and addressing fears related to further progression. Emotional support may be needed for coping with uncertainty, and physical rehabilitation may be required to manage lingering symptoms from aggressive treatments.

- **Profile Re-Ranking (Keyphrase-Based) :**

We extracted keyphrases from the patient's profile², which includes relevant attributes such as medical history, demographics, and treatment details. These keyphrases were then appended to the patient's stated need to generate a refined rank list. This approach aims to align the retrieved documents more closely with the patient's overall profile. We initially tried to use complete profile embeddings as well as all keyphrases but both of these methods performed poorly so we decided to restrict our keyphrases based on semantic types including drug, disease, therapeutic procedure, laboratory test.

- **Trajectory-Based Re-Ranking:** In this method, the patient's trajectory, or their current stage in the progression of their disease, was integrated into the recommendation process to provide more personalized information. A patient's disease trajectory refers to the stages they go through, from diagnosis to treatment and beyond, which significantly impact their information needs. For ovarian cancer patients, the type of information required can vary drastically depending on whether they are in early treatment, post-treatment survivorship, or dealing with a recurrence of the disease.

²Keyphrase Extraction was done using Scispacy with model en_core_sci_scibert. UMLS linking was utilized and semantic type of each keyphrase was extracted -<https://github.com/allenai/scispacy/tree/main>

To incorporate trajectory-based re-ranking, we created embeddings of the trajectory descriptions as listed in Table 4, which were used as query information to re-rank the top results. This approach ensured that documents related to the same information need but more specific to the patient’s trajectory were ranked higher. For example, if the patient was in the “Survivorship After Treatment” stage, their primary need might be managing side effects from chemotherapy. By embedding the description of this trajectory, the system would prioritize documents that address not only general chemotherapy side effects but those particularly relevant to patients who have completed treatment and are focusing on recovery and long-term management.

- **Combining Profile and Trajectory Re-Rankings:** We merged the rank lists generated from both the profile-based and trajectory-based re-rankings to produce a final combined rank list. This approach leverages the strengths of both methods to deliver a more comprehensive and contextually relevant set of results.

To generate a final, more comprehensive ranking, we combined the results from the profile-based and trajectory-based re-rankings using Reciprocal Rank Fusion (RRF). RRF is a technique used to merge multiple ranking lists by assigning a score to each document based on its position in each list. The RRF score for a document d in a ranking list r is calculated as:

$$\text{RRF_score}(d) = \sum_{r=1}^n \frac{1}{k + \text{rank}_r(d)}$$

where k is a constant (typically set to 60), and $\text{rank}_r(d)$ is the rank of document d in ranking list r . For our experiments, we gave equal weight to all ranking lists, ensuring a balanced integration of both profile-based and trajectory-based relevance. This combined approach allows us to generate a final ranking that takes into account both the patient’s detailed profile and their trajectory, leading to a more personalized and contextually appropriate set of recommendations. We restricted RRF to top 10 documents from each rank-list.

2.5. The Study Design

The study was designed to evaluate the effectiveness of different document recommendation algorithms within the HELPeR system, based on feedback from expert caregivers. By simulating real-world decision-making processes, the study aimed to assess how well these algorithms supported the selection of relevant documents for ovarian cancer patients and their caregivers. The study involved several key steps, as illustrated in Figure 3, guiding participants through a structured sequence of tasks that included document annotation, feedback, and selection.

Step 1: Pre-Survey

The study began with a pre-survey, which participants were required to complete. The survey collected key demographic information and assessed participants’ caregiving backgrounds, as summarized in Table 5. To reduce the number of articles for review experts were asked to look for a single information need of the patient. This limitation of the study is discussed in Section 4.1

Step 2: Persona Familiarization

Following the pre-survey, participants were introduced to a randomly selected hypothetical patient persona as discussed in Section 2.1. Participants were given time to familiarize themselves with the persona and were asked to confirm their understanding of the patient’s context. This step ensured that participants felt comfortable and confident in making document selections on behalf of the persona.

Step 3: Initial Document Feedback (Subtask 1)

The document list was created by first selecting the top ‘ n ’ ($n=10$) documents generated by each of the algorithms discussed in Section 2.3. These top documents were then combined to form a single list of unique documents. To eliminate any positional bias [23], the combined list was randomly sorted. Participants were subsequently presented with this randomized list of unique documents for their review and feedback. Every time a participant reviewed an article, we ensured that the persona was accessible to them (Refer 2). This approach allowed participants to refer back to the persona as needed,

HELPeR Expert Feedback System

Document #2 for Jennifer- [Done](#) [\[external link \]](#) [Go Back](#)

Title: Nerve Problems Peripheral Neuropathy and Cancer Treatment

1 of 2

Article

Nerve Problems (Peripheral Neuropathy) and Cancer Treatment

Some cancer treatments cause peripheral neuropathy, a result of damage to the peripheral nerves. These nerves carry information from the brain to other parts of the body. Side effects depend on which peripheral nerves (sensory, motor, or autonomic) are affected.

Damage to sensory nerves (nerves that help you feel pain, heat, cold, and pressure) can cause:

- tingling, numbness, or a pins-and-needles feeling in your feet and hands that may spread to your legs and arms
- inability to feel a hot or cold sensation, such as a hot stove
- inability to feel pain, such as from a cut or sore on your foot

Damage to motor nerves (nerves that help your muscles to move) can cause:

- weak or achy muscles that may cause you to lose your balance, trip easily, or have difficulty buttoning shirts or opening jars
- muscles that twitch and cramp or muscle wasting (if you don't use your muscles regularly)
- swallowing or breathing difficulties (if your chest or throat muscles are affected)

Damage to autonomic nerves (nerves that control functions)

People with nerve problems caused by cancer treatment need to take care to prevent falls. Sometimes integrative medicine practices, advised by your doctor, can also help you to feel better.

PERSONA

Jennifer's Age: 57 (57 when diagnosed)

Cancer Stage @ Diagnosis: Stage IIb

Current Disease Trajectory State: on chemo after initial diagnosis (survivorship)

Current Disease State: no evidence of disease

Jennifer's Current Needs: management of peripheral neuropathy, physical therapy, rehabilitation

Jennifer is a 57 y.o. who was diagnosed with ovarian cancer (Stage II-c clear cell) 3 months ago. She had sub-optimal debulking during surgery (with removal of a 40 cm tumor) but her last CT scan showed no evidence of disease and her Ca-125 is now in the normal range (14) from a high pre-surgery of 1372. Unfortunately, she developed severe peripheral neuropathy and the MD switched her to Taxotere at Cycle #3. She has had several falls and is now using a walker. She has gained 30 pounds and is hardly able to tolerate walks to the mailbox she is so deconditioned. Today is her next-to-last cycle of chemotherapy. The chemo nurses have been pushing her to ask for a referral to physical therapy or some kind of rehabilitation, but she has been so tired she couldn't face it. Now that she can see the light at the end of the tunnel, she wants to see what other people's experiences have been with PT or rehab for peripheral neuropathy.

Relevant to **Need**: ☐ Yes ☐ No

Relevant to **trajectory**: ☒ Yes ☐ No

Recommend this to User?: ☐ Yes ☒ No

If not to recommend, Please tell us the reason:

Submit **Reset** **FORM**

Figure 2: Feedback System used in the study. The left side displays the recommended article, while the right side shows the persona followed by the feedback form.

minimizing the risk of recall bias that could occur if they had to rely solely on their memory of the persona's details. For each document, participants were required to provide detailed feedback [24]. This feedback focused on the relevance of the document to the persona's needs and trajectory and the participant's assessment of the document's usefulness.

Step 4: Additional Document Feedback (Subtask 2)

Based on the feedback provided in Subtask 1, the HELPeR system generated a second set of documents. From the additional documents, the unique documents were selected to complement the initial set. These were shown to the participants for further feedback. This step was crucial in ensuring that the study captured feedback on a comprehensive set of relevant documents.

Step 5: Document Selection for Patient

In this step, participants were explicitly asked to assume the role of a nurse responsible for the patient. They were instructed to select up to three articles from the documents provided that they deemed most appropriate for the patient to read. Participants were required to justify their selections by indicating specific qualities that made these documents and how to perceive the understanding of this document for the presented patient. (Refer Table 5)

Step 6: Overall Feedback

Finally, participants were asked to provide overall feedback on the study. This included their impressions of the persona they interacted with, the quality and relevance of the documents encountered, and their thoughts on the HELPeR system. The design of this study was carefully structured to simulate a realistic caregiving scenario, allowing us to gather valuable insights into the usability and effectiveness of different recommendation algorithms within the HELPeR system, as judged by expert caregivers.

2.6. Participants

The study focused on assessing HELPeR from the perspective of expert caregivers. A total of **eleven** nurses from the School of Nursing were recruited for the study. These participants were selected based on their expertise and experience, with all participants possessing a minimum of **1 years** of nursing experience and all of them had care-giving experience. Additionally, each participant confirmed their understanding of ovarian cancer and their ability to provide care for ovarian cancer patients.

This study serves as a preliminary investigation, preceding a larger study involving patients and caregivers. The choice to use personas in the study was influenced by the fact that nurses typically

Table 5
Study Questionnaire

Question Group	Questions
Pre-Survey	1. Age 2. Education level 3. Employment status 4. Relationship to patients (caregivers) 5. Professional Experience with general caregiving 6. Experience with oncology caregiving in years 7. Experience with ovarian cancer caregiving in years
Subtask 1 & 2: Document Feedback	1. Relevance to Need 2. Relevance to Trajectory 3. Would you recommend this document?
Subtask 3: Document Selection for Patient	1. Which articles would you give to the patient? 2. What specific qualities of the document led you to select it? 3. Agreement with the statement: "The document is easy for the patient to understand"
Overall Feedback	1. Agreement with the statement: "This system would be beneficial for the care of my loved ones" 2. Agreement with the statement: "I would recommend this system to my loved ones"

Table 6
Participant Demographics and Experience

Category	Participants
Age Groups	
20-30 years	5
31-40 years	4
41+ years	2
Nursing Experience	
1-2 years	5
3-10 years	3
10+ years	3
Oncology Experience	
With Oncology Experience	8
Without Oncology Experience	3
Gyno-Oncology Experience	5

have limited information about a patient during a clinical visit, making the use of a persona a realistic representation of a real-world scenario.

To recruit participants, we employed a convenience sampling method due to the limited availability of qualified candidates. However, we ensured that the selected participants were representative of a diverse range of ages. Specifically, the group consisted of **eleven** participants aged between **20** and **44** years. The detailed statistics of the participants are provided in Table 6 Overall, this group of expert caregivers provided a representative sample for the study, allowing us to gather high quality gold-standard to evaluate recommender algorithms.

3. Experimental Results

3.1. Baseline Candidate Retrieval Performance

BM25 shows moderate performance, achieving a MAP of 0.4253 and an MRR of 0.2667 for patient need, but its effectiveness declines for patient trajectory (MAP: 0.2271, MRR: 0.2292) and expert recommendations (MAP: 0.1850, MRR: 0.1339). Among dense retrieval models, BioBERT outperforms BERT, especially in expert recommendation alignment (BioBERT MAP: 0.2289, MRR: 0.1339 vs. BERT MAP: 0.1736, MRR: 0.1250). However, both models fall short in trajectory relevance compared to sparse retrieval. As show in previous work [16], Fine-tuned models, ColBERT and MedCPT, lead in

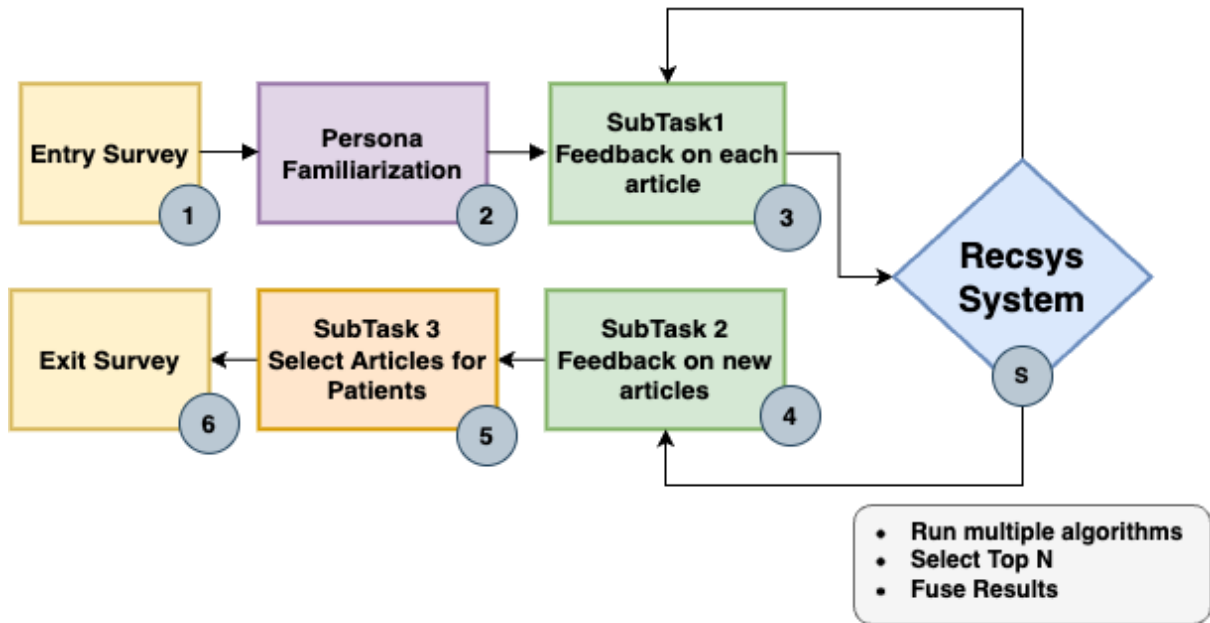


Figure 3: Flowchart of the User Study Design.

performance. MedCPT, in particular, excels across all criteria, with the highest MAP and MRR values (e.g., MAP: 0.5523 for need, 0.2815 for trajectory, and 0.4123 for expert recommendations), highlighting the benefits of fine-tuning for retrieval tasks.

The results demonstrate that fine-tuned dense retrieval models, especially MedCPT, offer superior accuracy in retrieving information relevant to patient needs, trajectories, and expert recommendations, surpassing both traditional sparse and standard dense retrieval approaches. For our further experiments, we will focus on BM25, ColBERT (general domain), and MedCPT (medical articles), as they are fine-tuned models and have demonstrated superior performance already.

3.2. Effect of integrating Patient Metadata

The performance of the re-ranking strategies across the BM25, ColBERT, and MedCPT models is summarized in the Table 8. MedCPT consistently demonstrates superior performance across all criteria, particularly in the combined profile and trajectory re-ranking approach, where it achieves a MAP of 0.7318 for Need, 0.6018 for Trajectory, and 0.6314 for Expert Recommendations. ColBERT also performs well, especially in the Trajectory-based re-ranking, indicating its effectiveness in integrating patient-specific trajectory information. In contrast, while BM25 shows reasonable performance, it falls behind the more advanced models, highlighting the significant benefits of incorporating both profile and trajectory data in enhancing retrieval relevance.

4. Implications

The use of personas, which are generic representations of patients at specific trajectory stages, offers a promising approach to addressing the cold-start problem in health recommender systems. These systems often struggle with limited patient data due to privacy concerns or patient reluctance to share information. Additionally, small population sizes in certain health contexts make collaborative filtering data scarce. By leveraging personas formed through representative patient stereotypes, health recommender systems can potentially provide personalized recommendations even in the absence of direct patient data, offering a viable solution for enhancing system effectiveness in these challenging scenarios.

