
JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 13 (2025) ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

| | |
|--|----|
| Patients' Experienced Usability and Satisfaction With Digital Health Solutions in a Home Setting: Instrument Validation Study (e63703) Susan Oudbier, Ellen Smets, Pythia Nieuwkerk, David Neal, S Nurmohamed, Hans Meij, Linda Dusseljee-Peute. | 2 |
| Autonomous International Classification of Diseases Coding Using Pretrained Language Models and Advanced Prompt Learning Techniques: Evaluation of an Automated Analysis System Using Medical Text (e63020) Yan Zhuang, Junyan Zhang, Xiuxing Li, Chao Liu, Yue Yu, Wei Dong, Kunlun He. | 18 |
| The Transformative Potential of Large Language Models in Mining Electronic Health Records Data: Content Analysis (e58457) Amadeo Wals Zurita, Hector Miras del Rio, Nerea Ugarte Ruiz de Aguirre, Cristina Nebrera Navarro, Maria Rubio Jimenez, David Muñoz Carmona, Carlos Miguez Sanchez. | 34 |
| Development and Evaluation of a Mental Health Chatbot Using ChatGPT 4.0: Mixed Methods User Experience Study With Korean Users (e63538) Boyoung Kang, Munpyo Hong. | 49 |

Patients' Experienced Usability and Satisfaction With Digital Health Solutions in a Home Setting: Instrument Validation Study

Susan J Oudbier^{1,2,3,4}, MD; Ellen MA Smets^{2,4,5}, PhD; Pythia T Nieuwkerk^{2,6}, PhD; David P Neal^{3,7}, PhD; S Azam Nurmohamed⁸, MD, PhD; Hans J Meij^{1,9}, PhD; Linda W Dusseljee-Peute⁷, PhD

¹Outpatient Division, Amsterdam University Medical Center, Meibergdreef 9, Amsterdam, the Netherlands

²Department of Medical Psychology, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, the Netherlands

³Digital Health, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands

⁴Quality of Care, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands

⁵Personalized Medicine, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands

⁶Amsterdam Institute for Infection and Immunity, Amsterdam, the Netherlands

⁷Department of Medical Informatics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, the Netherlands

⁸Department of Nephrology, Amsterdam University Medical Center, Amsterdam, the Netherlands

⁹Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Republic of Singapore

Corresponding Author:

Susan J Oudbier, MD

Outpatient Division, Amsterdam University Medical Center, Meibergdreef 9, Amsterdam, the Netherlands

Abstract

Background: The field of digital health solutions (DHS) has grown tremendously over the past years. DHS include tools for self-management, which support individuals to take charge of their own health. The usability of DHS, as experienced by patients, is pivotal to adoption. However, well-known questionnaires that evaluate usability and satisfaction use complex terminology derived from human-computer interaction and are therefore not well suited to assess experienced usability of patients using DHS in a home setting.

Objective: This study aimed to develop, validate, and assess an instrument that measures experienced usability and satisfaction of patients using DHS in a home setting.

Methods: The development of the “Experienced Usability and Satisfaction with Self-monitoring in the Home Setting” (GEMS) questionnaire followed several steps. Step I consisted of assessing the content validity, by conducting a literature review on current usability and satisfaction questionnaires, collecting statements and discussing these in an expert meeting, and translating each statement and adjusting it to the language level of the general population. This phase resulted in a draft version of the GEMS. Step II comprised assessing its face validity by pilot testing with Amsterdam University Medical Center’s patient panel. In step III, psychometric analysis was conducted and the GEMS was assessed for reliability.

Results: A total of 14 items were included for psychometric analysis and resulted in 4 reliable scales: convenience of use, perceived value, efficiency of use, and satisfaction.

Conclusions: Overall, the GEMS questionnaire demonstrated its reliability and validity in assessing experienced usability and satisfaction of DHS in a home setting. Further refinement of the instrument is necessary to confirm its applicability in other patient populations in order to promote the development of a steering mechanism that can be applied longitudinally throughout implementation, and can be used as a benchmarking instrument.

(*JMIR Med Inform* 2025;13:e63703) doi:[10.2196/63703](https://doi.org/10.2196/63703)

KEYWORDS

digital health solutions; questionnaire development; usability instruments; self-management; home setting; validation; reliability

Introduction

The number of digital health solutions (DHS) has increased rapidly, with the potential to significantly enhance the way health care is delivered [1]. DHS include, among others, tools for self-management of clinical data such as blood pressure

measurements, for medication adherence, and for education on health-related behaviours such as diet, smoking, and exercise [2]. These tools present the opportunity to increase access to health care and optimize disease management, and they ultimately aim to alleviate health care expenditure [3]. Self-management, as per the World Health Organization, encompasses the capacity of individuals to support and sustain

their own health, prevent diseases, and cope with illness and disability, whether independently or with the assistance of a health care professional (HCP) [4,5]. The use of DHS serves a dual purpose in patient self-management: (1) facilitating proactive engagement of individuals in their health journey to optimize treatment outcomes and (2) enhancing prevention of negative health outcomes [6,7]. Consequently, ensuring accessibility and adoption of DHS among target users is crucial for effective implementation [8]. The experienced usability of DHS is pivotal to their adoption, especially for individuals with disabilities or those living with chronic diseases who need to make frequent use of a DHS within their care journey [9-11]. Measuring DHS usability and patient satisfaction is crucial to understand and improve accessibility and use of DHS, thereby fostering patient engagement.

The international organization for standardization defines usability, as comprising effectiveness, efficiency, and satisfaction, given a specific user in a context [12]. In the context of DHS, effectiveness refers to the capacity for thorough and accurate task completion, such as logging into a patient portal or setting personal preferences for medication reminders. Efficiency, on the other hand, involves accomplishing these tasks with minimal effort. Finally, satisfaction is expressed as the comfort and acceptability experienced by patients when using a DHS tool. Usability is often measured by (validated) usability and satisfaction questionnaires, as they allow efficient collection and structured assessment of data from a large number of individual users [13,14]. Usability questionnaires originate from the field of human-computer interaction and user-centered design and have emerged as a means to evaluate the effectiveness, efficiency, and satisfaction of interactive systems, particularly software and digital interfaces from the perspective of end users [15]. Therefore, existing well-known and applied usability questionnaires, such as the System Usability Scale (SUS) and mHealth App Usability Questionnaire (MAUQ) apply software terminology such as the “*various functions* in this system,” or “*navigation* between screens” [16-18]. These statements are difficult to interpret for individuals lacking familiarity with software terminology, particularly for patients with low levels of digital literacy [19]. These statements are therefore not suited to measure the usability of self-management tools in healthcare practice by all users.

In addition, introducing DHS in a self-management care journey may increase disparities, as it requires particular skills to use it that comprise both health and digital literacy [20]. In terms of patient characteristics, patients with high health literacy, a higher educational level, and patients who are familiar with DHS find

it easier to use these tools [21]. Variability in digital literacy skills among patients are well-recognized, posing challenges in its utilization [22]. Comprehensive research on the specific patient groups for which DHS is relevant, and our understanding of usability in this domain are still in the nascent stages. Disparities arising within groups due to the utilization of technology might lead to one group adopting the technology, while the other group opts not to use it. With the increasing availability and reliance on DHS [26], these tools should be usable for the majority of the patient population. Evaluations of patient experiences with DHS should therefore also be accessible to diverse groups of patients. Thus, to optimize health outcomes and to deliver high quality care, evaluating patients' experienced DHS usability and satisfaction in a home setting is imperative for health care organizations and HCPs [1,23]. In order to ensure patient inclusivity, a general and accessible instrument is needed, which can be applied as a steering mechanism, deployed at multiple points in time to measure usability and satisfaction of DHS in a home setting.

The aim of this study is to develop, validate, and assess the reliability of an instrument that measures experienced usability of and satisfaction with DHS use, taking digital (language) literacy into account. When developing the Experienced Usability and Satisfaction with Self-monitoring in the Home Setting (GEMS) questionnaire, our goal is to find a middle ground between innovation and familiarity, drawing from established statements and questionnaires while tailoring them to be able to evaluate patients experiences with DHS from an inclusive perspective. In doing so, we aim to advance DHS implementation and expand our understanding of end users' needs, for efficient, effective and satisfied DHS use.

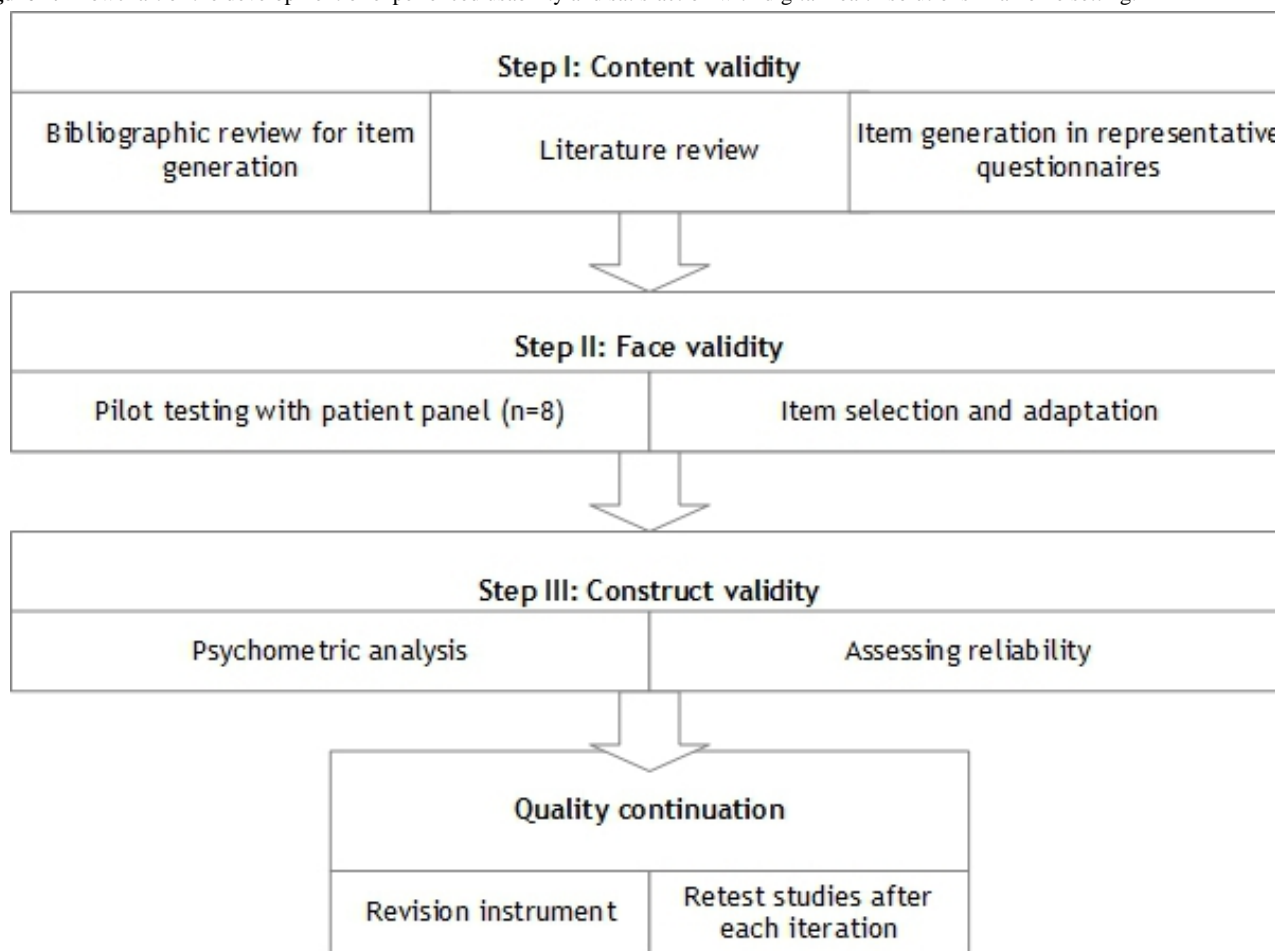
Methods

Ethical Considerations

The Medical Ethical Committee of Amsterdam University Medical Center (Academic Medical Center) declared that this study was not subject to the Medical Research Involving Human Subject Act and that further approval was not required (W22 291 # 22.352).

GEMS Questionnaire Development

To develop and validate the questionnaire “Gebruiksvriendelijkheid en Ervaring met Monitoren in de ThuisSetting,” translated as Experienced Usability and Satisfaction with self-monitoring in the Home Setting, we followed several steps, as depicted in [Figure 1](#).

Figure 1. Flowchart of the development of experienced usability and satisfaction with digital health solutions in a home setting.

Step I: Content Validity - Collecting User Experience Statements

To design the GEMS questionnaire, we first searched for published literature on user experience questionnaires in the context of DHS in PubMed using the keywords “Digital Health Solutions,” “Digital Health Technologies,” “Self-Management tools,” “Digital health apps,” “mHealth apps,” AND (“Usability” OR “Satisfaction”) [24]. We searched for questionnaires that measured end-user experiences, and restricted our search to studies published in the last 5 years due to the rapidly evolving nature of the field.

After the literature review, an expert meeting was held, for which we invited several usability experts in the field. We went through the domains and statements from the validated questionnaires retrieved from the literature search. The outcome of this meeting was a list of requirements for domains with items that should be included in the GEMS questionnaire. This is in line with the 6 domains of usability, according to the general guidelines for usability assessment [12,25]: “Effectiveness,” “Efficiency,” “Satisfaction,” “Learnability,” “Perceived value,” and “Privacy and Security Issues.”

After the selection of the items during the expert meeting, we translated the items that were only available in English into Dutch. We applied a forward-backward translation (English to Dutch) procedure for each item. This procedure was executed by 2 people who were native proficiency speakers of both Dutch

and English (DPN and Stephanie Medlock). A formal assessment of each item’s linguistic complexity using the Common European Framework of Reference for Language was conducted, including translating items as required to B1 level, by an expert that had experience in making patient instructions accessible (Marieke van Maanen) [26,27]. Items from 6 individual (validated) questionnaires were collected (Table S1 in [Multimedia Appendix 1](#)). In addition, insights from the article of the authors Berkman and Karahoca [28] were integrated into the process, as they describe that the change in sensitivity of a scale varies due to the responses, while in human-computer interaction, a scale is expected to be sensitive to the differences between systems instead of people. This insight enriched the questionnaire development with current research findings and best practices in usability metrics. We therefore maintained the item scores consistent with the current scoring methodology across responses. This has resulted in sufficient differentiation at the system level; however, further refinement is required to optimize the scoring of the GEMS.

