pHealth 2024
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI240075

Evaluating MediBetter: A Mobile Application for Health Monitoring and Medication Management

Bian YANG [a], M. Ali FAUZI [b,1] and Galih Wasis WICAKSONO [c]

[a] *Norwegian University of Science and Technology, Gjøvik, Norway*
[b] *Universitas Brawijaya, Malang, Indonesia*
[c] *Universitas Muhammadiyah Malang, Malang, Indonesia*

ORCiD ID: Bian Yang https://orcid.org/0000-0001-6189-1976,
M. Ali Fauzi https://orcid.org/0000-0002-9696-2807,
Galih Wasis Wicaksono https://orcid.org/0000-0002-8096-1762

Abstract. This study introduces MediBetter, a mobile application designed to empower patients undergoing routine medication in health monitoring and medication adherence. It is a mobile application designed to serve as a supportive health technology for patients to monitor their health status and manage their routine medication. It offers three main features: text-based daily self health report, AI-based summarization of the health report, and medication taking reminder. To evaluate the quality of generated summaries generated by both the user and AI (ChatGPT), we conducted human expert evaluation process. Furthermore, we also evaluated the usefulness of existing features in the app. The experiment results show that ChatGPT-generated summaries outperformed user-generated ones, demonstrating superior informativeness, coherence, fluency, consistency, and contradiction handling. Participants found the app's features highly useful for health monitoring and medication adherence, with strong agreement on their utility.

Keywords. mobile application, summary, AI, usefulness

1. Introduction

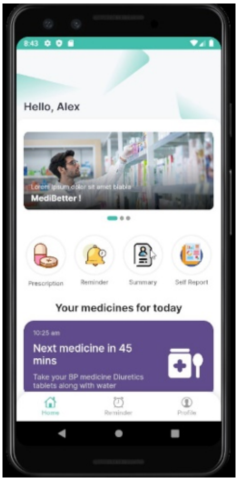
The rise of mobile technology in recent years has drastically changed many facets of healthcare, giving people the tools they need to take control of their health in a tailored and easily accessible way [1,2]. This shift towards mobile health solutions is particularly crucial for patients undergoing routine medication, as it addresses the need for consistent monitoring and management of their health status. Recording daily health status or diary entries allows patients to track their symptoms, activities, and vital signs, providing valuable data for both self-assessment and communication with healthcare providers [3]. Additionally, the ability to summarize these diary entries offers a convenient way for patients to review their health trends and patterns over time, facilitating informed decision-making and proactive health management. Furthermore, medication-taking reminders integrated into mobile apps ensure adherence, minimizing

[1] Corresponding Author: M. Ali Fauzi, moch.ali.fauzi@ub.ac.id

the risk of missed doses and improving medication adherence rates among patients [4]. The objective of this research is to evaluate the feature of a mobile application called MediBetter that is tailored for patients undergoing routine medication. This application aims to help patients undergoing routine medication with features for recording text-based health status, summarizing diary entries, and providing medication reminders. In this paper, we assess the usefulness of the application features. In addition, since we employ AI to do the health diary summarization, we also evaluate the efficacy of the generated summary.

2. Methods

MediBetter is a mobile application designed to serve as a supportive health technology for patients to monitor their health status and manage their routine medication. The app consists of three main functions related to medication and health monitoring: daily self health report, summarization of the health report, and medication taking reminder. Screenshots of the app are displayed in Figure 1.



Your medicines for today

10:25 am

Next medicine in 45 mins

Take your BP medicine Diuretics tablets along with water

Figure 1. The home screen of the app

In this research, we aim to evaluate the efficacy and user experience of MediBetter application among patients undergoing routine medication. Participants will be instructed to utilize the mobile app for recording daily health-related activities, symptoms, and vital signs. They will be prompted to summarize their health diary entries periodically. Additionally, ChatGPT, an AI-based tool, will be tasked with automatically summarizing participants' health diary entries, offering an alternative approach to data summarization. The summarization results produced by the user and ChatGPT will undergo expert evaluation by healthcare professionals.