Table 7

Performance of Candidate Retrieval Models on Biomedical Retrieval Tasks

Model	Need		Trajectory		Expert Rec.	
	MAP	MRR	MAP	MRR	MAP	MRR
Sparse Retrieval						
BM25	0.4253	0.2667	0.2271	0.2292	0.1850	0.1339
BERT	0.3820	0.2381	0.1426	0.1667	0.1736	0.1250
Dense Retrieval						
BioBERT	0.4195	0.2500	0.1926	0.2167	0.2289	0.1339
Dense Retrieval (fine-tuned for Retrieval)						
ColBERT	0.5476	0.2753	0.2458	0.2167	0.3173	0.2250
MedCPT	0.5523	0.2893	0.2815	0.2380	0.4123	0.2667

Table 8

Performance of Re-ranking based retrievals

Model	Need		Trajectory		Expert Rec.	
	MAP	MRR	MAP	MRR	MAP	MRR
Profile based re-ranking						
BM25	0.5371	0.3750	0.2678	0.2292	0.2645	0.1714
ColBERT	0.5397	0.2927	0.2458	0.2167	0.3863	0.2500
MedCPT	0.5611	0.2975	0.2381	0.2381	0.2381	0.2381
Trajectory based re-ranking						
BM25	0.6396	0.5018	0.5016	0.4917	0.4085	0.3518
ColBERT	0.6919	0.6676	0.5250	0.6636	0.5125	0.3321
MedCPT	0.7034	0.5327	0.5413	0.5191	0.6141	0.5163
Combine Profile and Trajectory re-ranking						
BM25	0.6431	0.5218	0.5127	0.5137	0.5251	0.3621
ColBERT	0.7092	0.6863	0.5750	0.6713	0.5365	0.3420
MedCPT	0.7318	0.6135	0.6018	0.5321	0.6314	0.5491

4.1. Limitations

While hybrid retrieval and fine-tuning approaches have shown potential for improving retrieval effectiveness [17], optimizing the retrieval pipeline was not the primary focus of this work. We utilized a basic zero-shot retrieval pipeline, which, while sufficient for our study’s objectives, may not fully leverage the potential of more advanced methods. In future work, we plan to explore and incorporate more sophisticated retrieval pipelines, including hybrid methods and fine-tuning techniques, to enhance the accuracy and relevance of health recommendations.

Recent advancements in large language models (LLMs) have demonstrated significant potential for enhancing personalization in recommender systems through zero-shot and few-shot prompting techniques [25]. These techniques have been particularly effective for tasks such as query expansion and user profile integration, allowing for more nuanced and context-aware recommendations [26, 27]. However, in this work, we did not explore these LLM-based approaches due to constraints on time and computational resources. While our current methods are effective for the study’s scope, integrating LLM techniques could potentially lead to more refined and personalized recommendations. In future research, we intend to investigate the application of LLMs to our recommendation framework, leveraging their capabilities to improve both query processing and user profile adaptation.

One of the primary limitations of this study is the relatively small sample size, with only 11 experts participating in the evaluation. This limited number of participants is insufficient to establish statistical significance for the results. Consequently, the findings should be interpreted with caution. However, it is important to note that this study serves as a preliminary investigation. Future research will involve a more comprehensive study with a larger sample size, including actual patients and caregivers. This expanded dataset will enable a more robust analysis and provide deeper insights into the effectiveness

and applicability of the proposed recommendation system.

5. Conclusion

In this study, we discussed and demonstrated the significance of integrating persona meta-data into health recommender systems. By leveraging personas—generic representations of patients at specific trajectory stages—we showed how these meta-data points can effectively address challenges in personalization, particularly in cold-start scenarios where patient data may be limited due to privacy concerns or patient reluctance. Our findings suggest that even minimal persona meta-data can enhance the relevance of recommendations, providing a strong foundation for further exploration in this area.

While we primarily focused on two specific meta-data points in this study, we recognize the potential of other persona attributes to further improve recommender systems. Our future research will explore how additional persona meta-data can be integrated to enrich recommendations and support more nuanced personalization.

Additionally, we collected qualitative feedback from nurse experts, which revealed a preference for selecting documents that are easy for patients to understand. This insight aligns with our ongoing research into knowledge adaptation [28], where we aim to tailor recommendations based on the patient's level of understanding and comprehension.

Acknowledgments

This work was supported by awards from the National Library of Medicine (NLM) of the National Institutes of Health (NIH) (Award Number: R01-LM013038). The content is solely the responsibility of the authors and NIH or NLM had no role in the design or conduct of the study; analysis, or interpretation of data; or preparation or review of the manuscript.

References

- [1] A. E. Anker, A. M. Reinhart, T. H. Feeley, Health information seeking: A review of measures and methods, *Patient Education and Counseling* 82 (2011) 346–354. URL: <https://www.sciencedirect.com/science/article/pii/S0738399110007470>. doi:<https://doi.org/10.1016/j.pec.2010.12.008>, methodology in Health Communication Research.
- [2] J. Susic, Nihseniorhealth classes for senior citizens at a public library in louisiana, *Journal of Consumer Health on the Internet* 13 (2009) 417–419.
- [3] A. O'Connor, D. Stacey, V. Entwistle, H. Llewellyn-Thomas, D. Rovner, M. Holmes-Rovner, V. Tait, J. Tetroe, V. Fiset, M. Barry, J. Jones, Decision aids for people facing health treatment or screening decisions., *The Cochrane database of systematic reviews* 2 (2003) CD001431.
- [4] E. L. Carter, G. Nunlee-Bland, C. Callender, A patient-centric, provider-assisted diabetes telehealth self-management intervention for urban minorities., *Perspectives in health information management* 8 (2011) 1b.
- [5] K. Lee, K. Hoti, J. Hughes, L. Emmerton, Consumer use of “dr google”: A survey on health information-seeking behaviors and navigational needs, *Journal of Medical Internet Research* 17 (2015).
- [6] K. Thaker, Y. Chi, S. Birkhoff, D. He, P. Brusilovsky, H. Donovan, Y. Lee, Exploring resource sharing behaviors for finding relevant health resources: An analysis of an online ovarian cancer community, *JMIR Cancer (Preprint) Accepted* 8 (2021) e33110.
- [7] S. Kanthawala, A. Vermeesch, B. Given, J. Huh, Answers to health questions: Internet search results versus online health community responses, *Journal of Medical Internet Research* 18 (2016).
- [8] A. C. Johnston, J. L. Worrell, P. Gangi, M. Wasko, Online health communities: An assessment of the influence of participation on patient empowerment outcomes, *Inf. Technol. People* 26 (2013) 213–235.

- [9] Jimmy, G. Zucco, B. Koopman, Choices in knowledge-base retrieval for consumer health search, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), *Advances in Information Retrieval*, Springer International Publishing, Cham, 2018, pp. 72–85.
- [10] D. He, Z. Wang, K. Thaker, N. Zou, Translation and expansion: Enabling laypeople access to the covid-19 academic collection, *Data and Information Management* 4 (2020) 177–190.
- [11] R. D. Croon, L. V. Houdt, N. N. Htun, G. Štiglic, V. V. Abeele, K. Verbert, Health recommender systems: Systematic review, *Journal of Medical Internet Research* 23 (2021).
- [12] M. Etemadi, S. B. Abkenar, A. Ahmadzadeh, M. H. Kashani, P. Asghari, M. Akbari, E. Mahdipour, A systematic review of healthcare recommender systems: Open issues, challenges, and techniques, *Expert Syst. Appl.* 213 (2023) 118823.
- [13] Y. Wang, K. Thaker, V. Hui, P. Brusilovsky, D. He, H. Donovan, Y. J. Lee, Utilizing digital twin to create personas representing ovarian cancer patients and their families, in: *Innovation in Applied Nursing Informatics*, IOS Press, 2024, pp. 754–756.
- [14] K. Thaker, V. Hui, Z. Luo, Y. Wang, B. Rahadari, P. Brusilovsky, D. He, H. Donovan, Y. J. Lee, Walk a mile in their shoes: Using personas to obtain health professional perspectives for an ovarian cancer patient-focused recommender system, in: *Proceedings of the AMIA Annual Symposium*, AMIA, New Orleans, USA, 2023.
- [15] B. Rahdari, P. Brusilovsky, D. He, K. M. Thaker, Z. Luo, Y. J. Lee, Helper: An interactive recommender system for ovarian cancer patients and caregivers, in: *Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 644–647.
- [16] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval, *Bioinformatics* 39 (2023).
- [17] N. Thakur, N. Reimers, A. Rüklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [18] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, TinyBERT: Distilling BERT for natural language understanding, in: T. Cohn, Y. He, Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, 2020, pp. 4163–4174. URL: <https://aclanthology.org/2020.findings-emnlp.372>. doi:10.18653/v1/2020.findings-emnlp.372.
- [19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics (Oxford, England)* 36 (2020) 1234–1240.
- [20] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, ColBERTv2: Effective and efficient retrieval via lightweight late interaction, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3715–3734. URL: <https://aclanthology.org/2022.naacl-main.272>. doi:10.18653/v1/2022.naacl-main.272.
- [21] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, Association for Computing Machinery, New York, NY, USA, 2021, p. 2356–2362.
- [22] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, *IEEE Transactions on Big Data* 7 (2019) 535–547.
- [23] Y. Yue, R. Patel, H. Roehrig, Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data, in: *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, Association for Computing Machinery, New York, NY, USA, 2010, p. 1011–1018.
- [24] M. Prince, 9 - epidemiology, in: P. Wright, J. Stern, M. Phelan (Eds.), *Core Psychiatry (Third*

Edition), third edition ed., W.B. Saunders, Oxford, 2012, pp. 115–129.

- [25] P. Liu, L. Zhang, J. A. Gulla, Pre-train, prompt, and recommendation: A comprehensive survey of language modeling paradigm adaptations in recommender systems, *Transactions of the Association for Computational Linguistics* 11 (2023) 1553–1571.
- [26] A. Acharya, B. Singh, N. Onoe, Llm based generation of item-description for recommendation system, in: *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23*, Association for Computing Machinery, New York, NY, USA, 2023, p. 1204–1207.
- [27] Y. Wu, R. Xie, Y. Zhu, F. Zhuang, X. Zhang, L. Lin, Q. He, Personalized prompt for sequential recommendation, *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [28] K. Thaker, Ka-recsys: Knowledge appropriate patient focused recommendation technologies, in: *Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 720–721. URL: <https://doi.org/10.1145/3523227.3547422>. doi:10.1145/3523227.3547422.

Personalizing Exercise Recommendations with Explanations using Multi-Armed Contextual Bandit and Reinforcement Learning^{*}

Parvati Naliyattthaliyazchayil^{1,†}, Deepishka Pemmasani², Navin Kaushal³, Donya Nemati⁴ and Saptarshi Purkayastha⁵

¹Dept. of Biomedical Engineering and Informatics, Indiana University Indianapolis, Indianapolis, Indiana, USA

²Dept. of Biomedical Engineering and Informatics, Indiana University Indianapolis, Indianapolis, Indiana, USA

³Dept. of Health Sciences, Indiana University Indianapolis, Indianapolis, Indiana, USA

⁴College of Nursing, The Ohio State University, Ohio, USA

⁵Dept. of Biomedical Engineering and Informatics, Indiana University Indianapolis, Indianapolis, Indiana, USA

Abstract

We present an innovative mobile exercise recommendation app that leverages clinical guidelines from authoritative sources to provide personalized, safe exercise suggestions. Our approach addresses two critical challenges in health-focused recommender systems: the cold start problem and user motivation through explainable AI. To overcome the initial lack of user data, we employ a two-stage process: We use Deep Q-Network (DQN) reinforcement learning to generate 2000 synthetic user profile. The DQN learns a reward function based on clinical guidelines, ensuring that the generated profiles align with established medical advice. These synthetic profiles bootstrap a multi-armed contextual bandit algorithm. This algorithm recommends the most suitable exercises for a given user persona, determined by a combination of comorbidities, age, and preferred exercise criteria. Our method's key innovation lies in its ability to mimic a large cohort of clinically safe user profiles without requiring real-world participants, effectively eliminating the cold start problem while maintaining medical appropriateness. To enhance user engagement and promote behavior change, we implement an explainability layer. Unlike black-box deep learning recommenders, our system provides transparent justifications for each recommendation. By highlighting the importance of specific features used in the decision-making process, we help users understand why a particular exercise is recommended for their persona. This recommender system is being incorporated into an existing mobile app, which will be trialed with healthy and cardiovascular disease patients.

Keywords

Mobile Health, Exercise Recommender System, Reinforcement Learning(RL), Explainable AI Deep Q-Network

1. Introduction

Physical exercise, widely recognized as a "miracle cure," remains underutilized despite its critical role in health maintenance and chronic disease management [1]. The World Health Organization reports that approximately one-third of the global adult population—1.8 billion individuals—are physically inactive [2]. A primary reason for this is that though people easily form habits around everyday activities, exercise is often something that is contemplated rather than consistently practiced. However, by turning exercise into a regular habit, individuals can significantly change their exercise behavior for the better and improve their health outcomes.

To address this, we have developed a novel mobile recommender that provides personalized exercise recommendations based on clinical guidelines, helping users build and sustain exercise habits. With mobile devices becoming integral to daily lives, mobile Recommender Systems (RS) have gained traction in healthcare interventions, though their application to physical activity promotion remains limited [3].

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author: parumenon.pm@gmail.com

✉ parumenon.pm@gmail.com (P. Naliyattthaliyazchayil); dpemmasa@iu.edu (D. Pemmasani); nkaushal@iu.edu (N. Kaushal); nemati.9@osu.edu (D. Nemati); saptpurk@iu.edu (S. Purkayastha)

ORCID 0009-0003-5917-4558 (P. Naliyattthaliyazchayil)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This is due to challenges such as the "user cold start" problem, limited data for analytics [4, 5], and the opaque nature of Deep Reinforcement Learning (DRL) algorithms [6]. Our research addresses these challenges by leveraging trusted clinical guidelines and incorporating an explainability layer, making the system reliable, safe, and personalized to the user's medical history and needs.

The user cold start problem, which arises when there is insufficient behavior information about new users, has been addressed in various ways in the literature. Approaches include clustering existing users to predict new user behavior [4], employing multi-phase algorithms for user classification, getting neighbours and outcome prediction [7], and few systemic reviews summarizing methods used including using auxiliary data to augment user profiles as data approach and using various Machine Learning algorithms along with content based algorithms as method driven approach.[8]. Our approach uniquely integrates clinical guidelines practices into synthetic user profiles that can be used to train the recommender, overcoming this issue. For this, we employ a two-stage process:

- Collecting the exercise recommendation guidelines for various medical conditions from their respective authoritative sources and structuring it into Machine Learning consumable format for further utilization.
- Utilizing DQN to generate 2,000 synthetic user profiles using structured clinical guidelines from step-1 ensuring the generated profiles align with established medical advice.

These synthetic profiles are used to train Multi-Armed Contextual Bandit (MAB) algorithm which recommends exercises making it extremely safe and useful when we don't have past behavior history. While these generated recommendations are clinically safe, they are also tailored to fit along with the other user choices like having a workout buddy or exercising at home etc., This effectively replaces the need for real-world data for training, thus addressing the cold start problem and lack of publicly available datasets, while ensuring medical accuracy.

To enhance user engagement and promote behavior change, we incorporate an explainability layer into our system. This approach aligns with Explainable Artificial Intelligence (XAI) principles, which aims to provide transparency in algorithmic decision-making processes [9]. Research indicates that an improved understanding of treatment correlates with better adherence [10, 11] and can enhance engagement and foster behavior change [12].

Our recommender system is being integrated into an existing mobile application and will be tested with both healthy individuals and those with cardiovascular conditions. This research contributes to the growing field of personalized digital health interventions by addressing key challenges in exercise recommendation systems.