Step II: Face Validity - Pilot Testing, Item Selection, and Adaptation

We recruited participants to take part in the evaluation of (1) the questionnaire itself, and (2) the evaluation of DHS using the draft GEMS instrument (Figure S1 in [Multimedia Appendix 1](#)). Round I consisted of an appreciative inquiry, to get feedback from stakeholders, to ensure that the instrument reflected their perspectives and values and that questions were understandable

[29]. We presented the questionnaire to the patient panel from the Amsterdam University Medical Center (n=8; Table S2 in [Multimedia Appendix 1](#)). After this round, an expert meeting including all authors (and Thomas Engelsma) was held to make adjustments to the language and wording of the questions.

Step III: Construct Validity - Psychometric Analysis

Round II consisted of the validation of the questionnaire by applying it with users of two self-management tools within the Amsterdam University Medical Center patient portal, which are available from the electronic health record for patients under the nephrology department: (1) entering home measurements of kidney transplant patients' vital statistics such as blood pressure, pulse, and temperature and (2) medication reminders. Patients were included when they participated in home measurements, or in the use of medication reminders, could read and understand the Dutch language, and downloaded the app from the patient portal in order to use one of these functionalities. Patients were invited to participate in this study by their HCP (physician or nurse practitioner). Informed consent of the participants was provided online (e-consent). Patients who agreed to participate were contacted by a researcher (SJO or a supportive researcher) to administer the GEMS questionnaire by email. Data were collected using Castor EDC [30]. Patients who did not return the questionnaire or did not fully complete the questionnaire received a reminder after 2 weeks, and, if necessary, a phone call after 4 weeks. After psychometric analysis, an expert meeting was held to discuss the findings, and if necessary, adjustments were made to the instrument.

Assessing Acceptability

The data from the questionnaire were analyzed using SPSS statistics (version 28.0.1.1, IBM) [31]. Respondents who missed more than one item of the GEMS were removed from the data set. Records missing other data, such as demographics, that were not part of the core of the GEMS questionnaire were not excluded. All items were recoded so that "1" was the most negative value on the Likert scale. In order to be able to perform factor analysis, the questions with scales ranging from "1-10" were recoded to "1 - 5" (1 and 2 were recoded to 1, 3 and 4 recoded to 2, and so on). The question with a Likert scale from "1-7" was recoded to "1 - 5," where the extremes are taken together (1 and 2 were recoded to 1; 6 and 7 were recoded to 5).

The Single Ease Questions (SEQ) is a single-item measure that assesses the complexity of a task for a user, such as entering home blood pressure measurements into the patient portal [32,33]. The SEQ aligns with the main features available in the system [33]. The different tasks that patients have to fulfil for the two separate DHS are difficult to compare, as logging into the system is the only task that is consistent across our analyses. Consequently, in psychometric evaluations, only the question regarding the ease or difficulty of "logging into the system" was included for both DHS assessments. For items where the nonresponse rate reached or exceeded 90%, it was inferred that patients chose not to answer the respective question. Consequently, the item in question was deemed unnecessary and was subsequently removed from the GEMS questionnaire

[34]. With regard to the distribution of item scores, a skewness of 90% was considered to indicate redundancy for inclusion of the item in the GEMS questionnaire [34].

Assessing Construct Validity

An item correlation analysis was performed using the Spearman rank-order correlation coefficient. All items were compared with each other to find inter-item overlap, with a score of $r_s > 0.70$ meaning that there could be singularity. Prior to performing a factor analysis, we tested whether the data set was suitable by assessing the Kaiser-Meyer-Olkin test of sampling adequacy (> 0.60), and Bartlett test of sphericity ($\alpha < .05$) [35,36]. A principal component analysis (PCA) with direct Oblimin rotation was used for factor analysis (FA). In addition, a scree plot was made of the PCA results. The number of values above the scree plateau were taken as the number of factors the items contributed to. In case of no clear scree plateau, a threshold of 1.0 was used.

Assessing Reliability and Internal Consistency

For all factors, extracted with PCA, the reliability and internal consistency were assessed by using the Cronbach α (> 0.70) and item-total correlations (> 0.40). Per factor, the items were dropped one by one to see whether items had to be removed to increase the Cronbach α to the threshold of 0.70. Finally, the items were scrutinised in an expert meeting (SJO, LWDP, DPN, SAN, HJM, and EMAS) using the results of the aforementioned analyses to determine which items were to be dropped and which should remain. In addition, we assigned labels to the constructs.

Results

Step I: Content Validity

In evaluations of DHS, researchers readily access numerous validated questionnaires from the literature, using them as tools for assessing usability and satisfaction in order to improve the product or system. Drawing from our literature review, the SUS is the most widely used usability evaluation instrument in the digital health industry [10,11]. For a long time, it has been a standard procedure to evaluate the usability of digital technology using general benchmarking tools, which has led to the adoption of generic tools like the SUS [11]. However, this questionnaire was developed in the early stages of the human-computer interaction field, at a time when digital health did not yet exist [16,37]. Newer questionnaires in the field such as the MAUQ and eHealth Usability Benchmarking Instrument try to be more specific within their domain; however, these questionnaires are still extensive, not easy to deploy, and using terminology derived from human-computer interaction [11,18]. In addition, as questionnaires such as SUS and Usability Metric for User Experience (UMUX) are primarily designed for software development, they use complex software-related terminology, such as functionalities of a system, that is often not understood by the general population [11].

We excluded statements regarding software interaction due to their complexity, which could potentially hinder understanding. We collected 14 unique statements from the identified questionnaires [12,25]. We chose to incorporate the 4-item UMUX (with Likert scale 1 - 5), along with SEQ (Likert scale 1 - 7). To include learnability, we added a question from the

SUS on whether patients had to learn a lot about the specific DHS before they could use it (Likert scale 1 - 5). Regarding perceived value, we added 2 questions from the MAUQ on whether the DHS contributed to the patient's health, and whether patients had the feeling that the DHS improved health care (both Likert scale 1 - 7). Finally, for perceived value, we added a question from Timmermans et al [38] on whether using the DHS reminded patients of being sick (Likert scale 1 - 5). To assess privacy and security, we added a question from Timmermans et al [38] (Likert scale 1 - 5). Regarding satisfaction, we opted to include the Net Promoter Score (NPS; Likert scale 1 - 10), the Customer Satisfaction Score (CSAT; Likert scale 1 - 5), and continued use, as we aimed to investigate whether satisfaction had an influence on continued use and vice versa (Likert scale 1 - 10). We added demographics such as gender, age, educational level, and health literacy [39,40]. At a later stage, we also added one question on digital literacy. The final GEMS questionnaire for validation consisted of 14 items (Table S4 in [Multimedia Appendix 1](#)).

Step II: Face Validity

In total, 92 patients participated in the validation: 65.2% (n=58) were male, 38% (n=35) were aged between 40 and 59 years, and 32.6% (n=30) had a higher professional education (Table S3 in [Multimedia Appendix 1](#)). A total of 92 patients were included for the psychometric analysis. All items presented to patients had a response rate of over 95%. For item skewness, no score was answered more than 90% for any of the answered questions. In the distribution of scores, we noticed that the highest value not applicable was entered with 10.9% on Q5 (question 5; "Q#" represents the questions involved in this study). The highest missing value with 17.4% was on Q13. No items of the GEMS were removed. Not all patients completed the question about digital literacy as this question was added to the demographics later (n=43). Patients' remarks and suggestions for improvement mainly focused on Q5, with some patients being unfamiliar with the nondigital method of filling in home measurements on paper. Therefore, some patients were unable to fill in this question. In addition, with Q8, patients indicated that the disease process is much more intense for some people than others, and that this question is difficult to answer in the home setting ([Table 1](#)).

Table . Description of each measurement instrument found in explorative literature search.

| Measurement instrument | Abbreviation | Author | Items, n | Population validated | Scale | Reference where questionnaire has been used in health care context |
|---|--------------|---------------------------------|----------|---|--------|--|
| Usability | | | | | | |
| Questionnaire for User Interaction Satisfaction | QUIS | Chin et al [41] | 27 | 150 users | 1 - 9 | [42,43] |
| System Usability Scale | SUS | Brooke [16] | 10 | 184, aimed to include a diverse range of participants | 1 - 5 | [44-46] |
| mHealth App Usability Questionnaire | MAUQ | Zhou et al [18] | 20 | 128, majority included were students with a bachelor's degree | 1 - 7 | [47,48] |
| The Usability Metric for User Experience | UMUX | Finstad [49] | 4 | 255, not extensively described | 1 - 7 | [50] |
| Poststudy System Usability Questionnaire | PSSUQ | Lewis [37] Lewis [51] | 16 | 48, and 210 in second validation study | 1 - 7 | [44,52] |
| Technology Acceptance Model questionnaire | TAM | Davis [53] | 12 | 107 users | 1 - 7 | [54] |
| User version of the Mobile App Rating Scale | uMARS | Stoyanov et al [55] | 20 | 164 young people | 1 - 5 | [56,57] |
| Mobile app rating scale | MARS | Terhorst et al [58] | 23 | 1299 mobile health apps | 1 - 5 | [59] |
| eHealth Usability Benchmarking Instrument | HUBBI | Broekhuis et al [11] | 18 | 148 persons | 1 - 5 | [60] |
| Satisfaction | | | | | | |
| Net Promoter Score | NPS | Reichheld [61] Mekonnen [62] | 1 | Not described | 1 - 10 | [63,64] |
| Client Satisfaction Questionnaire | CSQ-8 | Larsen et al [65] | 8 | Different populations, also in health care setting | 1 - 4 | [66] |
| Patient satisfaction questionnaire III | PSQ-III | Ware et al [67] | 50 | Various populations, in individuals with various medical conditions | 1 - 5 | [68] |
| Other | | | | | | |
| Single Ease Questionnaire | SEQ | Nielsen and Molich [25] | 1 | Not described | 1 - 7 | [69] |

Step III: Construct Validity

Spearman's rank correlation coefficient indicated Q8 as redundant as it showed a negative correlation on almost all items. The calculated UMUX score was also taken into consideration but did not show a significant correlation with items other than its own questions (Q1-Q4). None of the items was extremely skewed. Since none of the items were completed by less than 95% of the respondents, all items were included for psychometric analyses. The data set consisted of 14 items

that were used for psychometric analysis (Table S4 in [Multimedia Appendix 1](#) presents the Dutch original items). Kaiser-Meyer-Olkin was 0.72, and Bartlett Test of Sphericity was $P < .01$. PCA suggested a 5-factor solution. However, the fifth factor had an eigenvalue of 1.05, and we, therefore, decided to not include this factor. Q1 did not load to any factor. Common factor analysis using 4 factors with a factor loading threshold of 0.40 resulted in Q1 and Q5 not loading to any factors. Q7 cross loaded into factors 3 and 4. Q7 was dropped from factor 4 because this lowered the Cronbach α . Q8, Q10, and Q13 were

also dropped because these items lowered the Cronbach α for the respective factor. As shown in [Table 2](#), item-total correlation was considered sufficient (>0.40) for all items. Factors 1 and 3 had the lowest Cronbach α (0.66 and 0.67, respectively) and factors 2 and 4 the highest (0.77 and 0.78, respectively).

Table . Results of the GEMS validation.

| Item description | NA ^a ≥25% | $r_s^b > 0.70$ | CFA ^c loading | ITC ^d | Cronbach α^e |
|--|----------------------|----------------|--------------------------|------------------|---------------------|
| Factor 1: Convenience of use (Cronbach α of scale=0.66; 95% CI^f 0.49 - 0.78) | | | | | |
| Q2: "Using [this DHS] ^g is a frustrating experience." ^h Het is vervelend om [digitale tool] te gebruiken. | — ⁱ | — | 0.85 | 0.52 | — |
| Q6: "I needed to learn a lot of things before I could get going with [this DHS]." ^h Ik moest veel over [digitale tool] leren voordat ik het goed kon gebruiken. | — | — | 0.85 | 0.52 | — |
| Factor 2: Satisfaction (Cronbach α of scale=0.77; 95% CI 0.67 - 0.84) | | | | | |
| Q11: "Overall, how satisfied were you with [DHS]?" ^h Hoe tevreden bent u over digitale tool? | — | — | -0.61 | 0.60 | 0.70 |
| Q12: "How likely is it that you would recommend [DHS] to a friend or colleague?" ^h Hoe waarschijnlijk is het dat u [digitale tool] aan iemand anders die deze zorg nodig heeft aanraadt? | — | — | -0.59 | 0.63 | 0.67 |
| Q14: "I would use [this DHS] again." ^h Hoe waarschijnlijk is het dat u de [digitale tool] blijft gebruiken? | — | — | -0.44 | 0.62 | 0.70 |
| Factor 3: Perceived value (Cronbach α of scale=0.67; 95% CI 0.51 - 0.79) | | | | | |
| Q7: "The [DHS] would be useful for my health and well-being." ^h Het gebruik van [digitale tool] draagt bij aan mijn gezondheid. | — | — | 0.50 | 0.53 | — |
| Q9: "The [DHS] improved my access to health care services." Ik denk dat [digitale tool] de zorg verbetert. | — | — | 0.53 | 0.53 | — |
| Factor 4: Efficiency in use (Cronbach α of scale=0.78; 95% CI 0.67 - 0.86) | | | | | |
| Q3: "[This DHS] is easy to use." ^h [Digitale tool] is makkelijk te gebruiken. | — | — | -0.62 | 0.65 | — |
| Q4: "I have to spend too much time correcting things with [this DHS]." ^h Ik ben te veel tijd kwijt aan het gebruik [van digitale tool]. | — | — | -0.43 | 0.65 | — |

^aNA: "I do not know or not applicable" responses ≥25%.

^b r_s : Spearman rank correlation coefficient between items >0.70.

^cCFA: confirmatory factor analysis

^dITC: item-total correlation.

^eCronbach α of scale if item is deleted.

^fSee Baumgartner and Chung [29].

^gDHS: Digital Health Solution.

^hOriginal English item from questionnaire.

ⁱNot applicable.

After PCA, a collaborate expert meeting was held to determine the most appropriate labels for these factors based on existing usability terminology: convenience of use, perceived value, efficiency of use, and satisfaction. These constructs are known

in the field of human-computer interaction. A more complete definition of the 4 factors applied to the home setting are shown in [Textbox 1](#). The final constructs of the GEMS are outlined in [Figure 2](#).

Textbox 1. Textbox 1. Description of the constructs of the GEMS questionnaire.

Constructs and their explanations

Convenience of use

This highlights the ease and comfort with which users can interact with the digital health solutions at home. Convenience of use is a component of usability, emphasizing aspects that contribute to making the user experience more convenient, pleasant, and smooth [70]. This means tailoring it to fit to patient preferences and expectations for self-management at home.

Perceived value

Perceived value refers to the extent to which a system or product fulfills users' needs and goals, addressing the pragmatic utility it offers to its intended users [70]. It encompasses the relevance and value of the digital health solutions features and functionalities in addressing user requirements in a home setting. In a health care setting, perceived value ultimately determines the practical utility and adoption of the digital health solutions by patients [71,72].