To evaluate the quality of generated summaries generated by both the user and ChatGPT, we employed extensive human assessments. We adopted standardized metrics commonly used in previous studies [5,6,7,8] . These metrics encompass the following dimensions: (1) Informativeness or Relevance: Quantifying the ability of the summary to retain important and relevant facts and information. (2) Coherence:

Assessing the presence of smooth logical transitions between summary sentences or paragraphs. (3) Redundancy: Evaluating whether the summary contains repeated information or facts. This metric measures the summary should contain few repetition, with a higher score indicating a lower level of redundancy. (4) Fluency: Measuring the grammatical correctness and linguistic fluency of the summary text. (5) Consistency or Factuality: Verifying the factual accuracy of the summary in comparison to the source article. (6) Contradiction: Identifying instances where information within the summaries contradicts other information or disagrees with another piece of information. This metric measures the summary should contain few contradiction, with a higher score indicating a lower level of contradiction.

In our human evaluation process, both the generated summaries and their corresponding reference health diaries are needed. We utilized all generated summaries and their corresponding reference texts from the dataset were utilized. We enlisted the participation of four expert volunteers for evaluation, comprising one physicians and three nurses. They were tasked with rating these summaries across six aspects using a scale ranging from 1 (very bad) to 5 (very good). Furthermore, we also asked the human evaluator about their preference, whether they prefer summary by user or ChatGPT.

Furthermore, participants will be surveyed to gather feedback on the usefulness of the app's features in health and medication management. We adopted usefulness items from previous studies [9]. After using the app for four weeks, participants were tasked with rating the usefulness of the app's features using a Likert scale with responses ranged from 1 to 5, spanning from strongly disagree to strongly agree. The following are the items:

- The diary / daily health status report feature helps in health monitoring.
- The diary / daily health status report summary feature helps in health monitoring.
- The medication taking reminder feature helps in health monitoring.

By incorporating both user-generated and AI-generated summarizations, along with expert evaluations and user feedback, this research aims to provide a comprehensive assessment of the mobile app's utility in facilitating health diary management and improving user engagement in personal health monitoring and medication management.

3. Result

3.1. Participant Characteristics

In this study, 23 patients were participated in using the application and providing the user summary. However, only 19 participants were filling the demographic and app usefulness survey. Table 1 summarizes demographic characteristics of 19 participants included in this study. We had more female participants (12) than the male one (7). More than half of the participants are 21-30 years old (52.63%) while there are 3 participants for each age category of 31-40, 41-50, and over 50. In term of education, most of them got their last education at bachelor level (63.16%).

Table 1. Participant Characteristics

Variable	Category	n	%
Gender	Female	12	63.16 %
	Male	7	36.84 %
Age	21-30	10	52.63 %
	31-40	3	15.79 %
	41-50	3	15.79 %
	Over 50	3	15.79 %
Education	Elementary school	1	5.26 %
	High school	3	15.79 %
	Bachelor	12	63.16 %
	Master	3	15.79 %

Table 2. Summary evaluation results using human expert evaluation

Aspect	Summary generated by user			Summary generated by ChatGPT			t-test result
	mean	std	kappa	mean	std	kappa	
Informativeness	3.20	1.30	0.066	4.60	0.70	-0.067	t(91)=-8.494, p<.001
Coherence	3.39	1.31	0.134	4.45	0.80	-0.038	t(91)=-5.907, p<.001
Redundancy	3.83	1.20	-0.074	3.52	1.52	-0.127	t(91)=2.233, p=.028
Fluency	3.47	1.34	0.022	4.49	0.78	-0.009	t(91)=-5.690, p<.001
Consistency	3.52	1.30	0.038	4.52	0.72	-0.080	t(91)=-6.086, p<.001
Contradiction	3.83	1.41	0.045	4.30	1.10	-0.076	t(91)=-3.118, p=.002

3.2. Summarization

We obtained 23 pair of summaries by user and ChatGPT. We asked four expert to rate the summary result on six aspects based on the corresponding health diaries. The results of the human expert evaluation are depicted in Table 2. We computed the mean and standard deviation for each aspect. In addition, we also measured the agreement between experts using Fleiss's kappas. Furthermore, we evaluated the mean score difference between summary generated by user and summary generated by ChatGPT. A two-tailed paired t-test with p<.001 was employed to test whether the difference is significant.