2. Methodology

The design of our novel exercise recommendation system comprises three main components: guideline structuring, Deep Q-Network (DQN) for synthetic data generation, and a Multi-Armed Contextual Bandit (MAB) algorithm for personalized recommendations. We also incorporate an explainable AI (XAI) layer to enhance transparency and user engagement.

2.1. Guideline Structuring and Base File Creation

The goal of this step is to structure exercise guidelines from authoritative sources (e.g., American Heart Association, National Kidney Foundation) into a standardized and consumable format by DQN. The first step to structuring unstructured guidelines is to identify key attributes available commonly across most guidelines. Key attributes that were identified were age, gender, medical history, exercise preference, frequency, and duration. Next, the value of each attribute was collected from each guideline used. For e.g., value of attribute 'exercise preference' is cardio if the guideline refers to preferred exercise being walking or running. Each unique combination of attribute values was assigned a reward score (0-1) based on adherence to guidelines. For e.g., the American Heart Association(AHA) recommends at

least 150 minutes of moderate-intensity aerobic activity per week, with additional benefits for patients with cardiac diseases who engage in at least 300 minutes (5 hours) per week [13]. Consequently while structuring this guideline a record with 30-45 minutes of aerobic activity per day for 5 days a week will receive a higher "reward" compared to a record with 15-30 minutes of activity per day for 3 days a week because 30-45 minutes of activity for 5 days a week closely aligns with AHA guideline in this example. Other examples of exercise guidelines include the National Kidney Foundation's recommendation for continuous activity involving large muscle groups, aiming for 30-minute sessions [14], and the American Diabetes Association's guideline of 150 minutes of moderate-intensity exercise per week [15]. This structured dataset, termed the "base file," formed the foundation for subsequent steps. Figure 1 shows couple rows from base file showing how the structure looks:

age	gender	medical history	exercise preference	exercise frequency	exercise duration	reward
35	Male	Cardiac arrest	Cardio	5 days	30-45 mins	0.7
42	Male	Cardiac arrest	Cardio	3 days	15-30 mins	0.4

Figure 1: Example rows from base file

2.2. DQN for synthetic data generation

We employed a Deep Q-Network (DQN), a model-free, off-policy reinforcement learning algorithm [16, 17], to generate synthetic user profiles. A user profile is referring to each user's demographics, pre-existing medical conditions along with exercise goals and exercise preferences. The DQN system comprised three main components: environment construction, state representation, and recommendation policy learning.

1. **Environment Construction:** This phase involves creating an environment based on user behavior history[17]. In cold start scenarios, as in our study, the base file, developed in the previous step according to specific guidelines, is used to construct this environment.
2. **State Representation:** The environment generates a state representation that typically includes user demographics and past behaviors[17]. Our study utilizes demographic data randomly generated from defined value sets along with the base file data supplementing the absence of historical behavior for user cold start cases.
3. **Recommendation Policy Learning:** Guided by rewards derived from the base file's "reward" column. This reward value is carefully designed to reflect how closely the exercise recommendations adhere to clinical guidelines based on user attributes such as age, gender, medical history, exercise preference, frequency, and duration. Using this structured reward system, the model ensures that the generated recommendations are safe and tailored to individual needs.

DRL has the unique ability to leverage deep learning to approximate the value function in RL and solve high-dimensional Markov Decision Processes (MDPs)[17]. The DQN agent selected actions (attribute-value pairs) according to the policy at a given state, with rewards determined by matching rows in the base data. The agent updated its Q-values based on received rewards, learning optimal actions to generate guideline-adherent profiles. Key hyperparameters included a learning rate of 0.001, a discount factor of 0.95, an initial epsilon of 1.0, an epsilon decay of 0.995, and a minimum epsilon of 0.01.

2.3. Multi-Armed Bandit (MAB) Algorithm Implementation

To develop a personalized exercise recommendation system, we implemented a Multi-Armed Contextual Bandit (MAB) algorithm using the LinUCB (Linear Upper Confidence Bound) approach [18]. The system was designed to learn from and adapt to individual user profiles and behaviors over time, based on a dataset of 2000 users' exercise profiles and characteristics.

In our study, the MAB model defined three arms corresponding to the main exercise types: cardio, strength, and flexibility. The context for each user was represented as a feature vector comprising demographic information (age, sex, race), medical history, exercise preferences (frequency, duration, location), and other relevant attributes. To improve the algorithm's performance, we applied feature engineering techniques, including normalization of numerical features, one-hot encoding of categorical variables, and creation of interaction terms.

The LinUCB algorithm was implemented to balance exploration and exploitation in recommendation selection [19]. Exploration refers to the algorithm's attempt to try different exercise routines for a user, even if it is uncertain about their effectiveness, to gather more data. Exploitation, on the other hand, involves recommending exercises that have already shown positive results[20]. For each arm a , we maintained a matrix A_a and vector b_a to estimate the coefficients θ_a . For instance: For the cardio arm, A_{cardio} a matrix that tracks features like age, sex, race, frequency, duration of the cardio activity, while b_{cardio} is a vector representing the corresponding observed rewards like exercise completion or adherence. At each interaction, the algorithm computed a score for each arm based on the current context and coefficient estimates, selecting the arm with the highest score. The model parameters were updated after each interaction using the observed reward, which was defined as a weighted combination of short-term engagement (exercise completion) and long-term health outcomes (progress towards weekly goals).

2.3.1. Incorporating Contextual Information

To further refine the recommendations, we incorporated a contextual bandit model, a variant of the MAB framework that allows the algorithm to consider additional contextual information before making decisions[20]. In our case, context included variables such as age, exercise preference, duration, frequency and strength preference. By including these contextual factors, the algorithm could tailor its recommendations more closely to the user's current state and environment, thereby increasing the likelihood of user engagement and adherence to the exercise plan. The performance of the MAB-based recommendation system was evaluated by simulating user interactions with the synthetic data.

2.4. Explainable AI Layer (XAI)

To enhance transparency and foster trust, we incorporated an explainable AI layer. This layer provides insights into the rationale behind specific exercise recommendations, considering user medical history, user choices, and relevant health guidelines. The XAI component aims to support clinical adoption, ensure greater accuracy, minimize risks associated with errors or biases, and enhance user engagement [21] fostering behaviour change.

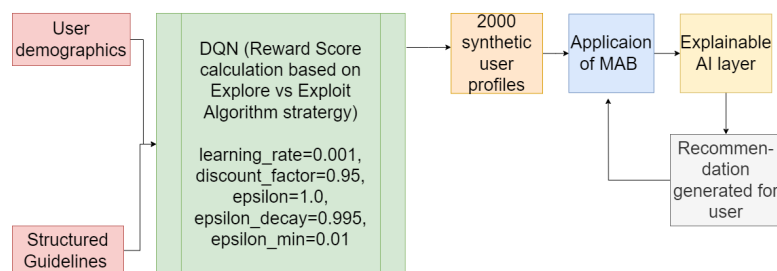


Figure 2: Methodology

3. Results and Discussion

As outlined in the methodology, we conducted an attribute analysis to identify the key variables and their respective valuesets necessary for structuring the exercise guidelines to create base file. Upon the

creation of the base file, it was utilized in the DQN model to generate profiles. Key parameters such as average reward per episode and epsilon decay were monitored to evaluate the Q-agent's learning progress and improvement over time. The results indicated a consistent average reward of 0.5 per episode, suggesting that the model effectively adhered to the guidelines and optimized its performance as training progressed as shown in 3.

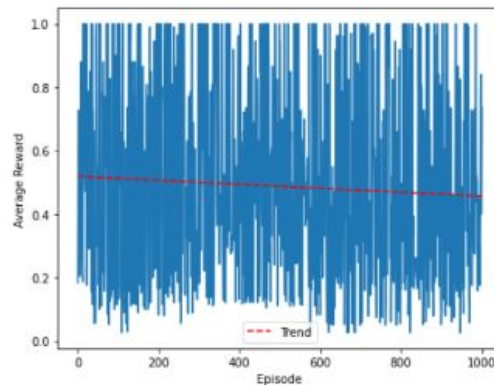


Figure 3: Average Reward per episode

Additionally, the epsilon value steadily decreased throughout the agent's lifecycle, demonstrating that the agent learned to act more optimally with experience, as depicted in the graph below Figure 4.

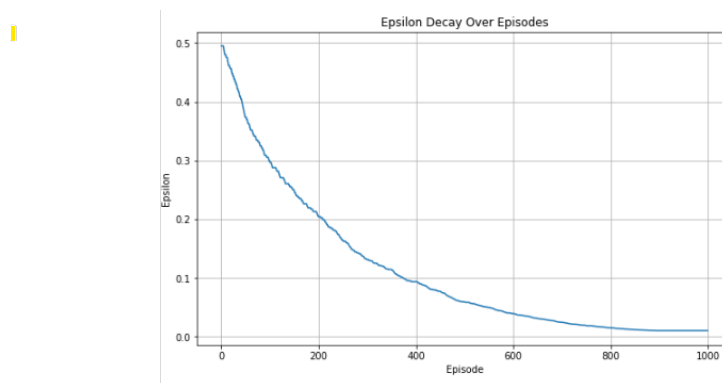


Figure 4: Epsilon Decay Over Episodes

To assess the similarity between real and synthetic data, we employed the Kolmogorov-Smirnov (KS) test, comparing the distributions of various features. Features such as Strength Training, Strength Preference, Exercise Location, Cardio Preference, Gender, Exercise Duration, Preferred Exercise, and Medical Exercise yielded high p-values (close to 1), indicating that their distributions in the synthetic data closely matched those in the real data. This result suggests that the DQN model successfully captured the essential patterns in the data while also adapting and improving its policy through exploration.

Additionally, we conducted a Feature Importance Similarity analysis to compare the importance of features when models were trained on real versus synthetic data. The analysis produced a Feature Importance Similarity score of 0.9787, indicating a high degree of similarity. This suggests that the synthetic data effectively captured the critical features. Figure 5 shows a Feature Importance Comparison, with Age used as the target variable.

3.1. MAB Algorithm Training and Evaluation

We evaluated the performance of our Multi-Armed Contextual Bandit (MAB) algorithm using the LinUCB approach over a simulated period of 30 days, with 2000 synthetic user profiles generated by

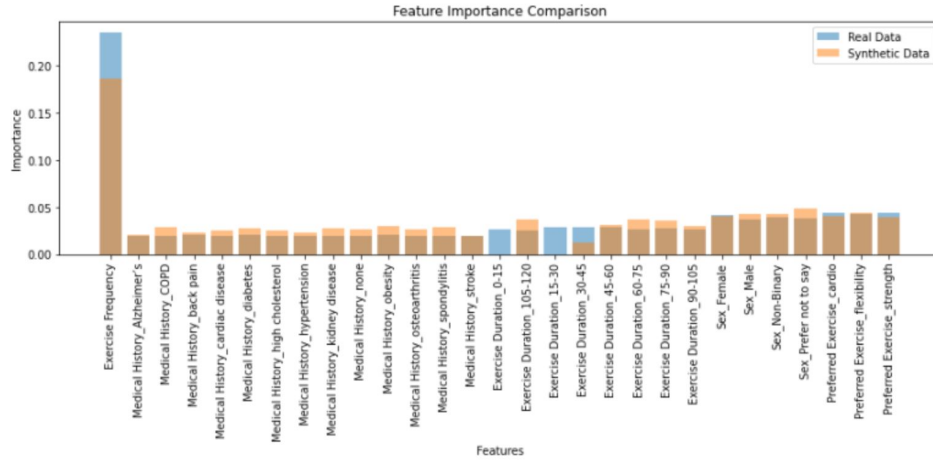


Figure 5: Feature Importance Comparison

the Deep Q-Network (DQN). The evaluation focused on the algorithm's ability to provide personalized exercise recommendations and adapt to user profiles over time.

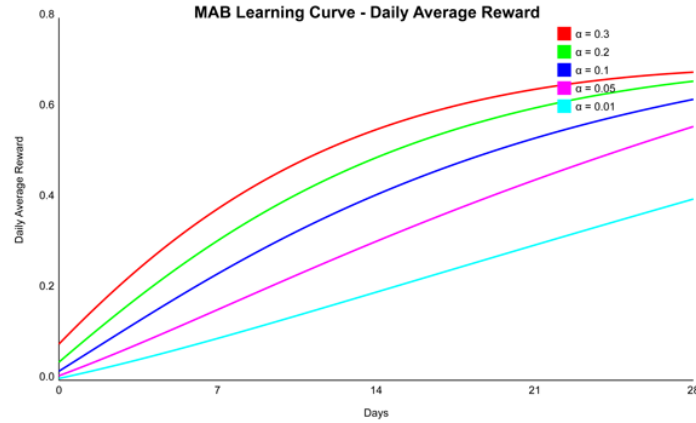


Figure 6: MAB Learning Curve - Daily Average Reward

3.1.1. Convergence and Learning Rate

The MAB algorithm demonstrated rapid convergence, with the average reward stabilizing after approximately 15 days of simulated users interactions. Figure 6 illustrates the learning curve, showing the daily average reward across all users.

The learning rate, α , was set to 0.1, which provided a balance between quick adaptation and stability. We observed that higher learning rates (e.g., 0.2, 0.3) led to faster initial convergence but increased volatility, while lower rates (e.g., 0.05, 0.01) resulted in slower learning but more stable long-term performance.

3.1.2. Cumulative Regret

Cumulative regret, a key metric for evaluating MAB algorithms, measures the difference between the optimal and actual rewards received over time. Our LinUCB implementation achieved a sub-linear cumulative regret, as shown in Figure 7.

The final cumulative regret after 30 days was 487.3, which is 18.9% lower than a standard ϵ -greedy approach (600.5) and 32.4% lower than a random selection baseline (721.6).

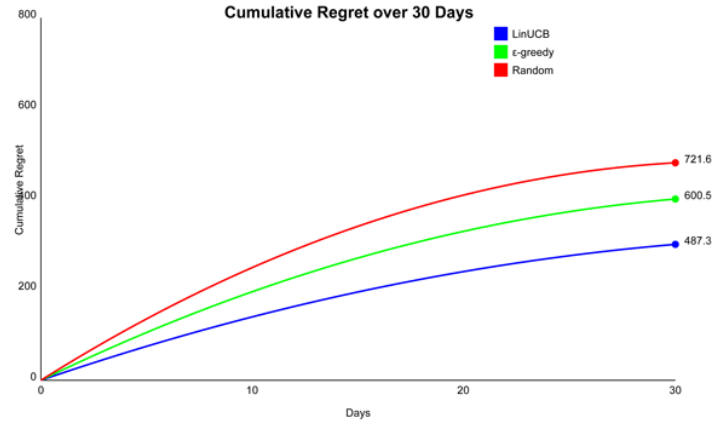


Figure 7: Cumulative Regret over 30 Days

3.1.3. Recommendation Accuracy and Diversity

To assess recommendation diversity, we calculated the Intra-List Distance (ILD) metric, which measures the dissimilarity between recommended items. The average ILD increased from 0.58 on day 1 to 0.73 on day 30, suggesting that the algorithm provided a more diverse range of recommendations as it learned user profiles.

3.1.4. Exploration vs. Exploitation Balance

We monitored the exploration-exploitation trade-off using the percentage of exploratory actions taken by the algorithm. Figure 8 shows how this percentage changed over time.