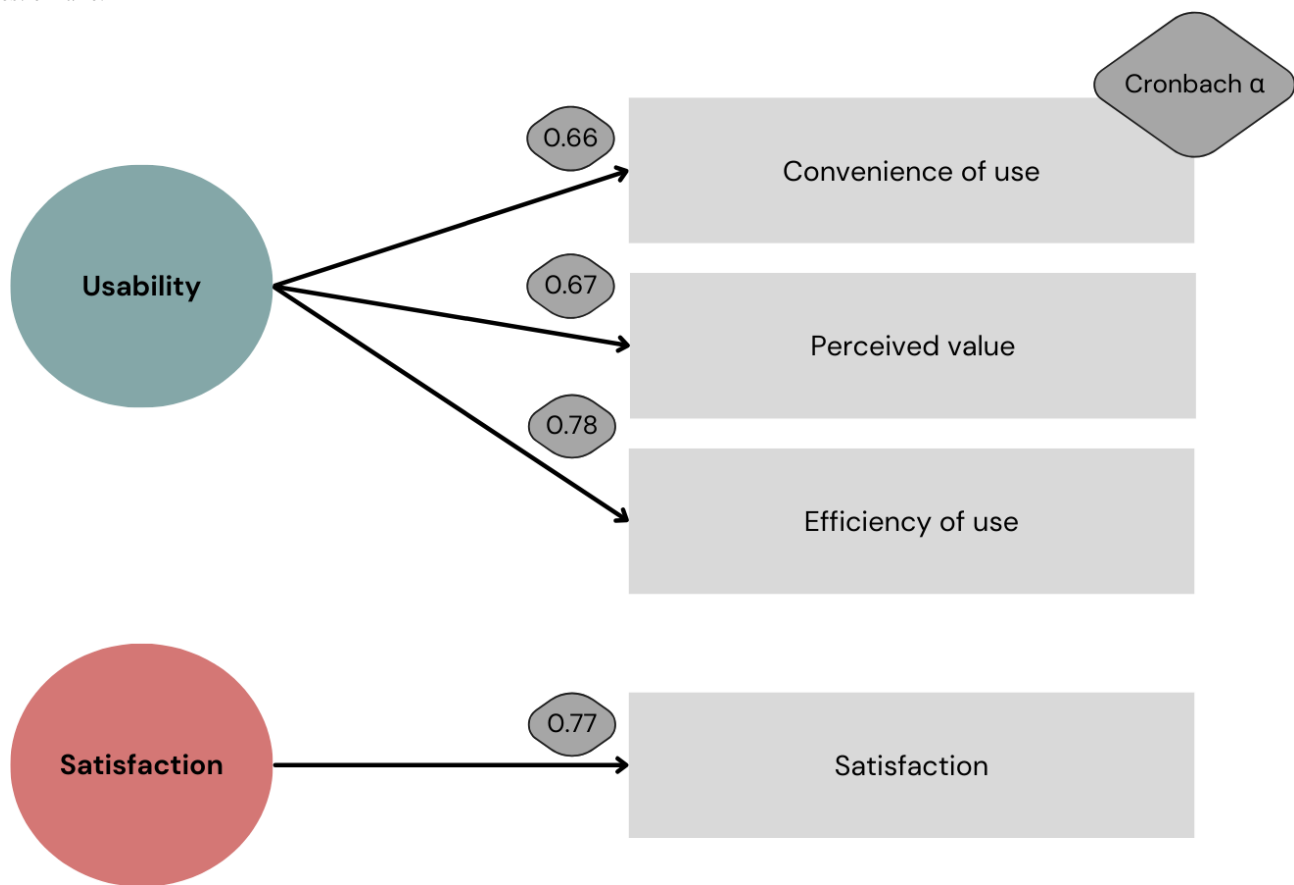
Efficiency of use

In a home setting, efficiency of use highlights how quickly users can perform tasks in a digital health solutions once they are familiar with it. Efficiency of use is influenced by factors such as learnability, memorability, and error prevention, as it pertains to how quickly and effortlessly users can achieve their goals when using a self-management tool in a home setting [12].

Satisfaction

According to International Organization for Standardization 9241, satisfaction is referred to as the degree to which users experience comfort and have positive attitudes toward using the product [12]. For self-management tools, satisfaction goes beyond mere functionality and usability, extending to factors such as efficacy, empowerment, and emotional well-being [73].

Figure 2. Visual abstract of final results and named constructs of Experienced Usability and Satisfaction With Self-Monitoring in the Home Setting Questionnaire.



Discussion

Principal Findings

Our aim, was to develop a steering instrument that enables the measurement of usability and satisfaction at various stages of adoption, with constructs that are relevant for a home setting, adapted to the language proficiency of the general population, and which might serve as a benchmarking instrument for usability and satisfaction with DHS. Following the initial translation phase of this study, it became evident that the items of the GEMS were easy to understand for patients. Although we designed the questionnaire for a broad population, our evaluation revealed that the majority of study participants had a higher level of education. In research, it is a known challenge to reach those with lower health and digital literacy levels for evaluation [74]. The applicability of the DHS varies depending on the specific needs and characteristics of different users. The GEMS questionnaire has been tailored to a B1 language proficiency level, which enhances its accessibility. However, there is a risk of obtaining biased outcomes of the GEMS depending on the demographic profile (eg, age, education, digital literacy, and health literacy) of the respondents. Therefore, collecting these demographic data are essential to understand if DHS users with different profiles assess the experienced usability and satisfaction differently. Gaining these insights may help in ensuring tailORIZATION of the DHS to the user needs based on GEMS outcomes. This necessitates further refinement of the DHS to ensure its suitability across diverse populations.

Internal consistency of the GEMS was sufficient and factor analysis confirmed 4 factors, to which we have assigned the following labels: convenience of use, perceived value, efficiency of use, and satisfaction. Internal consistency of the GEMS, as measured with the Cronbach α , was slightly lower compared with the minimum value of 0.7 [75]. A possible explanation could lie, in our sample characteristics, as several participants also used similar applications, such as smartwatches that provided reminders. This dual usage could have influenced their responses, leading to expressed preferences or aversions towards the usage of medication reminders.

Given that the NPS was integrated into our satisfaction metric within the GEMS questionnaire, we opted to use the raw NPS as a component within our scoring scale. This approach involves incorporating the absolute values of promoters, passives, and detractors, rather than calculating the traditional NPS by subtracting the percentage of detractors from the percentage of promoters [76]. In a manner similar to the SUS questionnaire, we reversed the scales in our questionnaire to enhance reliability and validity. This approach serves several key purposes: (1) mitigating response bias, (2) maintaining participant attention and engagement, (3) ensuring balance and consistency within the questionnaire, and (4) detecting random responses on the questions by participants [16]. For the factors and questions derived from the factor analysis, we carefully examined whether reversed scaling was still present in the questionnaire. We concluded that reversed scaling was still present in 2 out of the 4 constructs.

For the statements in the GEMS questionnaire, we decided to adopt, translate, and adapt the statements from the UMUX and adjust them to using DHS in a home setting. However, in some cases, we have labeled the factors differently from those in the UMUX. Specifically, the statement “It is frustrating to use this digital tool” is classified under “Convenience of use” in the GEMS questionnaire, while it is categorized under “Satisfaction” in the UMUX. The interrelationship with the other questions in GEMS aligns more closely with the definition of convenience. We decided to address the experiences related to the context in which the DHS are used, specifically the deployment of DHS in a home setting. First, the difficulty in using the technology due to lack of digital literacy or misunderstanding of terminology. Second, ease of use, as the primary concern in a home setting is how conveniently the DHS can be integrated into daily routines. In addition, we translated and modified the UMUX question “I spend too much time correcting things with this system” to make it applicable at a higher conceptual level. The revised question no longer concerns the correction of things (errors), but instead evaluates whether the DHS is usable within its intended context [28].

Closing the feedback loop between patients and HCP through the utilization of DHS represents a pivotal strategy in enhancing health care delivery with DHS. By enabling self-management of patients through communication and data exchange, digital tools foster a collaborative environment where patients can actively participate in their care and providers can make informed decisions [77]. Incorporating the GEMS questionnaire as part of a comprehensive evaluation of DHS may enhance usability and satisfaction, contributing to adoption and the overall effectiveness of the DHS in improving health outcomes. The GEMS is therefore of relevance and value to HCPs, decision makers, health insurance companies, and public health institutions. The outcomes of the GEMS can assist these stakeholders to identify important issues as perceived by patients, and to develop strategies to address these issues and improve the quality of their DHS.

Strengths and Limitations

The strength of the GEMS questionnaire lies in the convergence of the four factors: convenience of use, perceived value, efficiency of use, and satisfaction, its concise questionnaire format, its adaptation to the language proficiency of the general population, and its utility as a steering tool as it can be used longitudinally in DHS implementation. The main strength of this study is that we applied a 4-step structured methodology to develop the GEMS questionnaire, consisting of both qualitative and quantitative evaluation phases. We also included 2 functionalities of our electronic health records in our evaluation in order to assure that the GEMS is applicable to a range of self-management tools. One of the limitations of this study is that a subset of patients may have been unable to participate in these (digital) evaluations due to requirements such as internet access, concentration, self-confidence, and

proficient reading skills. We recognize that these evaluations cannot be used without considering potential issues of inequality [78]. According to the literature, this can be due to several reasons. First, the DHS may currently not be usable enough, for instance, by not involving the users during the design phase [79]. Second, health care professionals might be unfamiliar with the technology and not offering these tools to all patients [80]. Third, patients may feel having inadequate knowledge to use these tools [81], or have low (digital) literacy and therefore unable to use the tool [82]. Hence, we recommend further evaluating and refining the GEMS questionnaire in populations characterized by low (digital) literacy. Currently, we are conducting such a validation study within a demographic comprising individuals with low socioeconomic status and chronic obstructive pulmonary disease using a self-management tool. For these groups, we will conduct the evaluation on paper, using concept cards and translating the questions to graphics that visually support the questions [83]. By adopting this method, we aim to facilitate a comprehensive understanding of usability and satisfaction tailored to the needs and preferences of this specific population.

Because we used statements from various questionnaires, during the initial validation phase of the GEMS, some questions had different Likert scales. In order to ensure consistency in the analysis, the scales were converted. As a result, this might impact the interpretation of results, as the participants may interpret and respond to the items differently due to an expanded or contracted range of options [84,85]. Literature supports rescaling of 5- and 7-point scales for comparison, although it is noted that these scales may produce higher mean scores compared with a 10-point scale [84]. Finally, If the GEMS is used in another cultural setting, correct linguistic and cultural translation is needed to ensure content validity [86]. In order to facilitate this, an ongoing study is being conducted to assess a German translation of the GEMS questionnaire.

Conclusion

The GEMS questionnaire, comprising 9 items, has demonstrated its reliability and validity in assessing the usability and satisfaction of DHS within a home environment. It offers valuable insights into patient experiences with self-management tools, covering aspects of convenience of use, perceived value, efficiency of use and satisfaction. This development and validation study has been conducted with patient populations using medication reminders and home measurements. Further refinement is necessary in order to confirm the efficacy and applicability of the GEMS questionnaire in patient populations with low digital literacy. Using the GEMS questionnaire as a steering metric reflects a dedication to improving usability and satisfaction within DHS. In conclusion, the GEMS may promote development of a robust DHS, which enriches experienced usability and satisfaction and augments the efficacy of the DHS, thereby fostering positive health outcomes.

Acknowledgments

The authors would like to thank all experts who participated in the development rounds of the GEMS; Marieke van Maanen for her expertise on the adjustment of the items to the language proficiency of patients; Dr. Stephanie Medlock for contributing to the translation of the items; Hugo van Mens for his expertise on the Usability Metric for User Experience; Thomas Engelsma for the final expert meeting and questionnaire development process; and Ro Glasius for questionnaire support.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Flowchart of the inclusion process, original English or Dutch items and final version of translated Dutch version of GEMS, original GEMS, and demographics of included sample for validation (n=92).

[[PDF File, 221 KB - medinform_v13i1e63703_app1.pdf](#)]

References

1. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. *NPJ Digit Med* 2022 Jan 27;5(1):13. [doi: [10.1038/s41746-022-00557-1](https://doi.org/10.1038/s41746-022-00557-1)] [Medline: [35087160](https://pubmed.ncbi.nlm.nih.gov/35087160/)]
2. Li R, Liang N, Bu F, Hesketh T. The effectiveness of self-management of hypertension in adults using mobile health: systematic review and meta-analysis. *JMIR Mhealth Uhealth* 2020 Mar 27;8(3):e17776. [doi: [10.2196/17776](https://doi.org/10.2196/17776)] [Medline: [32217503](https://pubmed.ncbi.nlm.nih.gov/32217503/)]
3. van de Vijver S, Tensen P, Asiki G, et al. Digital health for all: how digital health could reduce inequality and increase universal health coverage. *Digit Health* 2023;9:20552076231185434. [doi: [10.1177/20552076231185434](https://doi.org/10.1177/20552076231185434)] [Medline: [37434727](https://pubmed.ncbi.nlm.nih.gov/37434727/)]
4. WHO guideline on self-care interventions for health and well-being, 2022 revision. World Health Organization. 2022. URL: <https://www.who.int/publications/i/item/9789240052192> [accessed 2024-12-16]
5. Global strategy on digital health 2020-2025. : World Health Organization; 2021. URL: <https://www.who.int/docs/default-source/documents/gS4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf> [accessed 2024-12-16]
6. Kraus S, Jones P, Kailer N, Weinmann A, Chaparro-Banegas N, Roig-Tierno N. Digital transformation: an overview of the current state of the art of research. *Sage Open* 2021 Jul;11(3):21582440211047576. [doi: [10.1177/21582440211047576](https://doi.org/10.1177/21582440211047576)]
7. Moqri M, Herzog C, Poganik JR, et al. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* 2023 Aug;186(18):3758-3775. [doi: [10.1016/j.cell.2023.08.003](https://doi.org/10.1016/j.cell.2023.08.003)]
8. Maqbool B, Herold S. Potential effectiveness and efficiency issues in usability evaluation within digital health: a systematic literature review. *J Syst Softw* 2024 Feb;208:111881. [doi: [10.1016/j.jss.2023.111881](https://doi.org/10.1016/j.jss.2023.111881)]
9. Jaspers MWM. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. *Int J Med Inform* 2009 May;78(5):340-353. [doi: [10.1016/j.ijmedinf.2008.10.002](https://doi.org/10.1016/j.ijmedinf.2008.10.002)]
10. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. *Int J Med Inform* 2019 Jun;126:95-104. [doi: [10.1016/j.ijmedinf.2019.03.018](https://doi.org/10.1016/j.ijmedinf.2019.03.018)] [Medline: [31029270](https://pubmed.ncbi.nlm.nih.gov/31029270/)]
11. Broekhuis M, van Velsen L, Hermens H. Assessing usability of eHealth technology: a comparison of usability benchmarking instruments. *Int J Med Inform* 2019 Aug;128:24-31. [doi: [10.1016/j.ijmedinf.2019.05.001](https://doi.org/10.1016/j.ijmedinf.2019.05.001)] [Medline: [31160008](https://pubmed.ncbi.nlm.nih.gov/31160008/)]
12. ISO 9241-11:20:2018 ergonomics of human-system interaction - part 11: usability: definitions and concepts. International Organization for Standardization. 2018. URL: <https://www.iso.org/standard/63500.html> [accessed 2024-12-16]
13. Albert B, Tullis T. *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics*: Morgan Kaufmann; 2022.
14. Simola S, Hörhammer I, Xu Y, et al. Patients' experiences of a national patient portal and its usability: cross-sectional survey study. *J Med Internet Res* 2023 Jun 30;25:e45974. [doi: [10.2196/45974](https://doi.org/10.2196/45974)] [Medline: [37389909](https://pubmed.ncbi.nlm.nih.gov/37389909/)]
15. ISO/IEC 25002:2024: systems and software engineering — systems and software quality requirements and evaluation (square) — quality model overview and usage. International Organization for Standardization. 2024. URL: <https://www.iso.org/standard/78175.html> [accessed 2024-12-16]
16. Brooke J. *SUS-a quick and dirty usability scale*. In: *Usability Evaluation In Industry*: Taylor & Francis; 1996:189-194.
17. Soltanzadeh L, Babazadeh Sangar A, Majidzadeh K. The review of usability evaluation methods on tele health or telemedicine systems. *Front Health Inform* 2022;11(1):112. [doi: [10.30699/fhi.v11i1.357](https://doi.org/10.30699/fhi.v11i1.357)]
18. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. *JMIR Mhealth Uhealth* 2019 Apr 11;7(4):e11500. [doi: [10.2196/11500](https://doi.org/10.2196/11500)] [Medline: [30973342](https://pubmed.ncbi.nlm.nih.gov/30973342/)]
19. Keogh A, Brennan C, Johnston W, et al. Six-month pilot testing of a digital health tool to support effective self-care in people with heart failure: mixed methods study. *JMIR Form Res* 2024 Mar 1;8:e52442. [doi: [10.2196/52442](https://doi.org/10.2196/52442)] [Medline: [38427410](https://pubmed.ncbi.nlm.nih.gov/38427410/)]