The results show that ChatGPT gave a better summary that user in five aspects, including informativeness, coherence, fluency, consistency, and contradiction. The t-test results indicates that summary generated by ChatGPT is significantly better that summary generated by user in four aspects, including informativeness, coherence, fluency, and consistency. For the contradiction aspect, even though it is not significant, the p value is 0.002, which is can be considered significant if we use p<.005. Summary by user is only better in one aspect, namely redundancy. However, the t-test result shows that the difference is not significant. Meanwhile, the kappas score indicates that the agreement between human evaluators are very poor for both the evaluation for summay generated by user and summary generated by ChatGPT.

Furthermore, we also asked the human evaluator about their preference, whether they prefer summary by user or ChatGPT in term of patient care support. The results that are shown in Table 3 indicates that all of the expert prefer summary by ChatGPT more than summary by user.

Table 3. The number of summary based on expert preference.

Preference	Expert I	Expert II	Expert III	Expert IV
Prefer summary by user	3	7	2	4
Prefer summary by ChatGPT	20	16	21	19

3.3. Usefulness

Majority of the participants perceived that of the three app features are useful. The mean score for health diary feature, summary feature, and medication taking reminder feature are 4.32 (SD 0.75), 4.21 (SD 0.71), and 4.42 (SD 0.69), respectively.

4. Conclusion

In conclusion, our findings indicate that the automated summarization by ChatGPT demonstrating superior performance in terms of informativeness, coherence, fluency, consistency, and contradiction compared to manual summarization by user. Expert evaluations also favored ChatGPT-generated summaries, highlighting the potential of AI-driven approaches in healthcare data summarization. Furthermore, participants perceived the app features, including the health diary, summary, and medication reminder, as highly useful. The high ratings obtained for these features underscore their importance in empowering patients to actively engage in their healthcare journey and adhere to prescribed treatment regimens. Further research and development efforts can focus on refining the app's features and enhancing its integration into routine clinical practice, ultimately improving patient outcomes and enhancing healthcare delivery.

References

[1] Siegler AJ, Knox J, Bauermeister JA, Golinkoff J, Hightow-Weidman L, Scott H. Mobile app development in health research: pitfalls and solutions. Mhealth. 2021;7.

[2] Sama PR, Eapen ZJ, Weinfurt KP, Shah BR, Schulman KA. An evaluation of mobile health application tools. JMIR mHealth and uHealth. 2014;2(2):e3088.

[3] Appelboom G, Camacho E, Abraham ME, Bruce SS, Dumont EL, Zacharia BE, et al. Smart wearable body sensors for patient self-assessment and monitoring. Archives of public health. 2014;72:1-9.

[4] Ahmed I, Ahmad NS, Ali S, Ali S, George A, Danish HS, et al. Medication adherence apps: review and content analysis. JMIR mHealth and uHealth. 2018;6(3):e6432.

[5] Cai X, Liu S, Yang L, Lu Y, Zhao J, Shen D, et al. COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers. Journal of Biomedical Informatics. 2022;127:103999.

[6] Jain R, Jangra A, Saha S, Jatowt A. A survey on medical document summarization. arXiv preprint arXiv:221201669. 2022.

[7] Shah D, Yu L, Lei T, Barzilay R. Nutri-bullets hybrid: Consensual multi-document summarization. Association for Computational Linguistics (ACL); 2021..

[8] Wallace BC, Saha S, Soboczenski F, Marshall IJ. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. AMIA Summits on Translational Science Proceedings. 2021;2021:605.

[9] Wang CJ, Chaovalit P, Pongnumkul S, et al. A breastfeed-promoting mobile app intervention: usability and usefulness study. JMIR mHealth and uHealth. 2018;6(1):e8337.