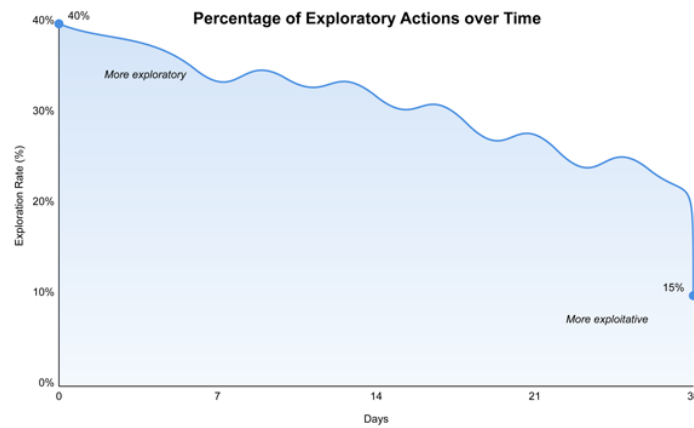


Figure 8: Percentage of Exploratory Actions over Time

The exploration rate decreased from an initial 40% to approximately 15% by day 30, indicating that the algorithm transitioned from a more exploratory phase to a more exploitative one as it gained confidence in its learned preferences.

3.1.5. Computational Efficiency

The average time to generate a recommendation was 12.3 milliseconds (ms) with a standard deviation of 2.1 ms, measured on a system with an AMD 5900X and 32GB RAM. This performance suggests that the algorithm is suitable for real-time recommendations in a mobile application setting. We plan to

transfer this to our mobile app, which might be slightly slower in its recommendations but still be fast enough for acceptable UX.

In summary, our MAB algorithm demonstrated effective learning, personalization, and adaptation capabilities in providing exercise recommendations. The results show improvements in recommendation accuracy and diversity over time, with successful contextual adaptation across different user segments. The sub-linear cumulative regret and efficient computational performance further support the viability of this approach for personalized exercise recommendation systems.

3.2. Results from the Explainable AI Layer

The integration of an explainable AI (XAI) layer into our exercise recommendation system yielded significant improvements in transparency, user understanding, and overall system effectiveness. We evaluated the XAI layer’s performance using only quantitative metrics, since our mobile app trial will be in the future.

3.2.1. Transparency and Interpretability

We analyzed the SHAP values for a sample of 2000 synthetic user profiles to understand the relative importance of different features in generating exercise recommendations from the MAB trial. Table 1 shows the average absolute SHAP values for the top 10 features:

Feature	Cardio	Strength	Flexibility
Age	0.3215	0.2987	0.2328
Exercise Frequency	0.2843	0.2765	0.2345
Exercise Duration	0.2567	0.2612	0.2132
Medical History	0.2456	0.2534	0.1946
Preferred Exercise	0.2234	0.2176	0.1905
Weekly Goal	0.2012	0.1987	0.1629
Cardio Preference	0.2345	0.1234	0.1383
Strength Preference	0.1234	0.2345	0.1192
Sex	0.1456	0.1567	0.1273
Race	0.1234	0.1345	0.1282
Start Preference	0.1123	0.1234	0.1456
Exercise Variety	0.1345	0.1456	0.1678
Exercise Location	0.1234	0.1345	0.1567
Exercise Buddy	0.1012	0.1123	0.1345
Coach Appearance	0.0901	0.0987	0.1123

Table 1
Average SHAP Values for Features Across MAB Recommendation Arms

Age, Exercise Frequency, and Exercise Duration are shown as the most important features across all three arms. Medical History and Preferred Exercise also have high SHAP values, indicating their significance in personalizing recommendations. Cardio Preference has a higher SHAP value for the Cardio arm, while Strength Preference has a higher value for the Strength arm, as would be expected. Some features, like Race and Coach Appearance, have lower SHAP values, suggesting they have less influence on the recommendations. The relative importance of features varies across the three arms, reflecting how different factors may be more or less relevant for different types of exercise recommendations.

3.2.2. Recommendation Consistency and Fairness assessment

To assess the consistency of recommendations across similar user profiles, we calculated the Jaccard similarity index for recommendations made to users with similar characteristics. For users with matching Age (± 5 years), Sex, and Medical History, the average Jaccard similarity of recommendations was 0.73, indicating a high degree of consistency while still allowing for personalization.

To ensure the model wasn't biased against particular demographic groups, we conducted a fairness assessment using the equal opportunity difference (EOD) metric [22]. The EOD values for key demographic features were:

1. Sex: 0.05
2. Race: 0.07
3. Age Groups (18-35, 36-55, 56+): 0.06

These values suggest relatively low levels of demographic bias in the recommendations, though there is still room for improvement. The addition of the SHAP-based XAI layer increased the average recommendation generation time from 12.3 ms to 89.7 ms (± 5.2 ms). This increase in latency is considered acceptable given the valuable insights provided by the explanations.

3.2.3. Example explanation

Here is a 62-year-old male user with hypertension, preferring strength training 7 times a week for 90-105 minutes.

Recommendations provided to user: A mix of moderate-intensity strength training and low-impact cardio exercises.

XAI rationale visible to user: "Exercise recommended for you, considering your age of 62 years and medical history of hypertension, where staying active with combination of cardio and strength training for 30-45 minutes a day, 4-5 days a week, is ideal. This follows the exercise guidelines to manage hypertension, keeping you healthy and strong."

The top 3 features influencing this recommendation by recommender were:

1. Age (SHAP value: +0.42): Increased the likelihood of recommending low-impact exercises
2. Medical History: Hypertension (SHAP value: -0.38): Decreased the intensity of recommended strength training
3. Exercise Frequency (SHAP value: +0.35): Increased the variety of recommended exercises

4. Future work

While these quantitative results provide valuable insights into the functioning of our Recommender System and the XAI layer, future work will include qualitative studies to evaluate user understanding and satisfaction with the explanations provided. We plan to conduct:

1. Semi-structured interviews with a diverse group of users to gather in-depth feedback on the clarity and usefulness of the explanations.
2. A longitudinal study to assess how the presence of explanations affects user adherence to recommended exercise routines over time
3. A comparative study between different explanation formats (e.g., natural language vs. visual representations) to determine the most effective way to communicate the reasoning behind recommendations.

Future iterations of this system will incorporate additional factors such as exercise intensity and explore more sophisticated feature interactions, further enhancing the personalization and effectiveness of the recommendations. Planned qualitative studies, including semi-structured interviews and longitudinal assessments, will provide crucial insights into user understanding, satisfaction, and long-term adherence to recommended exercise routines.

As we move forward with testing the system on both healthy individuals and those with cardiovascular conditions, we aim to validate its effectiveness in real-world scenarios.

5. Conclusion

This study introduces a novel recommendation system designed to address key challenges in personalized exercise interventions. By using DQN with MAB algorithm, we solve the user cold start problem [6, 4, 5] and improve the interpretability of deep learning models, ensuring exercise recommendations are personalized and aligned with medical advice.

The implementation of this recommender within a mobile application can not only promote regular physical activity but can also help users build lasting exercise habits in a mobile-driven world. With an average recommendation generation time including the XAI layer being 89.7 ms (\pm 5.2 ms), our system is both computationally efficient and practical for real-time use.

Our approach represents a significant step forward in digital health, combining advanced ML with XAI layer to promote guideline-based physical activity. By tackling the cold start problem and enhancing algorithm transparency, this system has the potential to foster lasting behavior change and improve public health.

6. Disclosure

Parvati Naliyattthaliyazchayil hereby discloses that she has volunteered at Indiana University and is currently employed by ConcertAI. This disclosure applies solely to Parvati Naliyattthaliyazchayil and does not extend to any of the other authors of this paper.

References

- [1] NHS, Benefits of exercise, 2024. URL: <https://www.nhs.uk/live-well/exercise/exercise-health-benefits/>, accessed: 2024-08-12.
- [2] WHO, Physical activity, 2024. URL: <https://www.who.int/news-room/fact-sheets/detail/physical-activity>, accessed: 2024-08-12.
- [3] U. Bhimavarapu, M. Sreedevi, N. Chintalapudi, G. Battineni, Physical activity recommendation system based on deep learning to prevent respiratory diseases, *Computers* 11 (2022) 150. doi:10.3390/computers11100150.
- [4] A. Panteli, B. Boutsinas, Addressing the cold-start problem in recommender systems based on frequent patterns, *Algorithms* 16 (2023) 182. doi:10.3390/a16040182.
- [5] Appier, 7 critical challenges of recommendation engines, <https://www.appier.com/en/blog/7-critical-challenges-of-recommendation-engines>, 2024. Accessed: 2024-08-12.
- [6] V. Hassija, V. Chamola, A. Mahapatra, et al., Interpreting black-box models: A review on explainable artificial intelligence, *Cognitive Computation* 16 (2024) 45–74. URL: <https://doi.org/10.1007/s12559-023-10179-8>. doi:10.1007/s12559-023-10179-8, published: 24 August 2023, Issue Date: January 2024.
- [7] B. Lika, K. Kolomvatsos, S. Hadjiefthymiades, Facing the cold start problem in recommender systems, *Expert Systems with Applications* 41 (2014) 2065–2073. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0957417413007240>. doi:10.1016/j.eswa.2013.09.005.
- [8] H. Yuan, A. A. Hernandez, User cold start problem in recommendation systems: A systematic review, *IEEE Access* (2023). URL: <https://doi.org/10.1109/ACCESS.2023.3338705>. doi:10.1109/ACCESS.2023.3338705, license: CC BY-NC-ND 4.0.
- [9] S. A., S. R., A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends, *Decision Analytics Journal* 7 (2023) 100230. URL: <https://www.sciencedirect.com/science/article/pii/S277266222300070X>. doi:10.1016/j.dajour.2023.100230.
- [10] O. Awwad, A. Akour, S. Al-Muhaissen, D. E. Morisky, The influence of patients' knowledge on adherence to their chronic medications: A cross-sectional study in jordan, *International*

- Journal of Clinical Pharmacy 37 (2015) 504–510. URL: <https://doi.org/10.1007/s11096-015-0086-3>. doi:10.1007/s11096-015-0086-3, epub 2015 Feb 24.
- [11] F. Folkvord, A.-R. U. Würth, K. van Houten, et al., A systematic review on experimental studies about patient adherence to treatment, *Pharmacology Research Perspectives* 12 (2024) e1166. doi:10.1002/prp2.1166.
 - [12] A. H. Krist, S. T. Tong, R. A. Aycock, D. R. Longo, Engaging patients in decision-making and behavior change to promote prevention, *Studies in Health Technology and Informatics* 240 (2017) 284–302. doi:10.3233/ISU-1708.
 - [13] American Heart Association, Aha recommendations for physical activity in adults, 2024. URL: <https://www.heart.org/en/healthy-living/fitness/fitness-basics/aha-recs-for-physical-activity-in-adults>, accessed: August 2024.
 - [14] National Kidney Foundation, Stay fit, 2024. URL: <https://www.kidney.org/atoz/content/stayfit>, accessed: August 2024.
 - [15] American Diabetes Association, Weekly exercise targets, 2024. URL: <https://diabetes.org/health-wellness/fitness/weekly-exercise-targets>, accessed: August 2024.
 - [16] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, et al., Human-level control through deep reinforcement learning, *Nature* 518 (2015) 529–533. URL: <https://www.nature.com/articles/nature14236>. doi:10.1038/nature14236.
 - [17] X. Chen, L. Yao, J. McAuley, G. Zhou, X. Wang, Deep reinforcement learning in recommender systems: A survey and new perspectives, *Knowledge-Based Systems* 264 (2023) 110335. URL: <https://www.sciencedirect.com/science/article/pii/S0950705123000850>. doi:10.1016/j.knosys.2023.110335.
 - [18] K.-H. Huang, H.-T. Lin, Linear upper confidence bound algorithm for contextual bandit problem with piled rewards, in: *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II* 20, Springer, 2016, pp. 143–155. doi:10.1007/978-3-319-31750-2_12.
 - [19] D. Bouneffouf, S. Upadhyay, Y. Khazaeni, Linear upper confident bound with missing reward: Online learning with less data, in: *2022 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2022, pp. 1–6. doi:10.1109/IJCNN55064.2022.9892856.
 - [20] A. Slivkins, Introduction to multi-armed bandits, *Foundations and Trends® in Machine Learning* 17 (2024) 1–143. URL: <https://arxiv.org/abs/1904.07272>. doi:10.1561/22000000068, first draft: January 2017; Published: November 2019; Latest version: April 2024.
 - [21] Z. Sadeghi, R. Alizadehsani, M. A. CIFCI, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladik, S. Nahavandi, P. M. Pardalos, A review of explainable artificial intelligence in healthcare, *Computers and Electrical Engineering* 118 (2024) 109370. URL: <https://www.sciencedirect.com/science/article/pii/S0045790624002982>. doi:10.1016/j.compeleceng.2024.109370.
 - [22] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, *Advances in neural information processing systems* 29 (2016). doi:10.48550/arXiv.1610.02413.

A. Online Resources

The code for this project is available on GitHub at the following repository: <https://github.com/iupui-soic/exercise-behavior-change-app/tree/recommendersys>

Health Document Presentation in Patient-Centered Recommender Systems with Carousel Interfaces

Behnam Rahdari¹, Peter Brusilovsky¹, Daqing He¹, Khushboo Thaker¹,
Mohammad Hassany¹, Youjia Wang², Young Ji Lee² and Heidi Donovan²

¹*School of Computing and Information, University of Pittsburgh, Pittsburgh, USA*

²*School of Nursing, University of Pittsburgh, Pittsburgh, USA*

Abstract

Despite the increasing availability of health information, many users still find it difficult to navigate and comprehend this content effectively. Addressing these challenges requires innovative approaches, including personalized recommendations and more efficient methods of information delivery. In this paper, we explore the use of generative AI to improve access to health article recommendations within a carousel-based interface, utilizing our system, HELPeR. Our focus is on both generating and evaluating these summaries through a three-stage online experiment with domain experts. The results reveal the potential and complexities of employing generative AI for summarizing recommended health articles for ovarian cancer patients and their caregivers.

Keywords

Health Recommender Systems, Generative AI, Personalized Health Information, User Interface Design

1. Introduction

The increased availability of health information online has transformed the way patients and caregivers access knowledge about diseases, treatments, and health management. However, navigating this vast amount of information can be overwhelming, especially for non-experts. The complexity of medical terminology, coupled with the sheer volume of available content, often leaves users struggling to find and understand the information most relevant to their needs. This challenge underscores the need for more intuitive methods of information delivery that cater to varying levels of information needs and health literacy.

In response to these challenges, personalized recommendations and advanced methods of information delivery have emerged as key strategies to bridge the gap between users and the information they seek. These approaches aim to adapt the content to individual users, considering factors such as their treatment history, disease trajectory, and cognitive abilities. By offering personalized recommendations, these systems can guide users through relevant information more efficiently, potentially improving their understanding and engagement with health content.

Despite the progress made in this area, important questions remain unanswered. One of the key issues is understanding how users decide whether to explore a recommendation further, especially when only minimal information is provided. While visual cues such as images are effective in capturing users' attention in domains like entertainment and e-commerce, the health domain primarily relies on textual content, with images often serving a decorative role in most online health articles. This raises the question of how to present health information in a way that encourages deeper exploration without the visual allure that is typical in other fields.