20. Smith B, Magnani JW. New technologies, new disparities: the intersection of electronic health and digital health literacy. *Int J Cardiol* 2019 Oct 1;292:280-282. [doi: [10.1016/j.ijcard.2019.05.066](https://doi.org/10.1016/j.ijcard.2019.05.066)] [Medline: [31171391](https://pubmed.ncbi.nlm.nih.gov/31171391/)]
21. Albert NM, Dinesen B, Spindler H, et al. Factors associated with telemonitoring use among patients with chronic heart failure. *J Telemed Telecare* 2017 Feb;23(2):283-291. [doi: [10.1177/1357633X16630444](https://doi.org/10.1177/1357633X16630444)] [Medline: [26869144](https://pubmed.ncbi.nlm.nih.gov/26869144/)]
22. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. *Front Dig Health* 2023;5:1264780. [doi: [10.3389/fdgth.2023.1264780](https://doi.org/10.3389/fdgth.2023.1264780)] [Medline: [38046643](https://pubmed.ncbi.nlm.nih.gov/38046643/)]
23. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. *NPJ Digit Med* 2019;2(1):38. [doi: [10.1038/s41746-019-0111-3](https://doi.org/10.1038/s41746-019-0111-3)] [Medline: [31304384](https://pubmed.ncbi.nlm.nih.gov/31304384/)]
24. White J. *PubMed 2.0. Med Ref Serv Q* 2020;39(4):382-387. [doi: [10.1080/02763869.2020.1826228](https://doi.org/10.1080/02763869.2020.1826228)] [Medline: [33085945](https://pubmed.ncbi.nlm.nih.gov/33085945/)]
25. Nielsen J, Molich R. Heuristic evaluation of user interfaces. Presented at: CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Apr 1-5, 1990; Seattle, WA. [doi: [10.1145/97243.97281](https://doi.org/10.1145/97243.97281)]
26. Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*: Cambridge University Press; 2001.
27. Figueras N. The CEFR, a lever for the improvement of language professionals in Europe. *The Mod Lang J* 2007 Dec;91(4):673-675. [doi: [10.1111/j.1540-4781.2007.00627_8.x](https://doi.org/10.1111/j.1540-4781.2007.00627_8.x)]
28. Berkman MI, Karahoca D. Re-assessing the Usability Metric for User Experience (UMUX) scale. *J Usability Stud* 2016;11(3) [[FREE Full text](#)]
29. Baumgartner TA, Chung H. Confidence Limits for Intraclass Reliability Coefficients. *Meas Phys Educ Exerc Sci* 2001 Sep;5(3):179-188. [doi: [10.1207/S15327841MPEE0503_4](https://doi.org/10.1207/S15327841MPEE0503_4)]
30. Castor electronic data capture. Castor EDC. 2019. URL: <https://castoredc.com> [accessed 2024-11-17]
31. SPSS Statistics 28.0.0 - IBM documentation. IBM. URL: <https://www.ibm.com/docs/en/spss-statistics/28.0.0> [accessed 2024-12-16]
32. Sauro J, Dumas JS. Comparison of three one-question, post-task usability questionnaires. Presented at: CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Apr 4-9, 2009; Boston, MA. [doi: [10.1145/1518701.1518946](https://doi.org/10.1145/1518701.1518946)]
33. Sauro J, Lewis JR. *Quantifying the User Experience: Practical Statistics for User Research*: Morgan Kaufmann; 2016. [doi: [10.1016/B978-0-12-802308-2.00002-3](https://doi.org/10.1016/B978-0-12-802308-2.00002-3)]
34. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*: Cambridge University Press; 2011.
35. Bartlett MS. Tests of significance in factor analysis. *Br J Stat Psychol* 1950 Jun;3(2):77-85. [doi: [10.1111/j.2044-8317.1950.tb00285.x](https://doi.org/10.1111/j.2044-8317.1950.tb00285.x)]
36. Kaiser HF, Michael WB. Little Jiffy factor scores and domain validities. *Educ Psychol Meas* 1977 Jul;37(2):363-365. [doi: [10.1177/001316447703700210](https://doi.org/10.1177/001316447703700210)]
37. Lewis JR. Psychometric evaluation of the post-study system usability questionnaire: the PSSUQ. *Proc Hum Factors Soc Annu Meet* 1992 Oct;36(16):1259-1260. [doi: [10.1177/154193129203601617](https://doi.org/10.1177/154193129203601617)]
38. Timmermans I, Meine M, Szendey I, et al. Remote monitoring of implantable cardioverter defibrillators: patient experiences and preferences for follow-up. *Pacing Clin Electrophysiol* 2019 Feb;42(2):120-129. [doi: [10.1111/pace.13574](https://doi.org/10.1111/pace.13574)] [Medline: [30536931](https://pubmed.ncbi.nlm.nih.gov/30536931/)]
39. Chew LD, Griffin JM, Partin MR, et al. Validation of screening questions for limited health literacy in a large VA outpatient population. *J Gen Intern Med* 2008 May;23(5):561-566. [doi: [10.1007/s11606-008-0520-5](https://doi.org/10.1007/s11606-008-0520-5)] [Medline: [18335281](https://pubmed.ncbi.nlm.nih.gov/18335281/)]
40. Franssen MP, Van Schaik TM, Twickler TB, Essink-Bot ML. Applicability of internationally available health literacy measures in the Netherlands. *J Health Commun* 2011;16 Suppl 3:134-149. [doi: [10.1080/10810730.2011.604383](https://doi.org/10.1080/10810730.2011.604383)] [Medline: [21951248](https://pubmed.ncbi.nlm.nih.gov/21951248/)]
41. Chin JP, Diehl VA, Norman LK. Development of an instrument measuring user satisfaction of the human-computer interface. Presented at: CHI '88: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; May 15-19, 1988; Washington, DC. [doi: [10.1145/57167.57203](https://doi.org/10.1145/57167.57203)]
42. Salmani H, Nahvijou A, Sheikhtaheri A. Smartphone-based application for self-management of patients with colorectal cancer: development and usability evaluation. *Support Care Cancer* 2022 Apr;30(4):3249-3258. [doi: [10.1007/s00520-021-06754-0](https://doi.org/10.1007/s00520-021-06754-0)] [Medline: [34984548](https://pubmed.ncbi.nlm.nih.gov/34984548/)]
43. Dinari F, Bahaadinbeigy K, Moulai K, Nemati A, Ershad Sarabi R. Designing and evaluating a mobile-based self-care application for patients with gastrointestinal cancer to manage chemotherapy side effects. *Med J Islam Repub Iran* 2022;36:14. [doi: [10.47176/mjiri.36.14](https://doi.org/10.47176/mjiri.36.14)] [Medline: [35991156](https://pubmed.ncbi.nlm.nih.gov/35991156/)]
44. Bhanvadia SB, Brar MS, Delavar A, et al. Assessing usability of smartwatch digital health devices for home blood pressure monitoring among glaucoma patients. *Informatics (MDPI)* 2022 Dec;9(4):79. [doi: [10.3390/informatics9040079](https://doi.org/10.3390/informatics9040079)] [Medline: [36873830](https://pubmed.ncbi.nlm.nih.gov/36873830/)]
45. Stamm-Balderjahn S, Bernert S, Rossek S. Promoting patient self-management following cardiac rehabilitation using a web-based application: a pilot study. *D Health* 2023;9:20552076231211546. [doi: [10.1177/20552076231211546](https://doi.org/10.1177/20552076231211546)] [Medline: [37954686](https://pubmed.ncbi.nlm.nih.gov/37954686/)]

46. Bostrøm K, Børøsund E, Varsi C, et al. Digital self-management in support of patients living with chronic pain: feasibility pilot study. *JMIR Form Res* 2020 Oct 23;4(10):e23893. [doi: [10.2196/23893](https://doi.org/10.2196/23893)] [Medline: [33094734](https://pubmed.ncbi.nlm.nih.gov/33094734/)]
47. Moorthy P, Weinert L, Harms BC, Anders C, Siegel F. German version of the mHealth app usability questionnaire in a cohort of patients with cancer: translation and validation study. *JMIR Hum Factors* 2023 Nov 1;10:e51090. [doi: [10.2196/51090](https://doi.org/10.2196/51090)] [Medline: [37910144](https://pubmed.ncbi.nlm.nih.gov/37910144/)]
48. Fedkov D, Berghofen A, Weiss C, et al. Efficacy and safety of a mobile app intervention in patients with inflammatory arthritis: a prospective pilot study. *Rheumatol Int* 2022 Dec;42(12):2177-2190. [doi: [10.1007/s00296-022-05175-4](https://doi.org/10.1007/s00296-022-05175-4)] [Medline: [36112186](https://pubmed.ncbi.nlm.nih.gov/36112186/)]
49. Finstad K. The usability metric for user experience. *Interact Comput* 2010 Sep;22(5):323-327. [doi: [10.1016/j.intcom.2010.04.004](https://doi.org/10.1016/j.intcom.2010.04.004)]
50. Sobnath D, Philip N, Kayyali R, Nabhani-Gebara S, Pierscionek B, Raptopoulos A. Mobile self-management application for COPD patients with comorbidities: ausability study. Presented at: 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom); Sep 14-16, 2016; Munich, Germany. [doi: [10.1109/HealthCom.2016.7749502](https://doi.org/10.1109/HealthCom.2016.7749502)]
51. Lewis JR. Psychometric evaluation of the PSSUQ using data from five years of usability studies. *Int J Hum Comput Interact* 2002 Sep;14(3-4):463-488. [doi: [10.1080/10447318.2002.9669130](https://doi.org/10.1080/10447318.2002.9669130)]
52. Bakogiannis C, Tsarouchas A, Mouselimis D, et al. A patient-oriented app (ThessHF) to improve self-care quality in heart failure: from evidence-based design to pilot study. *JMIR Mhealth Uhealth* 2021 Apr 13;9(4):e24271. [doi: [10.2196/24271](https://doi.org/10.2196/24271)] [Medline: [33847599](https://pubmed.ncbi.nlm.nih.gov/33847599/)]
53. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 1989 Sep;13(3):319. [doi: [10.2307/249008](https://doi.org/10.2307/249008)]
54. Greer DB, Abel WM. Exploring feasibility of mHealth to manage hypertension in rural black older adults: a convergent parallel mixed method study. *Pat Prefer Adherence* 2022;16:2135-2148. [doi: [10.2147/PPA.S361032](https://doi.org/10.2147/PPA.S361032)] [Medline: [35999840](https://pubmed.ncbi.nlm.nih.gov/35999840/)]
55. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user version of the Mobile Application Rating Scale (uMARS). *JMIR Mhealth Uhealth* 2016 Jun 10;4(2):e72. [doi: [10.2196/mhealth.5849](https://doi.org/10.2196/mhealth.5849)] [Medline: [27287964](https://pubmed.ncbi.nlm.nih.gov/27287964/)]
56. Agher D, Sedki K, Despres S, Albinet JP, Jaulent MC, Tsopra R. Encouraging behavior changes and preventing cardiovascular diseases using the prevent connect mobile health app: conception and evaluation of app quality. *J Med Internet Res* 2022 Jan 20;24(1):e25384. [doi: [10.2196/25384](https://doi.org/10.2196/25384)] [Medline: [35049508](https://pubmed.ncbi.nlm.nih.gov/35049508/)]
57. Wong W, Ming D, Pateras S, et al. Outcomes of end-user testing of a care coordination mobile app with families of children with special health care needs: simulation study. *JMIR Form Res* 2023 Aug 28;7:e43993. [doi: [10.2196/43993](https://doi.org/10.2196/43993)] [Medline: [37639303](https://pubmed.ncbi.nlm.nih.gov/37639303/)]
58. Terhorst Y, Philippi P, Sander LB, et al. Validation of the Mobile Application Rating Scale (MARS). *PLoS ONE* 2020;15(11):e0241480. [doi: [10.1371/journal.pone.0241480](https://doi.org/10.1371/journal.pone.0241480)]
59. Oakley-Girvan I, Yunis R, Fonda SJ, et al. Usability evaluation of mobile phone technologies for capturing cancer patient-reported outcomes and physical functions. *Dig Health* 2023;9:20552076231186515. [doi: [10.1177/20552076231186515](https://doi.org/10.1177/20552076231186515)] [Medline: [37456127](https://pubmed.ncbi.nlm.nih.gov/37456127/)]
60. Barbarossa F, Amabili G, Margaritini A, et al. Design, development, and usability evaluation of a dashboard for supporting formal caregivers in managing people with dementia. Presented at: PETRA '23: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments; Jul 5-7, 2023; Corfu, Greece. [doi: [10.1145/3594806.3594820](https://doi.org/10.1145/3594806.3594820)]
61. Reichheld FF. The one number you need to grow. *Harv Bus Rev* 2003 Dec;81(12):46-54. [Medline: [14712543](https://pubmed.ncbi.nlm.nih.gov/14712543/)]
62. Mekonnen A. The ultimate question: driving good profits and true growth. *J Target Meas Anal Mark* 2006 Jul;14(4):369-370. [doi: [10.1057/palgrave.jt.5740195](https://doi.org/10.1057/palgrave.jt.5740195)]
63. Jonker LT, Plas M, de Bock GH, Buskens E, van Leeuwen BL, Lahr MMH. Remote home monitoring of older surgical cancer patients: perspective on study implementation and feasibility. *Ann Surg Oncol* 2021 Jan;28(1):67-78. [doi: [10.1245/s10434-020-08705-1](https://doi.org/10.1245/s10434-020-08705-1)] [Medline: [32602060](https://pubmed.ncbi.nlm.nih.gov/32602060/)]
64. Gelbman BD, Reed CR. An integrated, multimodal, digital health solution for chronic obstructive pulmonary disease: prospective observational pilot study. *JMIR Form Res* 2022 Mar 17;6(3):e34758. [doi: [10.2196/34758](https://doi.org/10.2196/34758)] [Medline: [35142291](https://pubmed.ncbi.nlm.nih.gov/35142291/)]
65. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD. Assessment of client/patient satisfaction: development of a general scale. *Eval Program Plann* 1979;2(3). [doi: [10.1016/0149-7189\(79\)90094-6](https://doi.org/10.1016/0149-7189(79)90094-6)]
66. Shiraiishi M, Kamo T, Kumazawa R, et al. A multicenter, prospective, observational study to assess the satisfaction of an integrated digital platform of online medical care and remote patient monitoring in Parkinson's disease. *Neurology & Clinical Neurosc* 2023 May;11(3):152-163. [doi: [10.1111/ncn3.12709](https://doi.org/10.1111/ncn3.12709)]
67. Ware JE, Snyder MK, Wright WR, Davies AR. Defining and measuring patient satisfaction with medical care. *Eval Program Plann* 1983;6(3-4):247-263. [doi: [10.1016/0149-7189\(83\)90005-8](https://doi.org/10.1016/0149-7189(83)90005-8)] [Medline: [10267253](https://pubmed.ncbi.nlm.nih.gov/10267253/)]
68. Temple-Oberle C, Yakaback S, Webb C, Assadzadeh GE, Nelson G. Effect of smartphone app postoperative home monitoring after oncologic surgery on quality of recovery: a randomized clinical trial. *JAMA Surg* 2023 Jul 1;158(7):693-699. [doi: [10.1001/jamasurg.2023.0616](https://doi.org/10.1001/jamasurg.2023.0616)] [Medline: [37043216](https://pubmed.ncbi.nlm.nih.gov/37043216/)]