This challenge became particularly evident in our work with the HELPeR system [1] depicted in Figure 1, which used a carousel-based recommendation interface to present relevant health documents to cancer patients. The use of a carousel-based interface was important in our recommendation context, since the system has to data to decide which information need brought the user to HELPeR for each particular session. While the system maintains a fine-grain model of user interests and knowledge, the

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

✉ ber58@pitt.edu (B. Rahdari); peterb@pitt.edu (P. Brusilovsky); dah44@pitt.edu (D. He); k.thaker@pitt.edu (K. Thaker); moh70@pitt.edu (M. Hassany); yow14@pitt.edu (Y. Wang); leeyoung@pitt.edu (Y. J. Lee); donovanh@pitt.edu (H. Donovan)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

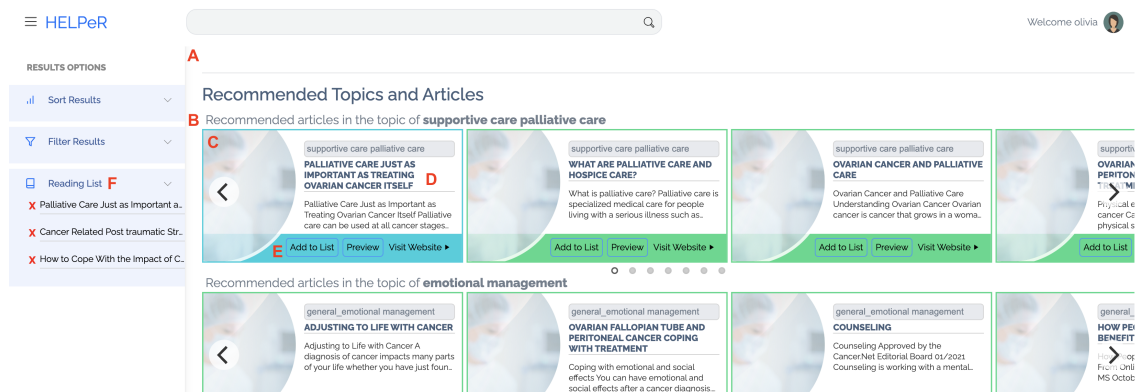


Figure 1: The HELPeR system interface design includes the following main components: A: Recommended Articles, B: A carousel containing recommended articles within the same topic, C: A specific recommended article, D: Details of the selected item, E: Options for user interaction with the recommendation, and F: Reading list (Personal Library)

majority of the users maintain their interests in several health topics, and choosing user chief concern reliably at each session is as hard as guessing which movie genre a user would like to watch on Netflix at each login. Just like Netflix uses a carousel-based interface to support human-AI collaboration and allow the user to choose the genre she preferred today, HELPeR's carousel-based interface allows the user to choose the priority topics from several most likely topics shown by the carousels.

The carousel format is familiar and effective in presenting multiple options in parallel, but in the context of health information, it presented unique difficulties. While carousels are useful for organizing and displaying visual content (such as movie posters or book covers), it is a challenge to choose textual content to present on a carousel card to ensure that this information is sufficient for the user to make an informed decision. This challenge was compounded by the fact that the recommended content was often not available in a cohesive, summarized format. The headers of health documents are usually too long to be readable or even fit on a card (Figure 1). Document summaries are either too complex or not available at all. To address this challenge, we turned to generative AI to create concise document overviews and summaries that could effectively fit within the carousel cards, helping users make informed decisions with the limited available space.

In this paper, we present our attempts to use Large Language Models (LLM) to generate brief but informative overviews and summaries of health articles related to ovarian cancer, which can fit to the cards of HELPeR's carousel-based recommendation interface. To assess the relevance, clarity, and informativeness of these AI-generated overviews and summaries, we performed a three-phase expert evaluation study. Our findings highlight the challenges of generating concise yet informative content that meets the needs of diverse users, offering valuable insights into the future of AI-driven health information delivery.

Our study contributes to this evolving landscape by evaluating the use of AI-generated summaries within a patient-centered health recommender system, specifically designed for ovarian cancer patients. Although previous research has established the potential of AI in personalizing health information, our focus is on understanding how these AI-generated summaries can be integrated into carousel interfaces to enhance user engagement and decision-making. Through a three-phase expert evaluation, we aim to provide insights that will inform the design of future recommender systems, making them more responsive to the diverse and changing needs of patients.

The paper is organized as follows. In the next section, we provide a review of related work in personalized health information systems. The methodology section explains the approach to generating summaries and a balanced document selection approach that we applied to minimize bias in our evaluation process. We then describe the evaluation study and present our findings. Finally, we discuss the implications of our results, outline the limitations of our study, and suggest directions for future work.

2. Related Work

The increasing availability of health information online has greatly influenced how patients and caregivers manage care and make informed decisions. Early recommender systems in this domain focused on providing personalized health information by tailoring content to often static user profiles. These systems reported in [2, 3] were important in demonstrating the benefits of personalized information delivery. However, as the complexity and volume of health information increased, so did the need for systems that could adapt to the evolving needs of users more effectively.

In response to these needs, the field has seen a shift towards more interactive and dynamic recommender systems. These systems known as conversational and critique-based recommenders [4, 5] allow users to actively engage with the recommendations, refining the information needs based on real-time feedback. Although these interactive models enhanced personalization, they also introduced new challenges, such as the cognitive load associated with continuous interaction, which could be particularly burdensome for users with lower health literacy.

To address these challenges, recent research explored visual interfaces for recommender systems [6], which offered a higher expressive power and a better opportunity to add transparency [7] and user control [8] to the recommendation process. Among other visual recommender interfaces, carousel-based interfaces became especially popular for their ability to display multiple pieces of content in a compact and navigable format [9, 10] and offer users a simple control over the recommendation process. These interfaces combined power and simplicity enabling users to quickly scan through recommendations and make informed decisions without feeling overwhelmed.

Despite these advancements, the challenge of effectively delivering personalized health information remains unsolved. As noted by Chi et al. [11] and Thaker et al. [12], users' information needs are not static; they could change frequently following the change in health status, the progress of treatment and personal circumstances. This variability requires a more dynamic approach to content presentation, one that can adapt to the immediate context of the user. The integration of generative AI into health recommender systems, as explored in our work, offers a promising solution by creating concise, contextually relevant summaries that fit within the limited space of carousel interfaces [1].

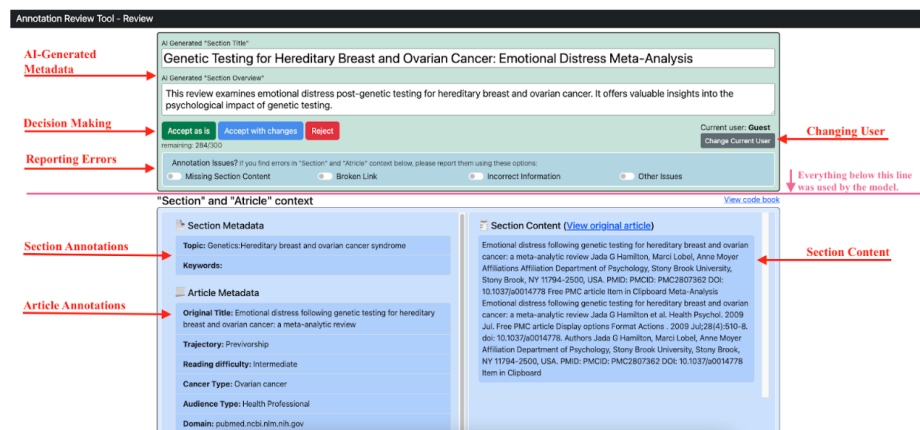


Figure 2: Evaluation Interface for AI-Generated Metadata

3. Method

3.1. Generating the title and overview

The HELPeR system [1, 13], is an interactive recommender system designed for ovarian cancer patients and their caregivers. It uses a knowledge base built from a curated collection of documents, which includes public health information, research articles, clinical trial results, and other relevant resources.

These documents undergo a rigorous curation process that involves sectioning, topic modeling, and key-phrase extraction to ensure that the information is reliable, up-to-date, and relevant to the needs of the users. For each document, HELPeR’s knowledge base includes a range of textual and metadata information, such as the article title, topics, difficulty, relevant key-phrase, and full text of each document section.

As mentioned above, neither document titles nor section content were immediately suitable to be displayed on carousel card. The titles were frequently long and confusing and summaries either very long or not available. To address this problem, the most recent version of our system explored the use of LLM¹ to generate a document representation (title and summary) that can fit on a carousel card providing concise information about the document behind the card.

The exact prompt used in our study to generate a title and summary for each document is shown below. This prompt was selected through a prompt-tuning process to maximize clarity and relevance within the constraints of a limited format. Note that we passed all information about each document to LLM as part of the prompt:

```
prompt = f""" Given the following section of an
article titled '{article_title}', with the
topic of {topic}, covering keywords: {keywords}
and containing this text: "{section_text}"
```

```
1- Generate a 4-7 word title reflecting the
section’s essence and aligning with the
article’s theme.
```

```
2- Write a 20-25 words, two-sentence summary
capturing key points, serving as an informative
overview of the article for readers.
```

```
Respond ONLY and PRECISELY in this format:
[{{ "title": "the generated title" }},
{{"summary": "the generated summary" }}]"""
```

3.2. Selecting a Diverse Subset of Annotations

Given the large number of documents in our collection related to ovarian cancer, selecting a manageable yet representative subset for our human-centered evaluation was essential. A subset that accurately represents the diversity of the full dataset is crucial to avoid biases that could skew our results, particularly in our collection, that certain topics or audience types are over-represented. To address this, we used a modified version of the Maximal Marginal Relevance (MMR) algorithm [14], traditionally used to balance relevance and novelty in information retrieval tasks.

The primary challenge in selecting this subset was to ensure that it not only reflected the diversity of topics, domains, and audience types present in the dataset, but also maintained the overall distribution of the documents and topics in our collection. We adapted the MMR algorithm to emphasize the intrinsic properties of the document, allowing it to select a subset that balanced these two critical factors.

Given a set of all documents D , where $n = |D|$ denotes the total number of documents, we represented the document set as a matrix X , with each document encoded as a one-hot vector across m categories. The subset S , initially empty, was populated by the execution of the MMR algorithm. The key parameter λ , which ranged from 0 to 1, modulated the trade-off between relevance and diversity.

The process began with the random selection of an initial document d_q from D , which served as a pseudo-query, establishing the initial subset $S = \{d_q\}$. Then each document d_i within D was assigned a relevance score, $sim(d_i)$, assumed for simplicity to follow a uniform distribution $sim(d_i) \sim U(0, 1)$.

¹OpenAI APIs - gpt-3.5-turbo-1106

During each iteration, the algorithm evaluated each candidate document d_c not yet included in S . The diversity score $div(d_c, S)$ was calculated based on the average cosine dissimilarity between d_c and all documents currently in S :

$$div(d_c, S) = 1 - \frac{1}{|S|} \sum_{d_s \in S} \text{cosine_similarity}(\mathbf{x}_{d_c}, \mathbf{x}_{d_s})$$

The MMR score for each candidate document was then calculated by combining the relevance and diversity scores:

$$MMR(d_c) = \lambda \cdot sim(d_c) + (1 - \lambda) \cdot div(d_c, S)$$

The document d_c with the highest MMR score was added to S , thereby progressively refining the document subset to be both *relevant* and *diverse*.

Our modified approach departs from the traditional reliance on a fixed query in the MMR algorithm by dynamically calculating diversity relative to the evolving subset S . This modification is particularly advantageous in environments where queries are undefined or fluid, such as unsupervised document clustering or information retrieval systems where query independence is crucial. By structuring the selection process around these principles, we ensured that the resulting subset was both representative of the dataset's diversity and suitable for our human-centered evaluation of the AI-generated summaries.

As illustrated in Figure 3, the subset selected using the Maximal Marginal Relevance (MMR) algorithm shows a more balanced distribution of features across various categories, such as article knowledge level, audience type, and domain, compared to a randomly selected subset. The MMR-based selection achieves a broader coverage of diverse topics, ensuring that underrepresented areas are included, thus mitigating the biases that are evident in the random sample where certain categories, such as "Survivorship" and "Patient and Caregiver," dominate.

4. Experiment

4.1. Experimental Design

To evaluate the quality and consistency of AI-generated *titles and summaries* for documents in HELPeR knowledge base we designed a multi-phase study. We alternate the articles and their order in which they are presented in each phase. This study design allows us to gain insights about both the quality of AI-generated content and the evaluation process itself.

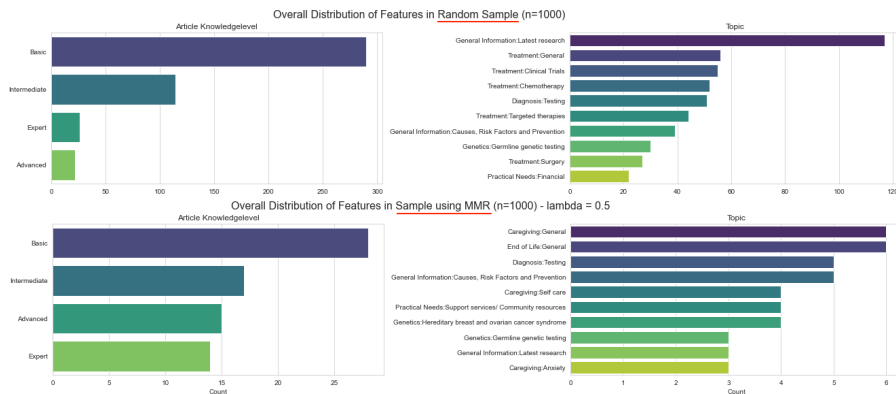


Figure 3: Comparison of Overall Distribution of Features in the Random Sample (top) vs. MMR-Selected Sample (bottom) - $n = 1000$, $\lambda = 0.5$ - Color intensity corresponds to feature frequency.

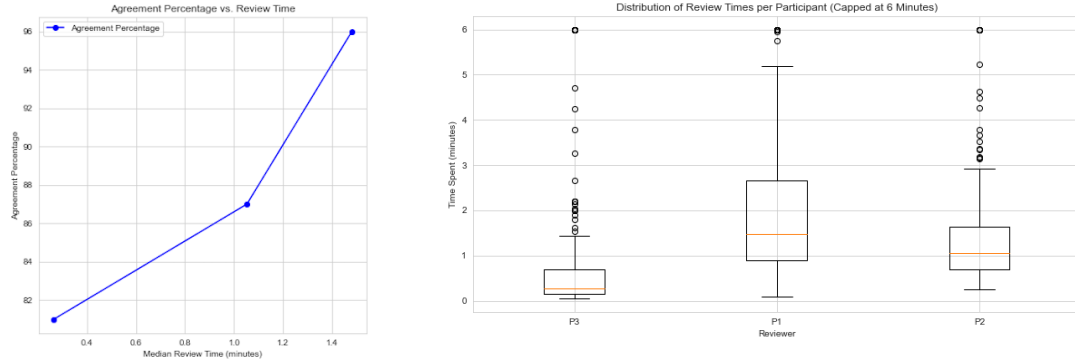


Figure 4: Left: Agreement percentage goes up with more time spend on reviewing. Right: Time spent by each reviewer.

4.2. Participants

The study involved two groups of participants: domain experts and nursing students with specialized knowledge in ovarian cancer. In the first phase, we have 100 unique documents to each of three domain experts in our research team. In the next phases, three nursing students (not belonging to the research team) were recruited to review the same sample in two rounds. Each student reviewed 300 instances, with 100 documents that overlapped between the two rounds. The use of nursing students in addition to domain experts was important to better represent the prospect of caregivers and reduce evaluation bias.

4.3. Procedure

4.3.1. Phase 1

In the first phase, AI-generated titles and overviews were created for a sub-sample of 300 documents selected by the modified MMR algorithm. Then these summaries were presented to three domain experts via a custom web interface (Figure 2). Each expert was tasked with evaluating 100 documents. To our surprise, only 55% of the summaries were accepted without modification, a number that raises many questions about the abilities of Generative-AI in summarizing health articles.