69. Nautiyal S, Shrivastava A. Designing a WhatsApp inspired healthcare application for older adults: a focus on ease of use. *Comp Hum Interact Res Appl* 2023. [doi: [10.1007/978-3-031-49368-3_9](https://doi.org/10.1007/978-3-031-49368-3_9)]
70. Shaw N, Sergueeva K. Convenient or useful? consumer adoption of smartphones for mobile commerce. 2016 Presented at: DIGIT 2016 Proceedings; Dec 11, 2016; Dublin, Ireland.
71. Jo HS, Jung SM. Factors influencing use of smartphone applications for healthcare self-management: an extended technology acceptance model. *Korean J Health Educ Promot* 2014 Oct 1;31(4):25-36. [doi: [10.14367/kjhhep.2014.31.4.25](https://doi.org/10.14367/kjhhep.2014.31.4.25)]
72. Cho H, Porras T, Flynn G, Schnall R. Usability of a consumer health informatics tool following completion of a clinical trial: focus group study. *J Med Internet Res* 2020 Jun 15;22(6):e17708. [doi: [10.2196/17708](https://doi.org/10.2196/17708)] [Medline: [32538796](https://pubmed.ncbi.nlm.nih.gov/32538796/)]
73. Peters D, Calvo RA, Ryan RM. Designing for motivation, engagement and wellbeing in digital experience. *Front Psychol* 2018;9:797. [doi: [10.3389/fpsyg.2018.00797](https://doi.org/10.3389/fpsyg.2018.00797)] [Medline: [29892246](https://pubmed.ncbi.nlm.nih.gov/29892246/)]
74. Arias López MDP, Ong BA, Borrat Frigola X, et al. Digital literacy as a new determinant of health: a scoping review. *PLOS Dig Health* 2023 Oct;2(10):e0000279. [doi: [10.1371/journal.pdig.0000279](https://doi.org/10.1371/journal.pdig.0000279)] [Medline: [37824584](https://pubmed.ncbi.nlm.nih.gov/37824584/)]
75. Gliem JA, Gliem RR. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. Presented at: Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education; Mar 27-29, 2003; Columbus, OH URL: <https://scholarworks.indianapolis.iu.edu/items/63734e75-1604-45b6-aed8-40ddd7036ee> [accessed 2024-12-16]
76. Adams C, Walpola R, Schembri AM, Harrison R. The ultimate question? Evaluating the use of Net Promoter Score in healthcare: a systematic review. *Health Expect* 2022 Oct;25(5):2328-2339. [doi: [10.1111/hex.13577](https://doi.org/10.1111/hex.13577)] [Medline: [35985676](https://pubmed.ncbi.nlm.nih.gov/35985676/)]
77. Weinhold I, Gastaldi L. From shared decision making to patient engagement in health care processes: the role of digital technologies. In: *Challenges and Opportunities in Health Care Management*; Springer; 2014:185-196. [doi: [10.1007/978-3-319-12178-9_15](https://doi.org/10.1007/978-3-319-12178-9_15)]
78. Goedhart NS, Verdonk P, Dedding C. "Never good enough." A situated understanding of the impact of digitalization on citizens living in a low socioeconomic position. *Pol & Int* 2022 Dec;14(4):824-844. [doi: [10.1002/poi3.315](https://doi.org/10.1002/poi3.315)]
79. Reynoldson C, Stones C, Allsop M, et al. Assessing the quality and usability of smartphone apps for pain self-management. *Pain Med* 2014 Jun;15(6):898-909. [doi: [10.1111/pme.12327](https://doi.org/10.1111/pme.12327)] [Medline: [24422990](https://pubmed.ncbi.nlm.nih.gov/24422990/)]
80. Cher BP, Kembhavi G, Toh KY, et al. Understanding the attitudes of clinicians and patients toward a self-management eHealth tool for atrial fibrillation: qualitative study. *JMIR Hum Factors* 2020 Sep 17;7(3):e15492. [doi: [10.2196/15492](https://doi.org/10.2196/15492)] [Medline: [32940611](https://pubmed.ncbi.nlm.nih.gov/32940611/)]
81. Halim NAA, Sopri NHA, Wong YY, Mustafa QM, Lean QY. Patients' perception towards chronic disease self-management and its program: a cross-sectional survey. *Chron Illn* 2023 Jul 4;17423953231185385. [doi: [10.1177/17423953231185385](https://doi.org/10.1177/17423953231185385)] [Medline: [37403449](https://pubmed.ncbi.nlm.nih.gov/37403449/)]
82. Marklund S, Tistad M, Lundell S, et al. Experiences and factors affecting usage of an ehealth tool for self-management among people with chronic obstructive pulmonary disease: qualitative study. *J Med Internet Res* 2021 Apr 30;23(4):e25672. [doi: [10.2196/25672](https://doi.org/10.2196/25672)] [Medline: [33929327](https://pubmed.ncbi.nlm.nih.gov/33929327/)]
83. Agúndez Del Castillo R, Ferro L, Silva E. The use of digital technologies in the co-creation process of photo elicitation. *QRJ* 2024. [doi: [10.1108/QRJ-06-2023-0101](https://doi.org/10.1108/QRJ-06-2023-0101)]
84. Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Int J Market Res* 2008 Jan;50(1):61-104. [doi: [10.1177/147078530805000106](https://doi.org/10.1177/147078530805000106)]
85. Oylum K, Arslan F. Impact of the number of scale points on data characteristics and respondents' evaluations: an experimental design approach using 5-point and 7-point Likert type scales. *J Fac Pol Sci* 2016;55:1-20. [doi: [10.2196/17708](https://doi.org/10.2196/17708)]
86. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* 2000 Dec 15;25(24):3186-3191. [doi: [10.1097/00007632-200012150-00014](https://doi.org/10.1097/00007632-200012150-00014)] [Medline: [11124735](https://pubmed.ncbi.nlm.nih.gov/11124735/)]

Abbreviations

- DHS:** digital health solutions
- HCP:** health care professional
- MAUQ:** mHealth App Usability Questionnaire
- PCA:** principal component analysis
- SEQ:** Single Ease Questions
- SUS:** System Usability Scale
- UMUX:** Usability Metric for User Experience

Edited by J Hefner; submitted 27.06.24; peer-reviewed by G Parmar, M Wright; revised version received 04.10.24; accepted 06.10.24; published 08.01.25.

Please cite as:

Oudbier SJ, Smets EMA, Nieuwkerk PT, Neal DP, Nurmohamed SA, Meij HJ, Dusseljee-Peute LW

Patients' Experienced Usability and Satisfaction With Digital Health Solutions in a Home Setting: Instrument Validation Study

JMIR Med Inform 2025;13:e63703

URL: <https://medinform.jmir.org/2025/1/e63703>

doi: [10.2196/63703](https://doi.org/10.2196/63703)

© Susan J Oudbier, Ellen M A Smets, Pythia T Nieuwkerk, David P Neal, S Azam Nurmohamed, Hans J Meij, Linda W Dusseljee-Peute. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 8.1.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Autonomous International Classification of Diseases Coding Using Pretrained Language Models and Advanced Prompt Learning Techniques: Evaluation of an Automated Analysis System Using Medical Text

Yan Zhuang^{1*}, PhD; Junyan Zhang^{1*}, ME; Xiuxing Li², PhD; Chao Liu³, PhD; Yue Yu³, BS; Wei Dong⁴, PhD; Kunlun He¹, PhD

¹Medical Big Data Research Center, Chinese PLA General Hospital, Beijing, China

²School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

³Digital Health China Technologies Co Ltd, Beijing, China

⁴Senior Department of Cardiology, The Sixth Medical Center of PLA General Hospital, Beijing, China

*these authors contributed equally

Corresponding Author:

Kunlun He, PhD

Medical Big Data Research Center

Chinese PLA General Hospital

28 Fuxing Road

Beijing, 100853

China

Phone: 86 13911232619

Email: kunlunhe@plagh.org

Abstract

Background: Machine learning models can reduce the burden on doctors by converting medical records into International Classification of Diseases (ICD) codes in real time, thereby enhancing the efficiency of diagnosis and treatment. However, it faces challenges such as small datasets, diverse writing styles, unstructured records, and the need for semimanual preprocessing. Existing approaches, such as naive Bayes, Word2Vec, and convolutional neural networks, have limitations in handling missing values and understanding the context of medical texts, leading to a high error rate. We developed a fully automated pipeline based on the Key–bidirectional encoder representations from transformers (BERT) approach and large-scale medical records for continued pretraining, which effectively converts long free text into standard ICD codes. By adjusting parameter settings, such as mixed templates and soft verbalizers, the model can adapt flexibly to different requirements, enabling task-specific prompt learning.

Objective: This study aims to propose a prompt learning real-time framework based on pretrained language models that can automatically label long free-text data with ICD-10 codes for cardiovascular diseases without the need for semiautomatic preprocessing.

Methods: We integrated 4 components into our framework: a medical pretrained BERT, a keyword filtration BERT in a functional order, a fine-tuning phase, and task-specific prompt learning utilizing mixed templates and soft verbalizers. This framework was validated on a multicenter medical dataset for the automated ICD coding of 13 common cardiovascular diseases (584,969 records). Its performance was compared against robustly optimized BERT pretraining approach, extreme language network, and various BERT-based fine-tuning pipelines. Additionally, we evaluated the framework's performance under different prompt learning and fine-tuning settings. Furthermore, few-shot learning experiments were conducted to assess the feasibility and efficacy of our framework in scenarios involving small- to mid-sized datasets.

Results: Compared with traditional pretraining and fine-tuning pipelines, our approach achieved a higher micro-F1-score of 0.838 and a macro-area under the receiver operating characteristic curve (macro-AUC) of 0.958, which is 10% higher than other methods. Among different prompt learning setups, the combination of mixed templates and soft verbalizers yielded the best performance. Few-shot experiments showed that performance stabilized and the AUC peaked at 500 shots.

Conclusions: These findings underscore the effectiveness and superior performance of prompt learning and fine-tuning for subtasks within pretrained language models in medical practice. Our real-time ICD coding pipeline efficiently converts detailed medical free text into standardized labels, offering promising applications in clinical decision-making. It can assist doctors unfamiliar with the ICD coding system in organizing medical record information, thereby accelerating the medical process and enhancing the efficiency of diagnosis and treatment.

(*JMIR Med Inform* 2025;13:e63020) doi:[10.2196/63020](https://doi.org/10.2196/63020)

KEYWORDS

BERT; bidirectional encoder representations from transformers; pretrained language models; prompt learning; ICD; International Classification of Diseases; cardiovascular disease; few-shot learning; multicenter medical data

Introduction

Background

The *International Classification of Diseases, 10th Revision (ICD-10)*, is a universally recognized diagnostic categorization system widely used in medical insurance reimbursements, health reporting, mortality assessments, and related fields [1]. The *ICD-10*'s automatic coding mechanism enables rapid and accurate classification and statistical analysis of medical data, offering a scientific foundation for effective hospital administration and decision-making. In addition, the *ICD-10* automatic coding system accelerates disease diagnosis and treatment planning for medical practitioners, thereby improving medical efficacy and quality. Compared with the original *ICD* code, *ICD-10* provides over 14,000 distinct disease codes (in contrast to the thousands in *ICD-9*), enabling more detailed disease classification. This comprehensive system offers clinicians enhanced patient information, supporting the development of more precise treatment plans and care programs, ultimately improving the quality of care and patient satisfaction. Moreover, as an internationally standardized code, *ICD-10* is essential for global public health surveillance, epidemiological research, and international medical cooperation. Consequently, ensuring accurate *ICD* coding remains a critical priority in clinical practice.

In hospital settings, the assignment of *ICD* codes to unstructured clinical narratives in medical records is a manual task performed by skilled medical coders based on the attending physician's clinical diagnosis. Despite its critical importance, this process is often hindered by inefficiencies such as time consumption, susceptibility to errors, and high costs. Additionally, manual coding cannot always ensure the accuracy of *ICD* codes due to the complexity of code assignment, which requires a thorough consideration of the patient's overall health condition, including medical history, coexisting conditions, complications, surgical interventions, and specialized diagnostic procedures [2,3].