4.3.2. Phase 2

Following the insights gained from the first phase, we conducted unstructured interviews with domain experts on our team to understand the reason behind their high rejection rate. As it turns out, the reason for rejecting in many cases was related to a problem with missing or inaccurate annotations and uncertainty about the decision-making criteria. To address these issues, several enhancements were made in the second phase. A detailed code-book was developed to provide clear guidelines and decision-making criteria for participants, ensuring consistency across evaluations. Additionally, the web interface was improved to include an interactive tutorial and more user-friendly features including the ability to report inconsistent annotations and other issues that are not directly related to the quality of AI-Generated contents. In this phase, three external nursing students, each with specialized knowledge in ovarian cancer, were recruited to evaluate the same 300 documents. Each student reviewed 100 documents in this phase.

4.3.3. Phase 3

In the final phase, the consistency of the evaluations was tested. The external nursing students were asked to review an additional set of 200 documents. This set included 100 new documents and 100 documents that they had reviewed in the previous round. These documents were randomly distributed.

This phase aimed to measure the stability and reliability of their evaluations over time. At the end of this phase, each document had been reviewed by multiple experts, providing a comprehensive dataset for our analysis.

5. Results

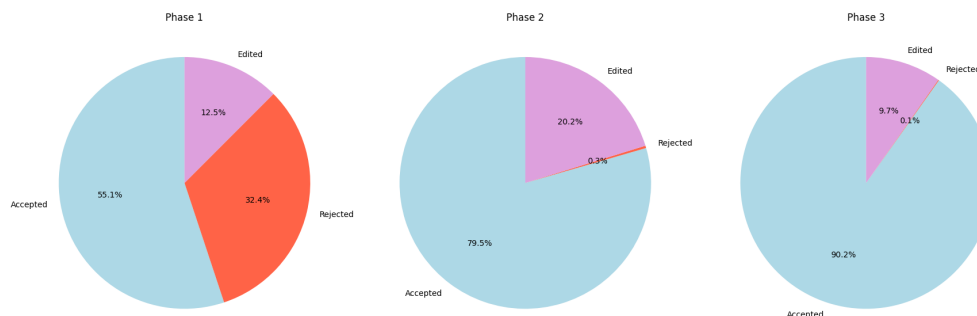


Figure 5: Document quality ratings before and after improvements (left to right)

As explained above, the initial evaluation phase, conducted by domain experts, indicated that only 55% of the AI-generated summaries were accepted without modifications. This lower-than-expected acceptance rate highlighted several issues in the evaluation process, which we attempted to address by introducing several enhancements reviewed above. Following these enhancements, the second phase of our study, in which nursing students with specialized knowledge in ovarian cancer were recruited as evaluators, showed a significant change, with 98.7% of the content being accepted either outright or with minor edits. Specifically, 78.7% of the summaries were accepted as they were, while 20% required slight modifications. Only a small fraction, 1.3%, of the content was rejected. These results demonstrated that the enhancements implemented after the first phase, such as the code-book and UI improvements, had a substantial positive impact on the perceived quality of the AI-generated summaries. Notably, our results suggest that the main factor contributing to this improvement was the added ability for users to report issues with the generated content separately from judging its overall quality. (Figure 5)

The final phase of the study focused on assessing the quality of evaluation itself by analyzing the consistency between our raters. The results from this ,depicted in Table 1,phase showed an 88% agreement in the reviews, indicating the reliability of the evaluation process. However, Cohen's Kappa scores of 0.37 suggested moderate to low consistency among reviewers, pointing to individual differences in judgment and the influence of subjective factors on the evaluation process. We are aware that our small sample size could have major influence of this results. Additionally, pairwise agreement percentages among reviewers varied, with the highest consistency observed between reviewers P2 and P3 at 92%.

To better understand the individual differences between our reviewers, we analyzed the ratio between the duration of the review session and the consistency with peers. Our analysis, illustrated in Figure 3-left, revealed a positive correlation between longer review times and higher consistency in evaluations as can be seen in Figure 4. This suggests that a more deliberate review process leads to more stable and reliable judgments. Reviewers who spent less time on evaluations exhibited lower consistency, emphasizing the importance of thoroughness in assessing AI-generated content.

Overall, the results of this study, although preliminary in nature, underscore the challenges inherent in evaluating AI-generated summaries, particularly in a complex domain like ovarian cancer information. The improvements made between phases significantly improved reviewer's judgment of acceptability of the AI outputs, yet individual differences among reviewers remain a critical factor in the consistency of evaluations. These findings highlight the need for clear guidelines and rigorous training to ensure reliable and consistent assessments in future AI evaluation tasks.

Table 1
Inter-Rating Consistency

Metric	Value
Total reviews analyzed	300
Unchanged reviews	264 (88.00%)
Changed reviews	36 (12.00%)
Changes frequency	
Edited → Accepted	32
Accepted → Edited	4
Inter-rater consistency	
Total sections	100
Some agreement	99%
No agreement	1% (actual AI mistake)
Full agreement	63 (63.00%)
Partial agreement	36 (36.00%)

6. Discussion, Limitations, and Future Work

The results of this study highlight both the potential and the challenges associated with using AI-generated summaries in the domain of ovarian cancer information. The improvements observed in the second phase, following the implementation of a comprehensive code-book² and user interface enhancements, underscore the importance of clear guidelines and a user-friendly evaluation environment. These adjustments significantly increased the acceptability of the AI outputs, demonstrating that careful attention to the evaluation process can substantially enhance the performance of AI systems in generating accurate and relevant content.

However, the study also revealed certain limitations. One of the primary concerns is the variability in reviewer judgments, as indicated by the moderate to low Cohen's Kappa scores. This suggests that individual differences in interpretation and the inherent subjectivity in content evaluation can impact the consistency of the results. Additionally, the correlation between review time and consistency points to the importance of thorough evaluations, but it also raises questions about the feasibility of scaling such processes for larger datasets.

Another limitation is related to the specific domain of ovarian cancer, which, while critical, may not encompass the full spectrum of challenges that AI-generated content could face in other medical or health-related domains. Therefore, while the findings of this study are valuable, they may not be fully generalizable to other areas without further validation.

For future work, several avenues can be explored. First, expanding the evaluation to include a more diverse set of medical topics could help to understand how AI performs in different domains. Furthermore, exploring automated methods to assess the consistency of evaluations, while keeping the human in the loop, could reduce the reliance on manual review processes, making the evaluation of AI-generated content more scalable.

Finally, addressing the subjective nature of content evaluation by developing more objective criteria or incorporating a larger pool of evaluators could improve the robustness of the evaluation process. Future studies could also examine the impact of these AI-generated summaries on end-users, such as patients and caregivers.

Acknowledgments

This work was supported by awards from the National Library of Medicine (NLM) of the National Institutes of Health (NIH) (Award Number: R01-LM013038).

²<https://bit.ly/helper-codebook>

References

- [1] B. Rahdari, P. Brusilovsky, D. He, K. M. Thaker, Z. Luo, Y. J. Lee, Helper: An interactive recommender system for ovarian cancer patients and caregivers, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 644–647.
- [2] K. Lee, K. Hoti, J. D. Hughes, L. M. Emmerton, Consumer use of “dr google”: a survey on health information-seeking behaviors and navigational needs, *Journal of medical Internet research* 17 (2015) e4345.
- [3] S. Kanthawala, A. Vermeesch, B. Given, J. Huh, Answers to health questions: Internet search results versus online health community responses, *Journal of medical Internet research* 18 (2016) e5369.
- [4] B. Smyth, L. McGinty, J. Reilly, K. McCarthy, Compound critiques for conversational recommender systems, in: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, WI '04, IEEE Computer Society, USA, 2004, p. 145–151.
- [5] L. Chen, P. Pu, Critiquing-based recommenders: survey and emerging trends, *User Modeling and User-Adapted Interaction* 22 (2012) 125–150.
- [6] C. He, D. Parra, K. Verbert, Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities, *Expert Systems with Applications* 56 (2016) 9–27.
- [7] K. Verbert, D. Parra-Santander, P. Brusilovsky, E. Duval, Visualizing recommendations to support exploration, transparency and controllability, in: the 2013 International Conference on Intelligent User Interfaces, IUI '2013, ACM Press, 2013, pp. 351–362.
- [8] D. Jannach, S. Naveed, M. Jugovac, User control in recommender systems: Overview and interaction challenges, in: E-Commerce and Web Technologies: 17th International Conference, EC-Web 2016, Porto, Portugal, September 5-8, 2016, Revised Selected Papers 17, Springer, 2017, pp. 21–33.
- [9] B. Rahdari, B. Kveton, P. Brusilovsky, The magic of carousels: Single vs. multi-list recommender systems, in: Proceedings of the 33rd ACM Conference on Hypertext and Social Media, 2022, pp. 166–174.
- [10] N. Felicioni, M. Ferrari Dacrema, P. Cremonesi, A methodology for the offline evaluation of recommender systems in a user interface with multiple carousels, in: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, 2021, pp. 10–15.
- [11] Y. Chi, D. He, W. Jeng, Laypeople’s source selection in online health information-seeking process, *Journal of the Association for Information Science and Technology* 71 (2020) 1484–1499.
- [12] K. Thaker, Y. Chi, S. Birkhoff, D. He, H. Donovan, L. Rosenblum, P. Brusilovsky, V. Hui, Y. J. Lee, Exploring resource-sharing behaviors for finding relevant health resources: analysis of an online ovarian cancer community, *JMIR cancer* 8 (2022) e33110.
- [13] K. Thaker, B. Rahdari, V. Hui, Z. Luo, Y. Wang, P. Brusilovsky, D. He, H. Donovan, Y. J. Lee, Helper: Interface design decision and evaluation, in: *Innovation in Applied Nursing Informatics*, IOS Press, 2024, pp. 750–751.
- [14] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 335–336.

Design and Assessment of Representative Hybrid Clinical Trials using Health Recommender System

Nafis Neehal^{1,*}, Vibha Anand² and Kristin P. Bennett³

¹Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, 12180 USA

²Biomedical AI-Healthcare and Life Sciences, IBM T.J. Watson Research Center, Cambridge, MA, 02141 USA

³The Institute of Data Exploration and Applications and the Mathematical Sciences and Computer Science Departments, Rensselaer Polytechnic Institute, Troy, NY, 12180 USA

Abstract

Incorporating real-world data (RWD) into clinical trials can enhance trial efficiency, diversity, and generalizability. This paper introduces the Framework for Research in Synthetic Control Arms (FRESKA), which uses a novel Recommender System combined with Equity Adjustment strategies to design and evaluate Representative Hybrid Clinical Trials (HCTs). FRESKA employs a novel matching algorithm through its recommendation system to select suitable patients from RWD while ensuring that the selected population is representative of the target demographic. This dual approach improves both patient selection and trial outcomes by balancing statistical appropriateness and equity. Simulations based on data from two existing randomized clinical trials (RCTs) show that using FRESKA to recommend patients from RWD and apply equity adjustments enhances internal validity and generalizability. Our analysis indicates that combining matching and equity adjustments yields more accurate treatment effect estimates and fair population representation, even with reduced RCT control group sizes. In contrast, using either method alone may result in biased outcomes. The flexibility of FRESKA to simulate various HCT scenarios makes it a valuable tool for advancing equitable and efficient clinical trial designs.

Keywords

Causal Inference, Equity, Hybrid Clinical Trials, Randomized Clinical Trial, Recommender Systems

1. Introduction

Enhancing the efficiency, diversity, and generalizability of clinical trials can be achieved by incorporating real-world data (RWD) [1]. This study introduces the Framework for Research in Synthetic Control Arms (FRESKA), which combines a novel Recommender System with strategies for Equity Adjustment to design and assess representative Hybrid Clinical Trials (HCTs). Synthetic control patients are patients created from pre-existing de-identified datasets, used to mimic the characteristics of a real control group in clinical trials. Synthetic control arms (SCAs) are especially useful in trials for rare diseases, where finding enough "in-trial" concurrent controls (CCs) can be difficult due to ethical and practical concerns [2] [3]. To address these challenges, HCTs use hybrid control arms (HCAs) that combine both concurrent and synthetic controls. FRESKA uses health recommender systems based on propensity score matching to recommend patients from external RWD who are suitable for inclusion in the trial, creating a hybrid population.

The health recommender system in FRESKA identifies patients from RWD who closely match the characteristics of those in the trial, enhancing the statistical power and reducing variance without extending the trial duration or increasing costs. However, integrating RWD with randomized control trial (RCT) data is challenging due to differences in their distributions [4]. To overcome this, FRESKA first uses its health recommender system to select appropriate patients and then applies equity adjustments to ensure the trial population accurately represents the target demographic. This combined approach

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

✉ neehan@rpi.edu (N. Neehal); anand@us.ibm.com (V. Anand); bennek@rpi.edu (K. P. Bennett)

🌐 <https://nafis-neehal.github.io/> (N. Neehal); <https://research.ibm.com/people/vibha-anand> (V. Anand);

<https://faculty.rpi.edu/kristin-bennett> (K. P. Bennett)

🆔 0000-0001-7015-8039 (N. Neehal); 0000-0001-8605-5712 (V. Anand); 0000-0002-8782-105X (K. P. Bennett)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

improves patient selection and ensures trial results are both statistically robust and demographically representative.

Our research aims to develop methods for conducting equitable HCTs and to provide a framework for evaluating them. In this context, equity means that the trial participants should represent a broader target population [5]. While RCTs can provide unbiased estimates for their specific cohorts, they often fail to represent larger, more diverse target populations. Researchers use data from these target populations to adjust RCT samples, ensuring that all relevant subgroups are included [6] [7]. Ensuring equity in hybrid trials is essential for generalizability and is a key focus for institutions like the NIH and FDA [8]. Our approach demonstrates that combining a health recommender system with equity adjustments creates a more balanced and representative trial population than using either method alone.

Table 1

Distribution of Protected Attributes in FRESKA Cohorts and Biased External Controls in ALLHAT and SPRINT trials along with NHANES Target Subgroup Rates

Attributes	ALLHAT			SPRINT			NHANES Target Rate
	TA (N=8116)	CC (N=4000)	Biased EC (N=9762)	TA (N=4234)	CC (N=2000)	Biased EC (N=2200)	
Age Group							
40-59	1549 (19.1%)	772 (19.3%)	5841 (59.8%)	923 (21.8%)	438 (21.9%)	334 (15.2%)	31.2%
59+	6567 (80.9%)	3228 (80.7%)	3921 (40.2%)	3311 (78.2%)	1562 (78.1%)	1866 (84.8%)	68.8%
Gender							
Female	3728 (45.9%)	1875 (46.9%)	2752 (28.2%)	1499 (35.4%)	691 (34.6%)	631 (28.7%)	55.4%
Male	4388 (54.1%)	2125 (53.1%)	7010 (71.8%)	2735 (64.6%)	1309 (65.5%)	1569 (71.3%)	44.6%
Race or Ethnicity							
Hispanic	1389 (17.1%)	740 (18.5%)	1671 (17.1%)	479 (11.3%)	225 (11.3%)	389 (17.7%)	10.0%
NH Asian	84 (1.0%)	41 (1.0%)	693 (7.1%)	42 (1.0%)	15 (0.8%)	138 (6.3%)	3.9%
NH Black	2525 (31.1%)	1265 (31.6%)	1937 (19.8%)	1232 (29.1%)	616 (30.8%)	468 (21.3%)	12.0%
NH White	4050 (49.9%)	1921 (48.0%)	5285 (54.1%)	2451 (57.9%)	1128 (56.4%)	1172 (53.3%)	69.3%
Other	68 (0.8%)	33 (0.8%)	176 (1.8%)	30 (0.7%)	16 (0.8%)	33 (1.5%)	4.8%

FRESKA was developed to design and evaluate these methods [9]. Before FRESKA, existing methods for HCTs did not explicitly address equity in patient selection. FRESKA uses real clinical trial data to simulate hypothetical trials, applying its methods to scenarios based on real RCTs, such as Systolic Blood Pressure Intervention Trial (SPRINT) [10] and Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) [11]. By integrating health recommender systems and equity adjustment techniques, FRESKA ensures that the selected patients not only meet statistical criteria but also represent the demographic characteristics of the target population. We define protected subgroups based on age, gender, and race/ethnicity, using NHANES data [12] to estimate the rates for these subgroups in a simulated target population. We will explore the use of additional protected attributes in future work. FRESKA includes five main functions: generating cohorts, simulating scenarios, calculating target subgroup rates, estimating treatment effects and equity, and providing a final assessment. It can evaluate any HCT method, including those that combine health recommender systems with equity adjustments. Detailed descriptions of the FRESKA framework and the trial configurations are provided in sections 2.4 and 2.5.

This paper makes several key contributions-

- It identifies equity issues in HCTs and proposes solutions to improve the representativeness of trial populations.
- It introduces an enhanced FRESKA framework with a modular architecture that supports multi-trial, multi-outcome, and multi-metric assessments for creating and evaluating HCTs.
- It evaluates HCT methods that combine health recommender systems based on novel matching algorithm with equity adjustments, showing that the best results come from combining propensity score matching with IPF [13] weighting.
- It demonstrates that using both health recommender systems and equity adjustments results in more equitable populations and precise estimates of the Population Hazard Ratio (PHR), even

with smaller CC sizes.

- It shows that variations in the sizes of treatment and control groups significantly affect the precision of treatment effect estimates, and that a balanced use of recommended synthetic controls is essential for accuracy.

There are several strategies for integrating synthetic control populations into trial populations. Most approaches use propensity score matching to select suitable external controls. Some methods focus on matching based on treatment propensity [4, 14, 15], while others use propensity to predict trial participation [3, 2, 1]. Bayesian approaches are also used to incorporate synthetic controls into trials [16, 17], but they typically do not consider equity adjustments for a target population. Future work with FRESCA could explore these and other methods to develop new strategies for equitable synthetic control arms and assess their effectiveness. Further theoretical exploration of these and other HCT algorithms is also planned.

2. Methodology

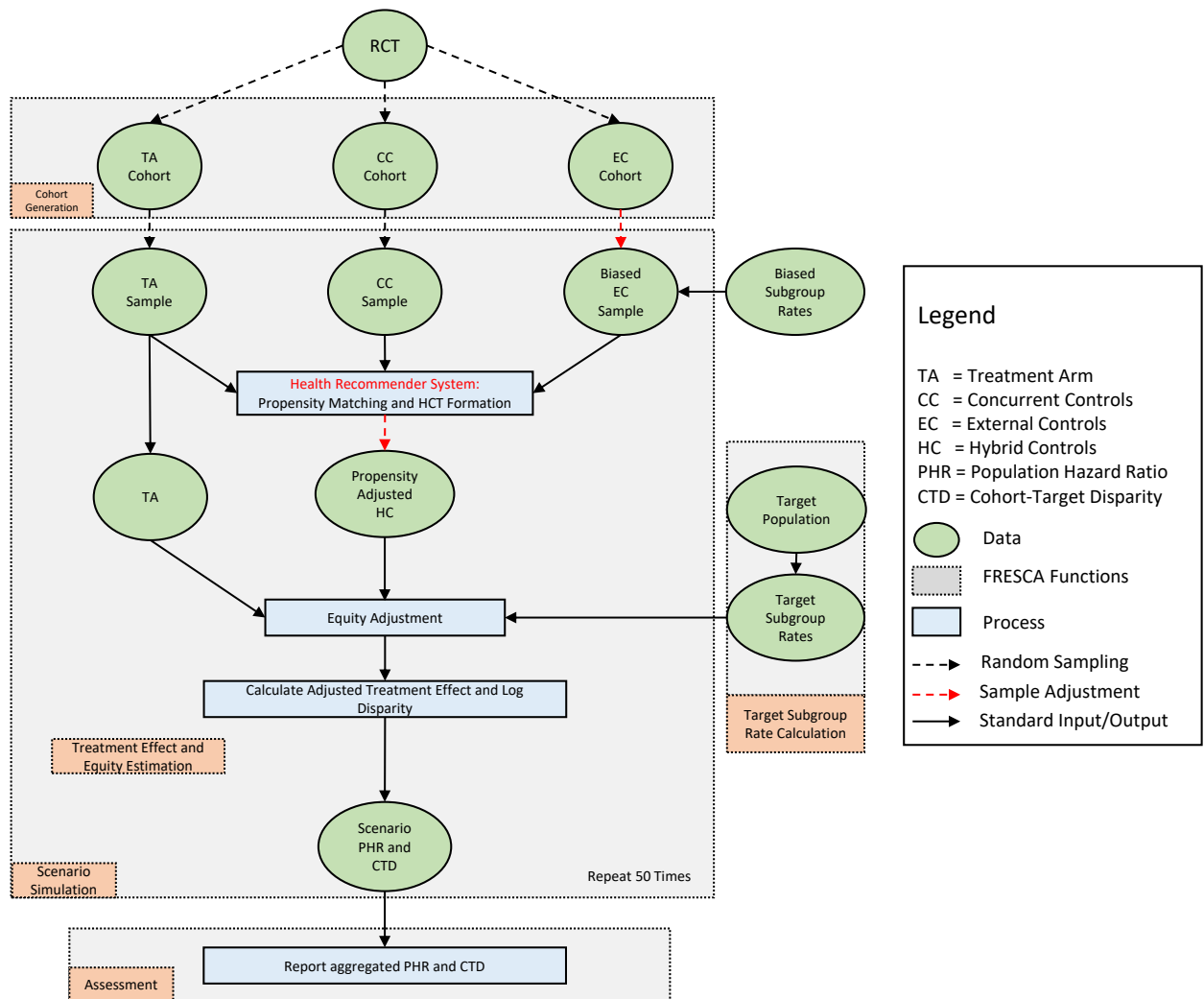


Figure 1: The FRESCA framework for hybrid clinical trials has five main functions. It utilizes a health recommender system for the “Propensity Matching and HC Formation” process and supports any standard method for distribution adjustment in the “Equity Adjustment” process.

2.1. Problem Definition

We define the problem using the potential outcomes framework from Neehal et al. [9]. Let Y_{st}^i represent the potential outcome for subject i in sample s under treatment t , where $s = 0$ is the target population, $s = 1$ the RCT population, $s = 2$ the RWD population, and $s = 3$ an adjusted sample combining RCT and RWD data. The Sample Hazard Ratio (SHR) in the RCT is $SHR = E(\text{effect}(Y_{11}, Y_{10})|S = 1)$, where $\text{effect}(Y_{11}, Y_{10})$ is the difference in treatment effects between treated and control groups. The Population Hazard Ratio (PHR) for the target population ($s = 0$) is defined as $PHR = E(\text{effect}(Y_{01}, Y_{00})|S = 0)$, representing the expected treatment effect in the target population. Two main issues arise: (1) *Equity*—the RCT may not represent the target population, leading to biased estimates ($PHR \neq SHR$), and (2) *Sample Size*—insufficient patients in the control group may require synthetic controls (SCs) from RWD. To accurately estimate PHR, we use a health recommender system with propensity score matching to augment RCT data with SCs, and then perform appropriate equity adjustment, forming an "equity-adjusted" sample ($s = 3$).

2.2. Data

We define the target population using the nationally representative hypertensive cohort from the National Health and Nutrition Examination Survey (NHANES) 2015-2016 [12]. Representativeness in the RCTs is assessed based on three protected attributes: Gender (Male, Female), Age Group (40-59, 59+), and Race/Ethnicity (Non-Hispanic Black, Non-Hispanic White, Non-Hispanic Asian, Hispanic, Other). These age groups align with the inclusion criteria of the SPRINT and ALLHAT hypertension studies. Target rates for each subgroup are calculated using survey-weighted analysis of US subjects aged 40+ with hypertension.

We use data from SPRINT and ALLHAT trials available from BioLINCC [18]. For ALLHAT, we focus on the Amlodipine vs. Clorthalidone group, as the results are similar for the Lisinopril vs. Clorthalidone group. After preprocessing, SPRINT includes 4,234 treated and 4,200 control patients, and ALLHAT includes 8,116 treated and 13,762 control patients. The primary outcome for SPRINT is a composite of major cardiovascular events, while for ALLHAT, the outcome is heart failure. Figure 1 illustrates how FRESKA divides the RCT data into Treatment Arm (TA), Concurrent Controls (CC), and External Controls (EC) cohorts. Table 1 presents the distribution of protected attributes in the RCT data and the Biased External Controls for both SPRINT and ALLHAT, showing their differences from the target NHANES population.

The NHANES surveillance data and clinical trial data from BioLINCC were used with appropriate approvals: BioLINCC approved ALLHAT and SPRINT data use under case 123537, and NHANES data is freely available and exempt from human subjects research regulations per Rensselaer Polytechnic Institute IRB 1863.

2.3. Adjustment Methods and Assessment Metrics

For balancing distributions between synthetic control and trial populations through the recommender system, we employ propensity score matching using the "MatchIt" R package [19]. Iterative Proportional Fitting (IPF) via the "IPFR" R package [20] is used for equity adjustment and Biased EC Cohort formation.

Treatment effects are assessed using Cox's Proportional Hazards Regression to estimate the Population Hazard Ratio (PHR). The "ground-truth" Target PHR is estimated by equity adjustment on the entire RCT dataset. For equity assessment, we use a variant of log disparity (LD)[5]:

$$\log \left\{ \frac{\text{odds}(g(x) = 1|y' = 1)}{\text{odds}(g(x) = 1|y = 1)} \right\} \quad (1)$$

where $g(x)$ is the protected group, y' is the observed cohort, and y is the target population. Absolute LD values between 0 and 0.22 are considered equitable [5]. We introduce Cohort-Target Disparity (CTD) as the mean of median absolute LD values across simulated runs, calculated for subgroups defined by age,

race, and gender. This provides a comprehensive measure of demographic representativeness between the study cohort and target population.

2.4. FRESCA Framework

FRESCA integrates a health recommender system to select suitable synthetic controls from external data and combine them with RCT data, forming a hybrid control population. This framework uses propensity score matching for patient selection and equity adjustments to ensure accurate estimates of the Population Hazard Ratio (PHR) while maintaining representativeness for any target population. FRESCA provides tools to assess the effectiveness of these methods, as shown in Fig 1, and comprises five main functions. We demonstrate FRESCA’s application using the SPRINT and ALLHAT trials with NHANES as the target population, but the framework is flexible and can be adapted to any RCT, RWD, or target population.

2.4.1. Cohort Generation

In the Cohort Generation phase, FRESCA employs its health recommender system to generate three cohorts: TA (Treatment Arm), CC (Concurrent Controls), and EC (External Controls). The TA and CC cohorts are derived from RCT data, representing the treatment and control groups, respectively. The EC cohort is sourced from real-world data, providing a pool of synthetic controls recommended by the health recommender system to supplement the RCT. These cohorts collectively form the basis for subsequent analyses.

2.4.2. Scenario Simulation

FRESCA facilitates the creation of diverse simulated trial scenarios to calculate adjusted PHRs and equity metrics. This involves two stages: first, generating unbiased samples from the TA and CC cohorts to simulate a randomized clinical trial; second, creating a Biased EC Sample from the EC cohort, reflecting biased real-world data as the source of synthetic controls. The health recommender system plays a key role in selecting these controls. Further details on simulation configurations are provided in section 2.5.

Table 2

Comparison of PHR and CTD across different trials, outcomes and methods. We show this for ALLHAT ($N_{TA} = 4000$, $N_{CC} = 2000$) and SPRINT ($N_{TA} = 2000$, $N_{CC} = 1000$) respectively. Symbol (†) in Cohort-Target Disparity column indicates measured CTD not being within equitable range ($CTD > 0.22$). Bold font indicates the best performing method.

Trial (Study)	Outcome Examined	Control Population	Adjustment Method	Target PHR [95% CI]	Estimated PHR [95% CI]	Cohort-Target Disparity [95% CI]
ALLHAT (Hypertension)	Secondary (Heart Failure)	CC	None	1.38 [1.36, 1.41]	1.39 [1.36, 1.43]	0.89 [0.84, 0.94] [†]
		Hybrid	NC Matching		1.42 [1.37, 1.48]	0.87 [0.81, 0.94] [†]
		Hybrid	Propensity Matching + IPF Sampling		1.43 [1.32, 1.49]	0.03 [0.02, 0.04]
		Hybrid	Propensity Matching + IPF Weighting		1.39 [1.33, 1.46]	0.04 [0.03, 0.05]
SPRINT (Hypertension)	Primary	CC	None	0.79 [0.77, 0.82]	0.75 [0.73, 0.78]	0.91 [0.86, 0.97] [†]
		Hybrid	NC Matching		0.74 [0.72, 0.77]	0.89 [0.84, 0.96] [†]
		Hybrid	Propensity Matching + IPF Sampling		0.75 [0.67, 0.84]	0.01 [0.00, 0.01]
		Hybrid	Propensity Matching + IPF Weighting		0.78 [0.74, 0.81]	0.04 [0.03, 0.05]

2.4.3. Target Subgroup Rates Calculation

The target rates for each subgroup are calculated using a survey-weighted analysis of the desired target population from NHANES (e.g., adults with hypertension).

2.4.4. Treatment Effect and Equity Estimation

Once the scenario samples are created and target subgroup rates are determined, the next step is to construct an equity-adjusted HCT population and estimate the treatment effect. FRESCA utilizes a

health recommender system based on propensity score matching to recommend suitable synthetic controls (SCs) from the Biased EC Sample. A binary logistic regression model, incorporating TA, CC, and Biased EC samples, generates propensity scores to select SCs, thereby forming a matched Hybrid Control Arm (HCA). Equity adjustments are then applied to both the TA and HCA cohorts using the Iterative Proportional Fitting (IPF) technique to better align them with the target population. This process results in specific weight vectors, W_{IPF_TA} for the TA and W_{IPF_HCA} for HCA. Unlike the previously used approach [9], where these weights were used to generate random samples, we now directly compute the weighted and equity-adjusted treatment effect and equity value (LD) using these weights.

2.4.5. Assessment

In the Assessment phase, FRESKA evaluates various HCT construction methods by combining the health recommender system for propensity matching with two types of IPF-based equity adjustments: weighted and sampling. Baseline scenarios without any adjustments or inclusion of SC are also evaluated and compared with the NC Matching technique [14], with results summarized in Table 2. To assess the precision of PHR estimations, FRESKA compares them with a “ground-truth” target PHR, derived from equity adjustments applied to the complete RCT dataset (e.g., SPRINT/ALLHAT) using all treated and control subjects. The data set is divided into treated and control cohorts, bootstrapped to match the sizes, and adjusted to align with the NHANES population. The target PHR is calculated as an average across all bootstrap samples and scenarios. Equity is evaluated by checking if the Cohort-Target Disparity (CTD) falls within the $[0, 0.22]$ range, adhering to the 80% rule [5].

2.5. Simulation of HCT Scenarios

FRESKA simulates various clinical trial scenarios to evaluate the effects of different trial design parameters on outcomes using a health recommender system for patient selection.

2.5.1. ALLHAT

For ALLHAT, FRESKA creates a Concurrent Control (CC) Cohort by selecting 4,000 individuals from the ALLHAT trial’s original control group, leaving 9,762 in the External Control (EC) Cohort. The Treatment Arm (TA) remains unchanged with 8,116 participants. During simulation, biases are introduced in the EC Cohort using the Iterative Proportional Fitting (IPF) method to reflect biased subgroup rates for age, gender, and race, as well as smoker status, depression history, and HDLC history (Table 1). FRESKA explores different experimental scenarios with varying sample sizes for TA ($N_{TA}=4000$) and CC ($N_{CC}=0, 500, 1000, 2000, 4000$), conducting 50 bootstrap simulations for each scenario. The mean Population Hazard Ratio (PHR) and Cohort-Target Disparity (CTD) are calculated for each scenario, along with a 95% Confidence Interval.