Machine Learning Techniques

The need to enhance efficiency and reduce errors has driven the development of various machine learning techniques to automate the medical *ICD* coding process. These techniques can be broadly classified into 4 main categories: rule-based systems [4,5], traditional supervised algorithms [6,7], gate unit-based deep learning approaches [7-9], and pretrained language models (PLMs) [9-19].

First, rule-based systems for automatic *ICD* coding rely on the creation of explicit rules and knowledge bases to map medical records to the appropriate *ICD* codes [4,5]. Although these approaches have been used for decades and have provided a foundation for more advanced techniques, they are limited by their lack of adaptability and scalability.

Second, traditional supervised algorithms, such as gradient-boosted trees, have been utilized for *ICD* coding due to their efficiency in handling large-scale, high-dimensional datasets. These algorithms rely on semistructured preprocessing, which involves organizing and refining semistructured data into a format suitable for analysis [6,7]. For example, Diao et al [6] developed a light gradient boosting machine-based pipeline for automatically coding 168 primary diagnosis *ICD-10* codes from discharge records and procedure texts, achieving an accuracy of 95.2%. Another study integrated long short-term memory networks with attention mechanisms to predict mortality in ICU patients using electronic health records, achieving significantly higher area under the receiver operating characteristic curve (AUC) scores compared with traditional statistical models and stand-alone long short-term memory networks [7].

Third, PLMs are neural network models with fixed architectures trained on large corpora, which can be fine-tuned for specific downstream tasks such as question answering and entity recognition [10-13]. A notable example is bidirectional encoder representations from transformers (BERT), a prominent PLM designed to learn deep bidirectional representations from large-scale unlabeled text data. BERT effectively captures semantic relationships in clinical records and can be easily adapted to various natural language processing (NLP) tasks through task-specific layers [13]. Coutinho and Martins [14] proposed a BERT model with a fine-tuning method for automatic *ICD-10* coding of death certificates based on free-text descriptions and associated documents. Additionally, Yan et al [15] introduced RadBERT, an ensemble model combining BERT-base, Clinical-BERT, the robustly optimized BERT pretraining approach (RoBERTa), and BioMed-RoBERTa adapted for radiology. Liu et al [16] evaluated RadBERT across 3 NLP tasks: abnormal sentence classification, report coding, and report summarization, demonstrating significantly better performance compared with existing transformer language models. Unstructured patient-generated health data can be leveraged to support clinical decision-making, remote monitoring, and self-care, including medication adherence and chronic disease management. By applying named entity recognition and customizable information extraction methods

based on medical ontologies, NLP models can extract a wide range of clinical information from unstructured patient-generated health data, even in low-resource settings with limited patient notes or training data [17]. Textual analysis presents numerous opportunities for future medical applications. It can aid in extracting information from various sources of medical data, such as clinical reports, nursing notes, scientific literature, and user-generated content. Additionally, vector-based representation methods can transform textual data within clinical documents into formats suitable for machine learning and can be applied to sequence modeling tasks, including sentiment analysis [18].

Finally, XLNet is another type of PLM that captures both forward and backward contexts of text [19]. It combines the advantages of autoregressive models and autoencoding models while overcoming their limitations. XLNet utilizes a permutation-based objective function that maximizes the expected likelihood of a text across all possible word orderings. It also incorporates the Transformer-XL (Transformer-Extra-Long) architecture, which enables long-term dependency modeling and improved memory efficiency. XLNet has been shown to outperform BERT and other baseline models on several natural language understanding tasks.

Prompt Engineering Techniques

By contrast, prompt engineering is a technique that involves the careful construction of prompts or inputs for artificial intelligence models to improve their performance on specific tasks. This technique includes selecting appropriate words, phrases, symbols, and formats to guide a large language model in generating high-quality and relevant text. Numerous studies have used prompts for model tuning to bridge the gap between pretraining objectives and downstream tasks, demonstrating that both discrete and continuous prompts can improve performance in few-shot and zero-shot tasks [20,21]. Furthermore, this technique within PLMs has been shown to outperform fine-tuning in various clinical decision-making tasks [22]. It has the advantage of requiring less data and computational resources, making it especially suitable for clinical settings.

There are 2 primary categories of prompting methods: hard prompts and soft prompts [22-25]. Hard prompts involve using an actual text string as the prompt and include methods that automatically search for templates within a discrete space, such as mining-based, paraphrasing-based, and gradient-based approaches [26-28]. The advantages of hard prompts are interpretability, portability, flexibility, and simplicity. However, designing effective prompts for specific tasks requires significant effort and creativity.

Soft prompts, by contrast, are learnable tensors concatenated with the input embeddings and can be optimized for a given dataset. The main advantage of soft prompts is their ability to achieve better performance than hard prompts by adapting to the model and the data. However, they are not human-readable and lack portability across different models.

Prefix tuning and P-tuning are 2 methods of prompt engineering that can enhance performance beyond traditional fine-tuning

[22-24]. Prefix tuning is a lightweight approach that keeps the PLM parameters unchanged while optimizing a sequence of task-specific vectors called the prefix [23]. This prefix is added to the input and interacts with the model's hidden states at each layer. Its success depends on how effectively the prefix is initialized, particularly when data are limited. P-tuning is another prompt tuning strategy that performs comparably to fine-tuning across various tasks [24]. It reduces the number of PLM parameters through self-adaptive pruning and tunes a small number of continuous prompts at the beginning of each transformer layer.

The verbalizer is the final layer that defines the answer space and maps it to the target output. Typically, verbalizers are manually created, which can limit their coverage due to personal vocabulary biases [21,29]. To address this, some studies have proposed automatic verbalizer search methods to identify more effective verbalizers, also known as soft verbalizers [20,30-32].

Autonomous ICD Coding in Cardiovascular Disease

Cardiovascular disease (CVD) is currently a leading cause of death worldwide, posing a significant risk of mortality among patients [7]. Automatically labeling patients with CVD is essential for clinical decision-making and resource allocation. However, existing prediction models have limitations, including low accuracy, limited generalizability, and an inability to capture multicenter data. To address these challenges, we propose a prompt learning real-time framework based on PLMs that can automatically label long free-text data with *ICD-10* codes for CVDs without the need for semiautomatic preprocessing.

Our framework consists of 4 components: a medically oriented pretrained BERT, a keyword filtration BERT, a fine-tuning phase, and task-specific prompt learning facilitated by mixed templates and soft verbalizers. To validate the efficacy of our framework, we conducted comprehensive evaluations on a Chinese multicenter cardiovascular dataset that includes data from 13,000 patients with CVD. This deliberate choice of dataset ensures the robustness and wide applicability of our framework. We compared our framework with RoBERTa, XLNet, and various BERT-based fine-tuning pipelines to highlight its performance. Additionally, we conducted few-shot experiments to demonstrate its resilience. This work promises to provide valuable insights into enhancing medical knowledge extraction and its effective application, underscoring the need for continued research and development in this promising area. In future work, we plan to implement this fully automated *ICD* coding pipeline across various clinical applications, including clinical decision support systems, cohort studies, and disease early warning and diagnosis systems.

Methods

Ethical Considerations

The study was approved by the Ethics Committee of the Chinese PLA General Hospital (S2023-325-02). Ethical approval included a waiver for obtaining informed consent signatures from participants. The study posed no potential harm to participants and did not involve any compensation for their participation. To protect patient privacy, we used regular

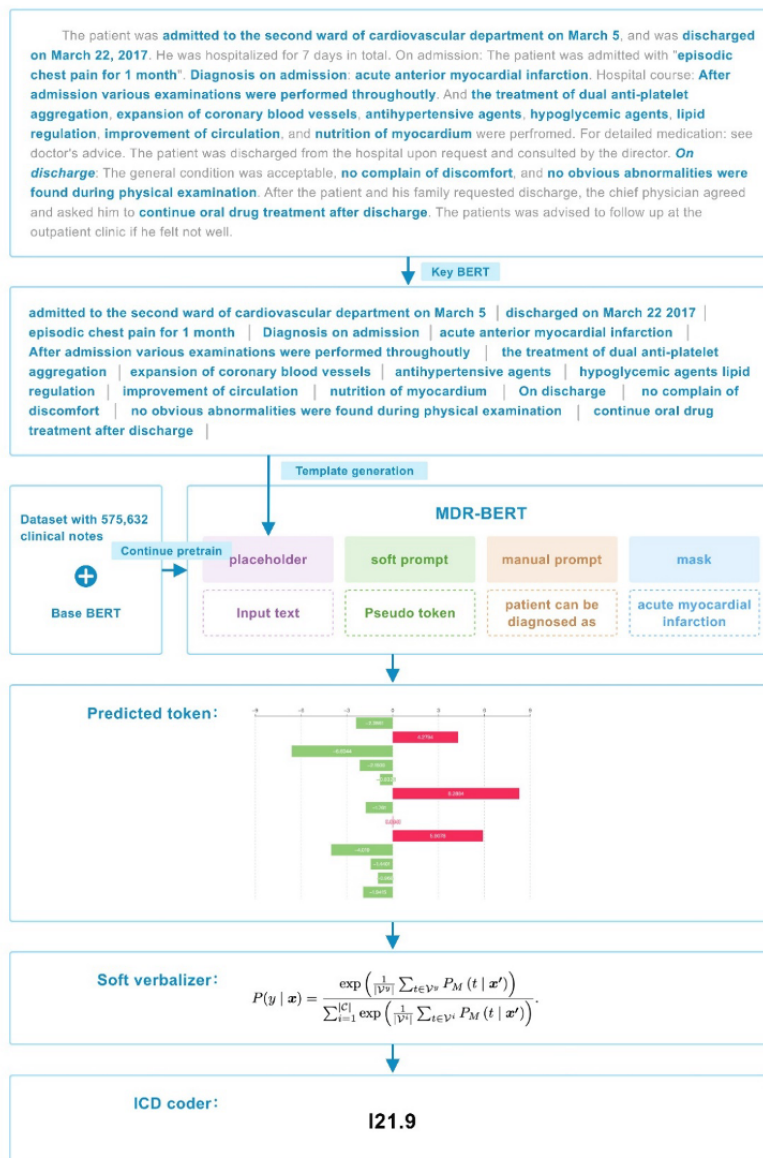
expressions to parse and redact basic identifying information from the medical records. As these records were created using a standardized template, we ensured that the excerpts extracted for this study did not contain patients' names.

Overview

The overall framework of the model is shown in Figure 1. We used a corpus dataset of 575,632 clinical notes to continue training the original BERT model, which we named medical domain refinement-BERT (MDR-BERT), as the PLM for our work. For the classification task, we first applied Key-BERT to filter the discharge summaries. This method extracts keywords and splits long free-text data into shorter sentences.

We then constructed the input template for fine-tuning and prompt learning using 3 components: the soft prompt, the manual prompt, and the mask part. The manual prompt was a handcrafted text prompt containing discrete tokens. The soft prompt was a learnable pseudo-token with a few continuous parameters. The mask part represented the ICD coding label. Finally, we used a trainable soft verbalizer to compute and apply the softmax function to the probabilities of the ICD classes, producing the output. By designing specific prompts, it is possible to incorporate the knowledge of medical experts into the model, helping it better understand and perform ICD coding. These prompts can direct the model to focus on critical sections of the input text, thereby enhancing performance.

Figure 1. Overall framework of MDR-BERT, Key-BERT, and prompt learning pipeline. BERT: bidirectional encoder representations from transformers; ICD: International Classification of Diseases; MDR: medical domain refinement.



Dataset Characteristics

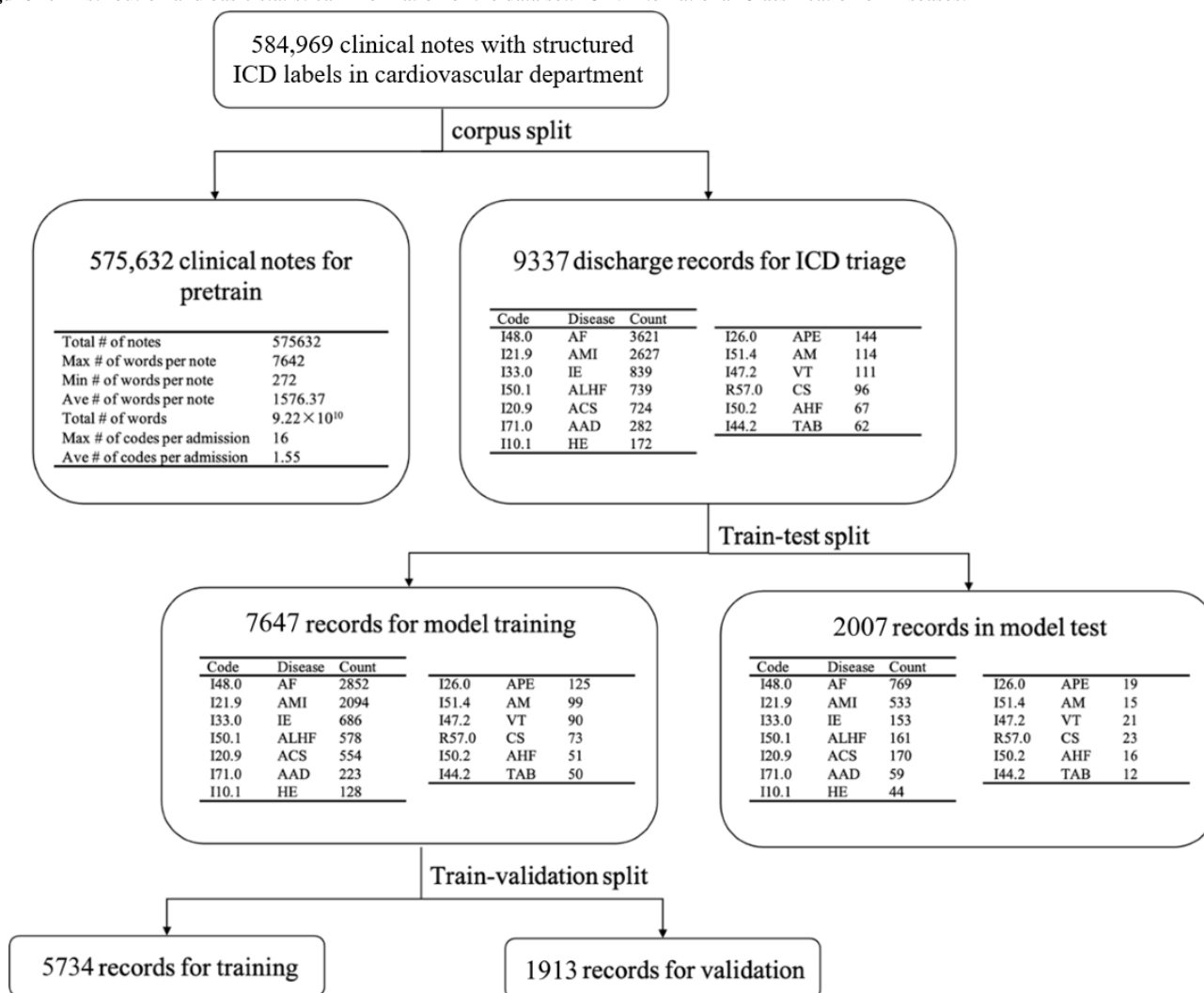
The cardiovascular dataset used in this study was obtained from the Cardiovascular Department of the Chinese PLA General Hospital's Medical Big Data Research Center in Beijing, China, which includes 9 medical centers with data aggregated into a

comprehensive medical big data platform. Additionally, the hospital is a key center for the treatment of CVDs, with numerous specialized physicians and detailed medical records, making its data highly practical and representative. To ensure privacy, patient names and addresses were desensitized. The data platform consists of electronic health records aggregated

from 8 affiliated medical centers. A total of 584,969 clinical notes with structured ICD labels were extracted from admission records and discharge summaries in the Cardiovascular Department. We ensured that each diagnosis included at least

50 cases and adopted a stratified sampling approach to divide each disease category into training, validation, and test sets in a 3:1:1 ratio. The detailed distribution and basic statistical information of the dataset are shown in Figure 2.