2.5.2. SPRINT

For SPRINT, FRESKA selects a CC Cohort of 2,000 from the control group (total $N=4200$), leaving 2,200 in the EC Cohort, with the TA consisting of 4,234 participants. IPF is used to adjust for biases in three protected attributes (Table 1) and additional factors such as Framingham Risk Score and Cardiovascular Disease (CVD) History. Simulations use a TA sample size ($N_{TA}=2000$) and vary CC sample sizes ($N_{CC}=0, 500, 1000, 1500, 2000$), running 50 bootstrap simulations per scenario. Results include mean PHR, CTD, and 95% Confidence Intervals, as reported in the final assessment.

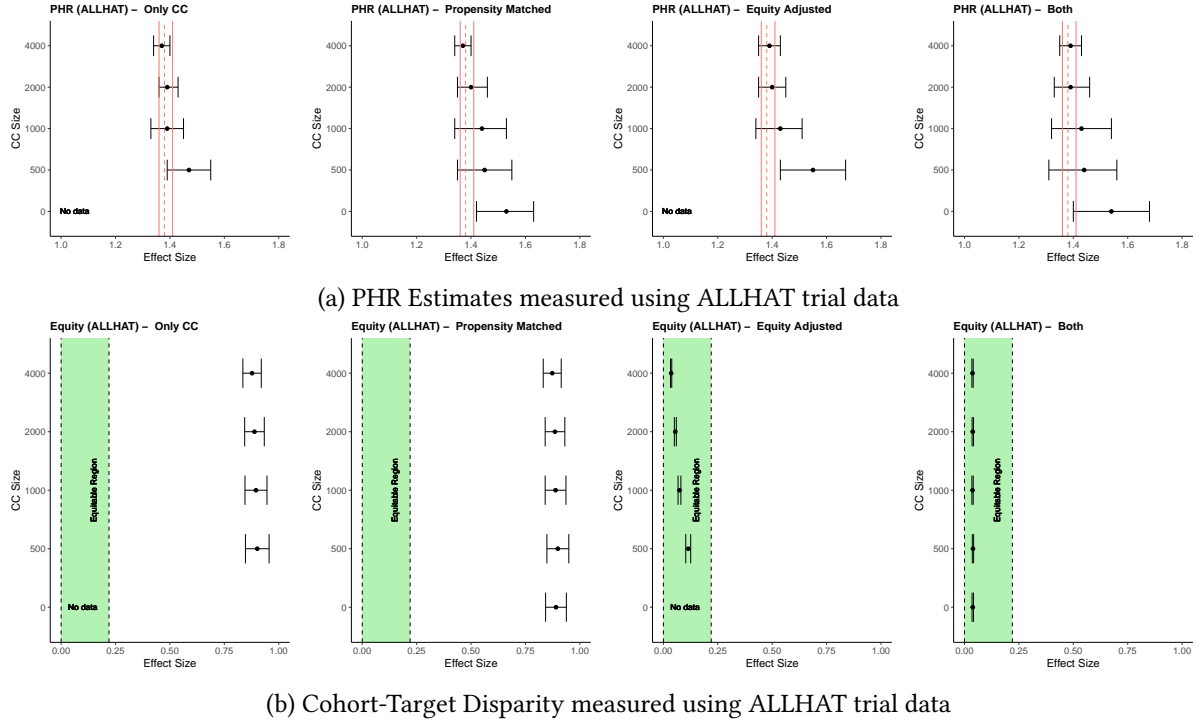


Figure 2: Variation of PHR Estimates and Cohort-Target Disparity in ALLHAT Trial with Different CC Sizes. This figure illustrates the influence of various CC sizes ($N_{CC} = 0, 500, 1000, 2000, 4000$) on the PHR estimates and equity measures for ALLHAT trial with a treatment arm size of $N_{TA} = 4000$. In panel (a), the target PHR is demarcated by two solid red lines, encompassing the red dashed line representing the 95% confidence interval. Panel (b) features a green shaded area delineated by two black dashed lines, indicating the range considered equitable.

3. Results and Discussion

3.1. Performance comparison of different methods for creating HCTs

Our study evaluates methods for creating HCTs using two metrics as described above: PHR and CTD. We analyze methods that combine propensity score matching with Iterative Proportional Fitting (IPF) equity-adjustment methods (weighting or random sampling) and compare them to two baseline scenarios: (i) no adjustments applied to the CC population, and (ii) SC added to trial via NC Matching [14] algorithm. The results for the ALLHAT and SPRINT trials are detailed in Table 2. Our analysis reveals two key findings: (i) *Necessity of both Propensity and Equity Adjustments*: PHR estimates without equity adjustments (either by IPF weighting or sampling) are typically inequitable as evident by CTD values, and they may produce biased treatment effects, as evident in high CTD values, and (ii) *Superiority of Weighting over Sampling*: Using IPF Weighting for equity adjustment improves the accuracy and consistency of PHR estimates compared to IPF Sampling. These findings, particularly the need for comprehensive adjustments and the effectiveness of sample weighting, have been observed consistently across multiple trials and outcomes, demonstrating the robustness of the FRESKA framework.

3.2. Examination of Variation in CC Size on PHR and CTD estimation

We studied the effect of varying CC population sizes on the estimated PHR and CTD in HCT. We examined four methods: (i) no adjustments, (ii) only propensity matching, (iii) only equity adjustment, and (iv) both propensity and equity adjustments. Fig 2 shows the results on ALLHAT for CC sizes varying from 0 to 4000. Two main findings emerged - (i) *Benefits of Synthetic Controls for Limited Data*: In scenarios with smaller CC sizes, missing subgroups were compensated by incorporating SC. This strategy, especially with both propensity and equity adjustments, yielded accurate estimates, and (ii)

Table 3

The effect and significance of several trial design parameters in predicting the bias in PHR estimation. Here the bias is defined to be the squared deviation of the estimated PHR from the target PHR. Star (*) symbol represents a significant effect with $p < 0.05$.

Predictors of Linear Model	Estimate	2.5%	97.5%	P Value
TA Size	0.022	0.009	0.036	0.001*
CC Size	-0.134	-0.157	-0.110	0.000*
Cohort-Target Disparity (CTD)	-0.062	-0.183	0.059	0.316
Cohort-RCT Disparity (CRD)	0.160	0.021	0.298	0.024*
Controls (Only Equity)	-0.132	-0.264	-0.001	0.048*
Controls (Both)	-0.089	-0.222	0.043	0.186
Controls (Only Propensity)	0.035	0.020	0.049	0.000*

PHR Accuracy and Acceptable Equity with Reduced CC Population Size: Reducing the CC population size by 50% (from 4000 to 2000) still produced PHR estimates close to the target PHR. However, a larger CC size is preferable for lower estimation variance, indicating a trade-off for trial designers. These patterns were consistent across both SPRINT and ALLHAT trials, affirming the robustness of our findings. We only show results for weighting equity adjustment for the ALLHAT trial for brevity, and additional results are available in the Supplementary.

3.3. Examining effects of multiple factors for predicting PHR estimation accuracy

We analyzed factors affecting the accuracy of PHR estimates in ALLHAT using a linear model with seven predictors. Bias in PHR estimation was quantified as the squared deviation from the target PHR. Predictors included the size of the treatment arm (TA Size), the control group size (CC Size), Cohort-Target Disparity (CTD), and Cohort-RCT Disparity (CRD), which measures the distribution differences between control populations and the RCT population. Controls were categorized by adjustment type: propensity, equity, both, or none. Results in Table 3 showed that TA Size, CC Size, and CRD significantly predicted PHR bias. Key findings include: (i) *Larger CC Size Reduces Bias:* Increasing CC size lowers bias, favoring a larger control group directly recruited over synthetic controls; (ii) *Impact of Adjustments:* "Only Equity" and "Propensity and Equity" adjustments reduce bias compared to the "Only CC" category, while "Only Propensity" adjustments increase bias, highlighting the importance of equity adjustments for accurate PHR estimates.

3.4. Examining the effect of TA and CC Size ratio for a Fixed Size Recruitment Trial

We examined the balance between TA and CC population sizes when supplemented by SC in clinical trials with a fixed recruitment size. Using ALLHAT and SPRINT trial data, we maintained a total recruited participant cap of 4000, varying TA and CC sizes with corresponding SC adjustments. Four

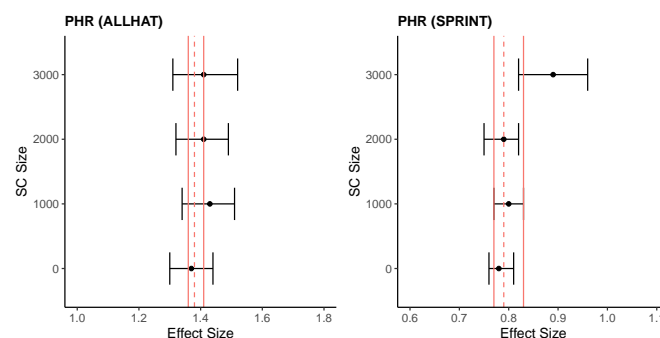


Figure 3: Examining the influence of varied TA, CC, and SC Sizes on PHR Estimation in Fixed Sized Recruitment Trials.

scenarios were considered with TA sizes of 3500, 3000, 2500, 2000 and CC sizes of 500, 1000, 1500, 2000, inversely adjusting SC sizes 3000, 2000, 1000, 0. The PHR estimates from these scenarios are shown in Figure 3. Key findings include: (i) The variance of PHR estimates increases with SC size, affecting the stability of treatment effect estimation, and (ii) The PHR estimate can significantly shift with a highly imbalanced ratio of real to synthetic data; especially observed in some scenarios with substantially high SC size. This investigation therefore highlights the importance of carefully balancing the ratio of CC and SC patients in HC to ensure accurate treatment effect estimates and avoiding erroneous conclusions about a trial's efficacy.

4. Conclusion

FRESCA offers a major advancement in equitable HCT methods and serves as a valuable tool for future research. It creates realistic HCT scenarios, using a health recommender system for propensity score matching and equity adjustments to provide more precise and equitable PHR estimates. Our simulations suggest that fewer patients may be needed to achieve results similar to full trials, but further research is required to determine the optimal balance of synthetic and concurrent controls in fixed-size trials. Future work will involve testing FRESCA with more realistic EHR data, with additional protected attributes and exploring the optimal size for CC recruitment during RCT design. Additionally, developing strategies that integrate matching and equity adjustments in a single step could enhance efficiency and reduce variance. These areas present opportunities for further refinement, making FRESCA a significant step forward in hybrid clinical research with potential for ongoing improvement.

Acknowledgments

This work was partially funded by IBM Research. This manuscript was prepared using SPRINT and ALLHAT study research materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of SPRINT, ALLHAT or NHLBI.

References

- [1] A. Sachdeva, R. C. Tiwari, S. Guha, A novel approach to augment single-arm clinical studies with real-world data, *Journal of Biopharmaceutical Statistics* 32 (2022) 141–157. doi:10.1080/10543406.2021.2011902.
- [2] J. Harton, B. Segal, R. Mamtani, N. Mitra, R. A. Hubbard, Combining real-world and randomized control trial data using data-adaptive weighting via the on-trial score, *Statistics in Biopharmaceutical Research* (2022) 1–13. doi:10.1080/19466315.2022.2071982.
- [3] X. Yin, P. S. Mishra-Kalyan, R. Sridhara, M. D. Stewart, E. A. Stuart, R. C. Davi, Exploring the potential of external control arms created from patient level data: a case study in non-small cell lung cancer, *Journal of Biopharmaceutical Statistics* 32 (2022) 204–218. doi:10.1080/10543406.2021.2011901.
- [4] E. A. Stuart, D. B. Rubin, Matching with multiple control groups with adjustment for group differences, *Journal of Educational and Behavioral Statistics* 33 (2008) 279–306. doi:10.3102/1076998607306.
- [5] M. Qi, O. Cahan, M. A. Foreman, D. M. Gruen, A. K. Das, K. P. Bennett, Quantifying representativeness in randomized clinical trials using machine learning fairness metrics, *JAMIA Open* 4 (2021) ooab077. doi:10.1093/jamiaopen/ooab077.
- [6] E. Hartman, R. Grieve, R. Ramsahai, J. S. Sekhon, From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (2015) 757–778. doi:10.1111/rssa.12094.

- [7] A. Y. Ling, M. E. Montez-Rath, P. Carita, K. J. Chandross, L. Lucats, Z. Meng, B. Sebastien, K. Kapphahn, M. Desai, An overview of current methods for real-world applications to generalize or transport clinical trial findings to target populations of interest, *Epidemiology* 34 (2023) 627–636. doi:10.1097/EDE.0000000000001633.
- [8] J. Petkovic, J. Jull, M. Yoganathan, O. Dewidar, S. Baird, J. M. Grimshaw, K. A. Johansson, E. Kristjansson, J. McGowan, D. Moher, et al., Reporting of health equity considerations in cluster and individually randomized trials, *Trials* 21 (2020) 1–12. doi:10.1186/s13063-020-4223-5.
- [9] N. Neehal, V. Anand, K. P. Bennett, Framework for research in equitable synthetic control arms, in: *AMIA Annual Symposium Proceedings*, volume 2023, American Medical Informatics Association, 2023, p. 530.
- [10] J. T. Wright Jr, et al., A randomized trial of intensive versus standard blood-pressure control, *New England Journal of Medicine* 373 (2015) 2103–2116. doi:10.1056/NEJMoa1511939.
- [11] C. D. Furberg, et al., Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (allhat), *JAMA* 288 (2002) 2981–2997. doi:10.1001/jama.288.23.2981.
- [12] Centers for Disease Control and Prevention, National health and nutrition examination survey (nhanes) data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2023. URL: <https://www.cdc.gov/nchs/nhanes/index.htm>, accessed on: October 2023.
- [13] W. E. Deming, F. F. Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *The Annals of Mathematical Statistics* 11 (1940) 427–444. doi:10.1214/aoms/1177731829.
- [14] J. Yuan, J. Liu, R. Zhu, Y. Lu, U. Palm, Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls, *Journal of Biopharmaceutical Statistics* 29 (2019) 558–573. doi:10.1080/10543406.2018.1559853.
- [15] Y. Liu, B. Lu, J. Foster, Y. Zhang, Z. J. Zhong, M.-H. Chen, P. Sun, Matching design for augmenting the control arm of a randomized controlled trial using real-world data, *Journal of Biopharmaceutical Statistics* 32 (2022) 124–140. doi:10.1080/10543406.2021.2011900.
- [16] K. Viele, et al., Use of historical control data for assessing treatment effects in clinical trials, *Pharmaceutical statistics* 13 (2014) 41–54. doi:10.1002/pst.1589.
- [17] X. Pang, et al., A bayesian alternative to synthetic control for comparative case studies, *Political Analysis* 30 (2022) 269–288. doi:10.1017/pan.2021.22.
- [18] National Heart, Lung, and Blood Institute, Biologic specimen and data repository information coordinating center, <https://biolincc.nhlbi.nih.gov/home/>, 2024. Accessed September 22, 2024.
- [19] D. E. Ho, K. Imai, G. King, E. A. Stuart, Matchit: Nonparametric preprocessing for parametric causal inference, *Journal of Statistical Software* 42 (2011) 1–28. doi:10.18637/jss.v042.i08.
- [20] K. Ward, IPFR: List Balancing for Reweighting and Population Synthesis, 2020. URL: <https://CRAN.R-project.org/package=ipfr>, r package version 1.0.2.