Figure 2. Distribution and basic statistical information of the data set. ICD: International Classification of Diseases.



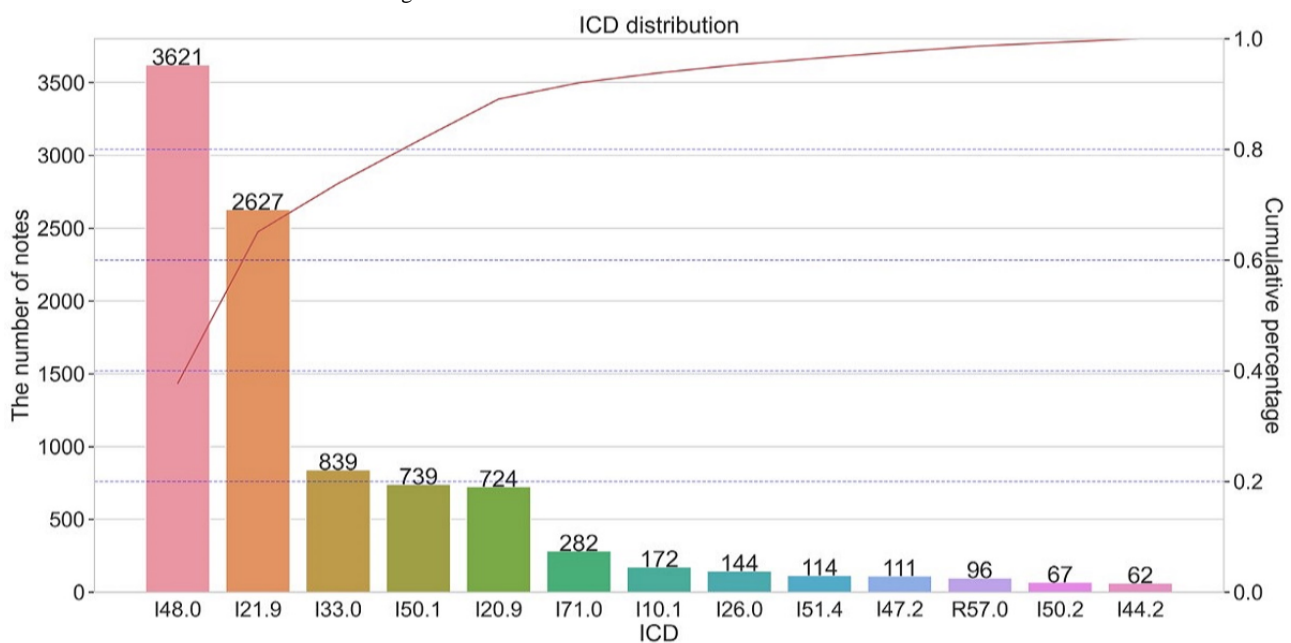
Based on the long-tailed distribution and clinician selection, 13 diseases were chosen for classification. These diseases include atrial fibrillation, acute myocardial infarction, infective endocarditis, acute left heart failure, acute coronary syndrome (ACS), acute aortic dissection, hypertensive emergency, acute pulmonary embolism, acute myocarditis, ventricular tachycardia, cardiogenic shock, acute heart failure, and third-degree atrioventricular block. The corresponding ICD-10 codes and abbreviations for these diseases are listed in Table 1. Despite the disparity in the number of cases for different diseases, the imbalance inherent in medical data accurately reflects real-world conditions, taking into account the clinical insights of medical professionals. This imbalance represents the varying frequency at which different diseases occur in clinical practice. By preserving the raw data distribution and avoiding artificial balancing, our training approach aligns more closely with real-world medical practice. As a result, this enhances the model’s generalization ability and its applicability in practical scenarios.

To ensure task independence and prevent data leakage, all clinical notes were divided into 2 parts: the pretraining corpus dataset and the ICD coding dataset. The pretraining corpus consisted of a total of 575,632 notes, while the ICD coding dataset included 9337 discharge records. The data were stratified by imbalanced ICD labels and randomly split into training, validation, and test sets in a 3:1:1 ratio. The sample sizes were as follows: 5734 in the training set, 1913 in the validation set, and 2007 in the test set. We applied regularization to truncate patients’ basic information, as this information could negatively impact the model’s fitting.

As shown in Figure 3, the distribution of the 13 ICD codes was imbalanced and exhibited a long-tail pattern. The dataset for ICD classification contains a total of 4.574×10^7 words, with an average of 490 words per note. The maximum and minimum lengths of the clinical notes are 5243 and 22 words, respectively.

Table 1. Overview of target International Classification of Diseases (ICD) codes and disease names.

| International Classification of Diseases code | Disease (abbreviation) |
|---|---|
| I48.0 | Atrial fibrillation (AF) |
| I21.9 | Acute myocardial infarction (AMI) |
| I33.0 | Infective endocarditis (IE) |
| I50.1 | Acute left heart failure (ALHF) |
| I20.9 | Acute coronary syndrome (ACS) |
| I71.0 | Acute aortic dissection (AAD) |
| I10.1 | Hypertensive emergency (HE) |
| I26.0 | Acute pulmonary embolism (APE) |
| I51.4 | Acute myocarditis (AM) |
| I47.2 | Ventricular tachycardia (VT) |
| R57.0 | Cardiogenic shock (CS) |
| I50.2 | Acute heart failure (AHF) |
| I44.2 | Third-degree atrioventricular block (TAB) |

Figure 3. Distribution of ICD codes for the triage task. ICD: International Classification of Diseases.

Pretraining

Our study's foundational framework is based on BERT, a multilayer bidirectional transformer encoder known for its conceptual simplicity and empirical effectiveness [33]. This architecture consists of 12 layers, a hidden size dimension of 768, and 12 self-attention heads [13]. BERT's inherent self-attention mechanism provides the versatility to handle various downstream tasks by allowing the interchange of relevant inputs and outputs, making it well-suited for our task involving ICD classification through clinical records.

To adapt BERT to the specific requirements of our task, we continued training the PLM using an extensive medical corpus, resulting in MDR-BERT. During the tuning process, we selected a batch size of 32, considering the constraint of a maximum

sequence length of 512 tokens. The Adam optimization algorithm was used with a conservative learning rate of 2×10^{-5} . The training was carried out over 15 epochs, an empirically determined figure based on the characteristics of the clinical dataset.

Key-BERT

The Key-BERT method offers a novel self-supervised framework for extracting keywords and keyphrases from textual content using deep learning techniques [34]. This approach leverages the contextual and semantic features provided by bidirectional transformers, with a particular focus on the influential BERT model. The method's architecture is designed for end-to-end training, utilizing a contextually self-annotated corpus that enables the model to develop a nuanced understanding of the complex relationships between words and

their semantic meanings. In the *ICD* coding task, Key-BERT leverages BERT's context-aware capabilities to extract keywords from the document, quickly identify the sections relevant to *ICD* coding, and reduce the risk of miscoding caused by misinterpreting or overlooking critical information in the text.

A distinctive feature of Key-BERT lies in its automated keyword labeling process. This process effectively utilizes contextual insights from bidirectional transformers to construct a carefully curated ground truth dataset. This approach bypasses the labor-intensive task of manual labeling and eliminates the need for domain-specific expertise.

The repository of self-labeled data generated by Key-BERT is partially shared with the NLP community, contributing to a deeper and more comprehensive understanding of keyword extraction techniques across various domains. This collaborative effort enhances the landscape of knowledge and expertise, driving advancements in the field of NLP and semantic information extraction.

To extract keywords using Key-BERT, the contextual feature vector for each word in a sentence is obtained by passing the sentence through the pretrained BERT model. Let $S = [w_1, w_2, \dots, w_n]$ be a sentence consisting of n words, where w_i is the i th word in the sentence and E_i is the contextual feature vector of the i th word in the sentence. The sentence embedding vector, denoted as E_s , is obtained by averaging the feature vectors of all the words in the sentence:

$$E_i = \text{BERT_Embedding}(w_i) \quad (1)$$

$$E_s = (E_1 + E_2 + \dots + E_n)/n \quad (2)$$

The cosine similarity metric is used to measure the similarity between the sentence embedding vector and the feature vectors of candidate keywords or keyphrases.

$$\text{Cos_SIM}(E_i, E_s) = (E_i \times E_s) / (\|E_i\| \times \|E_s\|) \quad (3)$$

The top-scoring keywords or keyphrases are returned as the most relevant to the document. Additionally, key medical terms are directly extracted using the medical diagnostic table, ensuring that essential terminology is accurately identified and applied.

Fine-Tuning and Prompt Learning

To fully leverage the clinical knowledge embedded within the dataset, our fine-tuning approach mirrors the unsupervised task used in the initial pretraining phase, known as masked language modeling (MLM). MLM involves randomly masking a predetermined proportion of input tokens, and the model then attempts to predict these masked tokens based on context. This process, commonly called a Cloze task, helps the model learn contextual relationships effectively.

For the fine-tuning phase in this study, we maintained the MLM framework to align with the pretraining procedure. A consistent masking rate of 15% was applied across the dataset. In addition to the fine-tuning process, we introduced prompt learning during parameter tuning. This approach involved the construction of a template comprising 4 distinct components: the input text, a soft prompt, a manual prompt, and a masking component. The

manual prompt included discrete tokens that reflected the downstream task expected by the PLM. By contrast, the soft prompt comprised trainable continuous vectors, which enhanced the model's adaptability.

Formally, automatic *ICD* coding, as a text multiclassification task, can be denoted as (x, y) , where x is the set of discharge summaries and y is the *ICD* code set of the 17 chosen discharge diagnoses as labels. Given a clinical record $x \in X$, it can be annotated with *ICD* codes of discharge diagnosis $y^x \in Y$ and a sequence of discrete input tokens $x = (x_0, x_1, \dots, x_k)$, where k is the number of tokens in the sequence. Prompt learning can be achieved via modifying the x to a prompt format $x = \text{fp}(x)$, where the template $f_p(\cdot)$ will insert a number of extra embeddings to x along with a masked token, denoted by $\langle[\text{MASK}] \rangle$. Compared with hard prompts, soft prompts replace some fixed manual components with trainable embeddings (continuous vectors) of the same dimension as the PLM. After that, x is fed into M , to predict the masked token, which is in accordance with the objective of M . The output of M will be a distribution over the fixed vocabulary V of M . The next crucial step is to map tokens in V to y for the downstream task with a mapping \square , known as verbalization. In a word, there are 2 essential components to be studied, the template of prompt $x' = f_p(x)$ and the mapping of verbalizer \square .

A mixed template of prompts in this paper is used. For simplicity, the prompt function $x' = f_p(x)$ is denoted as a sequence template:

$$x' = [P_0, P_1, \dots, P_j], x, [P_{j+1}, P_{j+2}, \dots, P_t], [\text{MASK}] \quad (4)$$

where P_i refers to the i th token in the template and t is the number of prompt tokens beyond x . P_i does not necessarily meet $P_i \in V$ other than manual hard prompt. As x' is fed to the PLM, the prompt tokens are also mapped to the embedding space, where we can assume that the tokens denoted as $\langle[\text{soft}] \rangle$ in the template can be tuned during training as pseudo-tokens. A simple example of a prompt template for automatic *ICD* coding could be as generated as follows:

$$x' = \langle x \rangle \langle [\text{soft}] \rangle \text{be encoded as } \langle [\text{MASK}] \rangle \quad (5)$$

Once these templates were formulated, the model inputs, along with the established templates, were processed through the trainable MDR-BERT model. Notably, in the final layer of the most advanced pipeline, a soft verbalizer mode was used. This mode manages the mapping process between the predicted tokens and the final *ICD* codes. The innovative feature of the soft verbalizer is its substitution of tokens in the verbalizer with trainable vectors, each tailored to a specific class. Generally, the verbalizer maps the probabilities of infrequent words in the vocabulary to the probabilities of the labels. The set of label words is denoted as V , the label space is Y , and V_y represents the subset of label words for label y . The final estimation of the probability for label y is calculated using equation 5, where g is utilized to convert the probability of label words to the probability of the label:

$$P(y|x_p) = g(P_M([\text{MASK}] = v|X_p)|v \in V_y) \quad (6)$$

This strategy enhances the precision and semantic accuracy of the generated outputs, enabling a more precise alignment between predicted tokens and the definitive *ICD* codes. Consequently, it is unnecessary to manually build an explicit mapping ϕ for the soft verbalizer, as the trainable vectors do not have explainable semantic meaning. A matrix operator can represent the soft verbalizer as \mathbf{W} [22-25], where n represents the size of y and m represents the dimension of output embeddings from M . For the verbalizer, θ_i denotes the i th row of \mathbf{W} as the trainable vector of the i th class. The soft verbalizer replaces the original decoder head of M by mapping the embeddings of x' from M , denoted as $e(x')$, to the distribution over the classes of y . We denote the resulting mapping from $e(x')$ to the prediction of the embedding of $\langle \text{[MASK]} \rangle$ as \hat{y} , where l is the sequence length of x' . And then, the probability of class y can be calculated as follows:

$$P(y) = \frac{\exp(\theta_i \cdot e(x'))}{\sum_j \exp(\theta_j \cdot e(x'))}$$

The loss from the automatic *ICD* coding task can be backpropagated to tune only the embeddings for the prompt template and the verbalizer. The loss function can be expressed as follows:

$$L = -\sum_y P(y) \log P(y)$$

Ultimately, the model learns to generate and map the most appropriate *ICD* codes to the corresponding discharge record.

The experiments were conducted using the OpenPrompt framework [22-25]. For prompt learning, we utilized the Adafactor optimizer for soft and mixed prompt templates, while the AdamW optimizer was used for the PLMs and soft verbalizers. In conventional fine-tuning, we applied the AdamW optimizer to the MLP heads and PLMs. To expedite the experiments, we used 2 Nvidia TESLA V100 GPUs, each with 16-GB memory, and set the batch size to 32 due to memory constraints.

The model's performance is influenced by variations in hyperparameters. In the comparisons presented, hyperparameters were carefully optimized for each model. To determine the optimal configuration, we used a random search strategy. This approach involves generating multiple random combinations of parameters, evaluating the performance of each combination, and selecting the one that yields the best results. Accuracy and AUC were chosen as the primary optimization objectives during the random search, as they intuitively reflect the model's classification performance. The strategy involved 100 training runs, each using randomly generated hyperparameters from the defined search space. To effectively address model overfitting, we carefully adjusted the dropout rate within a range of 0.1-0.5. After numerous training iterations, we found that the optimal dropout rate for the prompt learning phase is 0.382, while for the prompt tuning phase, it is 0.1563. In the prompt learning phase, a higher dropout rate contributes to improved generalization, serving as an effective safeguard against overfitting. In the subsequent fine-tuning phase, a lower dropout rate is used to ensure the model retains its learned attributes while enabling further performance enhancement. The optimal hyperparameters for the models are detailed in Table 2.

Table 2. The optimal hyperparameters and their search space.

| Hyperparameters | Search space | Optimal hyperparameter | |
|-----------------------------|------------------------------|------------------------|----------------|
| | | Prompt learning | Fine-tuning |
| Learning rate | log.uniform [1*10-5, 3*10-1] | 0.0048 | 0.0121 |
| Batch size | 4 | 4 | 4 |
| Gradient accumulation steps | range[2,10] | 4 | 3 |
| Dropout | range[0.1,0.5] | 0.382 | 0.1563 |
| Optimizer | [adamw, adafactor] | adamw | adafactor |
| Prompt learning rate | log.uniform[1*10-5, 3*10-1] | 0.3 | — ^a |
| Verbalizer learning rate | log.uniform[1*10-5, 1*10-1] | 0.007 | — |

^aNot available.

Evaluation Metrics

To thoroughly evaluate and compare the performance of the models, we used a range of metrics, including micro- F_1 -score, macro-AUC, and accuracy. The definitions for micro-averaged precision and micro- F_1 -score are provided in equations 9-11, while the macro-AUC is defined in equations 12 and 13.

$$P = \frac{TP_i}{TP_i + FP_i}$$

$$R = \frac{TP_i}{TP_i + FN_i}$$

$$\text{Micro-}F_1\text{-score} = [2 \times (\text{micro-}P) \times (\text{micro-}R)] / [(\text{micro-}P) + (\text{micro-}R)] \quad (11)$$

where TP_i , FP_i , and FN_i represent true positives (correctly assigned instances), false positives (incorrect assignments by automated methods), and false negatives (correct instances omitted by automated methods), respectively, of code i , and l is the size of the sample space. The micro- F_1 -score is the harmonic mean of micro- P and micro- R , and a bigger value of micro- F_1 -score indicates a better performance.



where n is the number of thresholds and K is the number of classes.

Data and Code Availability

Data acquisition requests can be made by contacting the corresponding author (KH). Given the sensitive nature of the hospital data, it cannot be released publicly. However, part of the downstream subtask data is currently undergoing desensitization and approval processes. The source code for this study is publicly available on GitHub [35].

Results

Performance of Different Pipelines

To evaluate the performance of different methods, we implemented 4 state-of-the-art algorithms: BERT [15], XLNet [18], RoBERTa [19,36], and prompt learning [22]. These PLMs were integrated with various algorithms to create 6 main pipelines: BERT with fine-tuning, XLNet with fine-tuning, RoBERTa with fine-tuning, BERT with prompt learning, MDR-BERT with prompt learning, and MDR-BERT with both fine-tuning and prompt learning. MDR-BERT is a PLM developed by further pretraining BERT on our medical corpus.

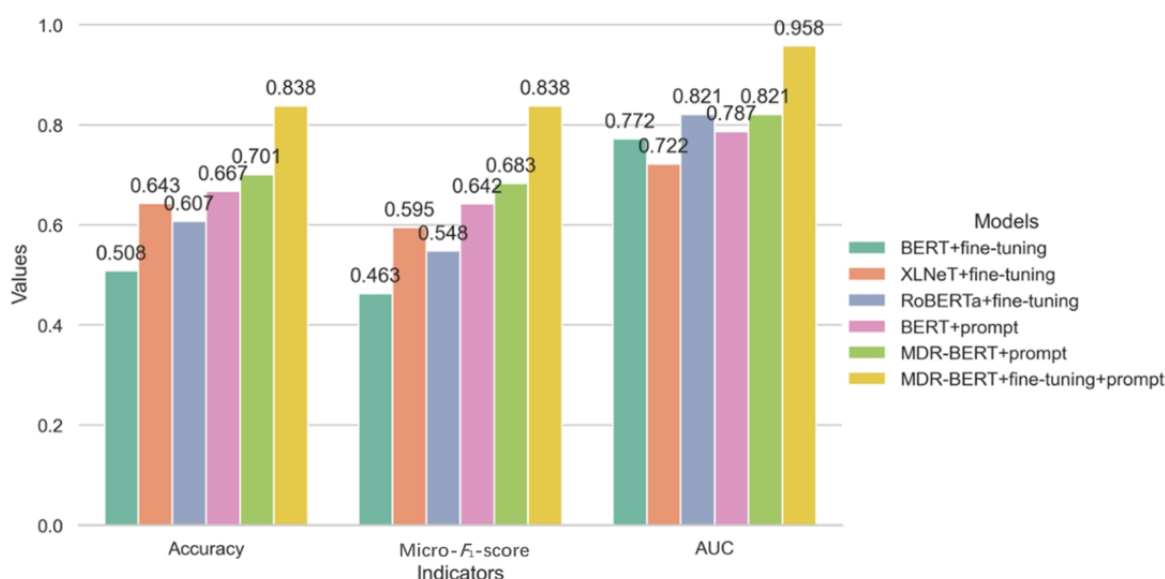
As shown in Figure 4, MDR-BERT with fine-tuning and prompt learning achieved the highest performance across all evaluation metrics, with a micro- F_1 -score of 0.838, a macro-AUC of 0.958, and an accuracy of 0.838. MDR-BERT with prompt learning alone performed slightly worse than the combined fine-tuning

and prompt learning approach, but both outperformed the other pipelines by a significant margin. This suggests that continued pretraining on clinical records can significantly enhance the performance of the PLM for the task, while freezing parameters may hinder the adaptation of smaller PLMs to the task.

Among the other pipelines, BERT with prompt learning achieved the highest accuracy (0.67) and the highest micro- F_1 -score (0.64), though its macro-AUC (0.79) was slightly lower than that of RoBERTa with fine-tuning. This suggests that prompt learning, as a lightweight tuning approach, can match or even surpass traditional fine-tuning methods, aligning with the findings of Taylor et al [22].

We also conducted a comparison with state-of-the-art methods and selected 2 prominent models: mt5-xxl (11B) and Qwen2.5-72B-Instruct. Among these, mt5-xxl demonstrated the best performance in text classification, while Qwen2.5-72B-Instruct excelled as a large language model. For mt5-xxl, we fine-tuned the model using the training and validation sets from our fine-tuning dataset, setting the “prefix_text” to “Classify the following text:”. For Qwen2.5-72B-Instruct, we conducted experiments using both zero-shot and retrieval augmented generation methods. In the zero-shot setting, we used prompts to constrain the diagnostic scope, allowing the model to make inferences based on the input information. For the retrieval augmented generation approach, we first encoded the training set using BGE-M3 (BAAI general embedding multilinguality, multigranularity, and multifunctionality) and stored it in a Faiss vector database. During the testing phase, we retrieved cases and classification results relevant to the input content and concatenated them with the prompt to enhance model performance.

Figure 4. The optimal hyperparameters and their search space. AUC: area under the receiver operating characteristic curve; BERT: bidirectional encoder representations from transformers; MDR: medical domain refinement; RoBERTa: robustly optimized BERT pretraining approach; XLNet: extreme language network.



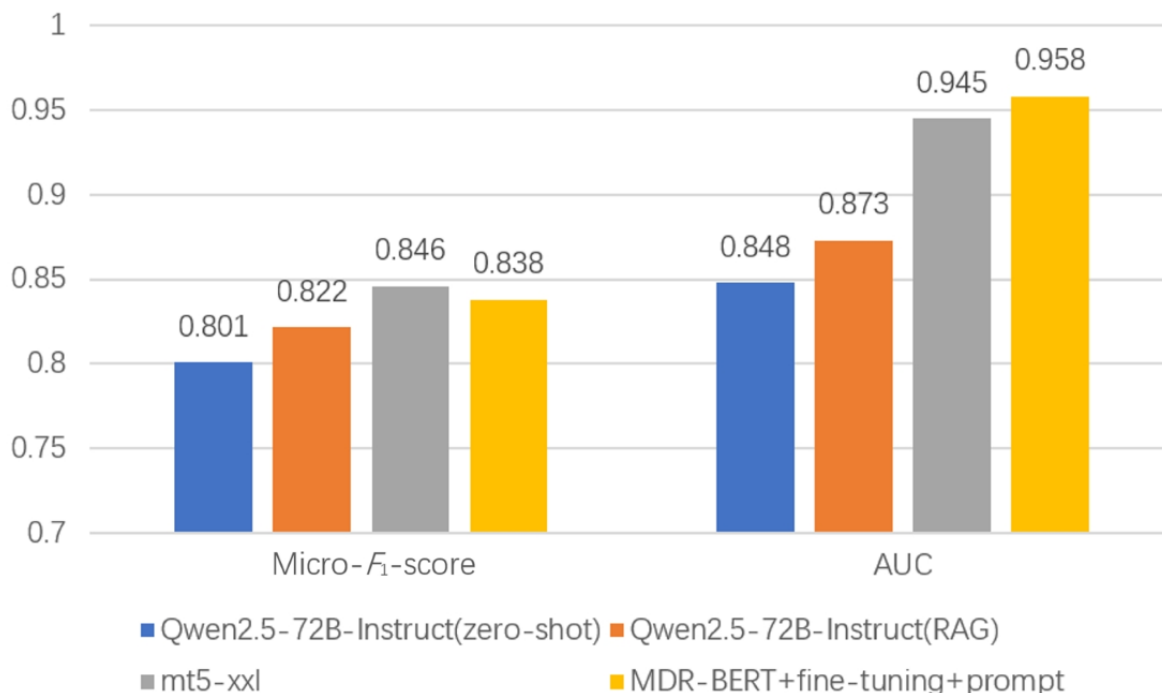
The experimental results indicate that the micro- F_1 -score for the mt5-xxl method is 0.846, and the AUC value is 0.945. In comparison, the micro- F_1 -score for the Qwen2.5-72B-Instruct

method was 0.822, and the AUC value was 0.848. However, the accuracy of both methods does not surpass that of our MDR-BERT model (Figure 5). After a series of strategic

optimizations, our MDR-BERT model achieved results comparable to the fine-tuned mt5-xxl on specific tasks. This is primarily due to the specific structure of the medical records,

which can be effectively captured by models with fewer parameters, meaning that overly complex models are not necessary to achieve good performance.

Figure 5. Micro- F_1 -score and AUC values for the MDR-BERT model versus the QWEN2.5 and mt5-xxl models. AUC: area under the receiver operating characteristic curve; MDR: medical domain refinement; BERT: bidirectional encoder representations from transformers.

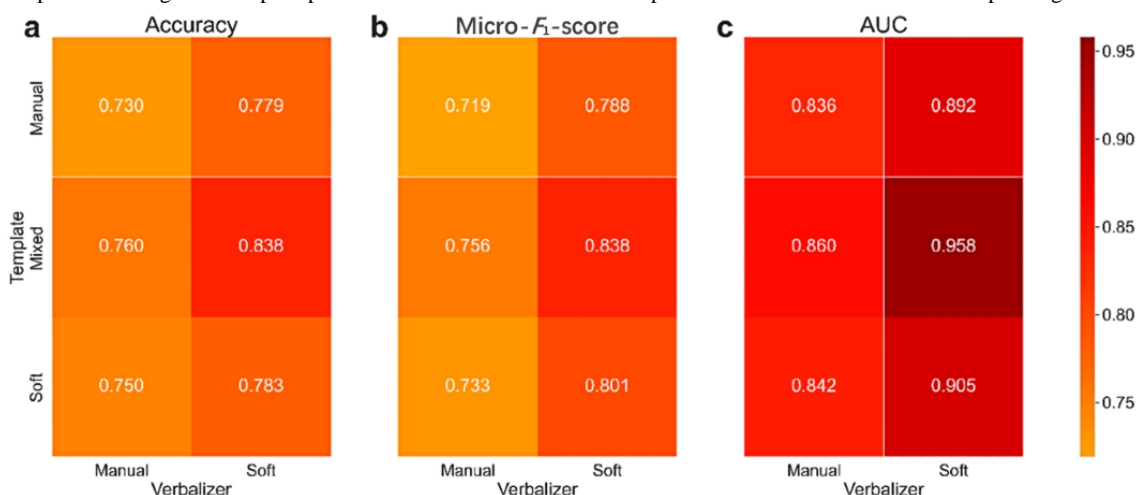


Performance of Different Prompt Learning Modes

We evaluated the performance of MDR-BERT under various settings of prompt learning and fine-tuning, using 3 types of templates (manual, soft, and mixed) and 2 types of verbalizers (manual and soft) as hyperparameters.

For templates, both scripted and self-adaptive patterns performed well independently, and their combination had a cumulative positive effect on performance. For verbalizers, the self-adaptive type outperformed the traditional manual vectors and had a greater impact on overall performance. As shown in Figure 6, the combination of mixed templates and the soft verbalizer achieved the best results.

Figure 6. Comparison among different prompt combinations in verbalizer and template. AUC: area under the receiver operating characteristic curve.



Take the following prompt template as an example:

Mixed template: {"placeholder": "text_a"} patient {"soft": "can be diagnosed as "} {"mask"}.

For the following case:

The patient was discovered to have bradycardia and unconscious disturbance 7 days ago as a result of physical examination. After consultation with the director, lipid-lowering drugs were added. No diarrhea was detected, and no medication was administered at home. Permanent cardiac pacemaker

implantation under local anesthesia was carried out, and after the surgery, cephalosporin for injection was utilized to prevent infection.

The classification result by our model is as follows: “The patient can be diagnosed as {third-degree atrioventricular block}.”

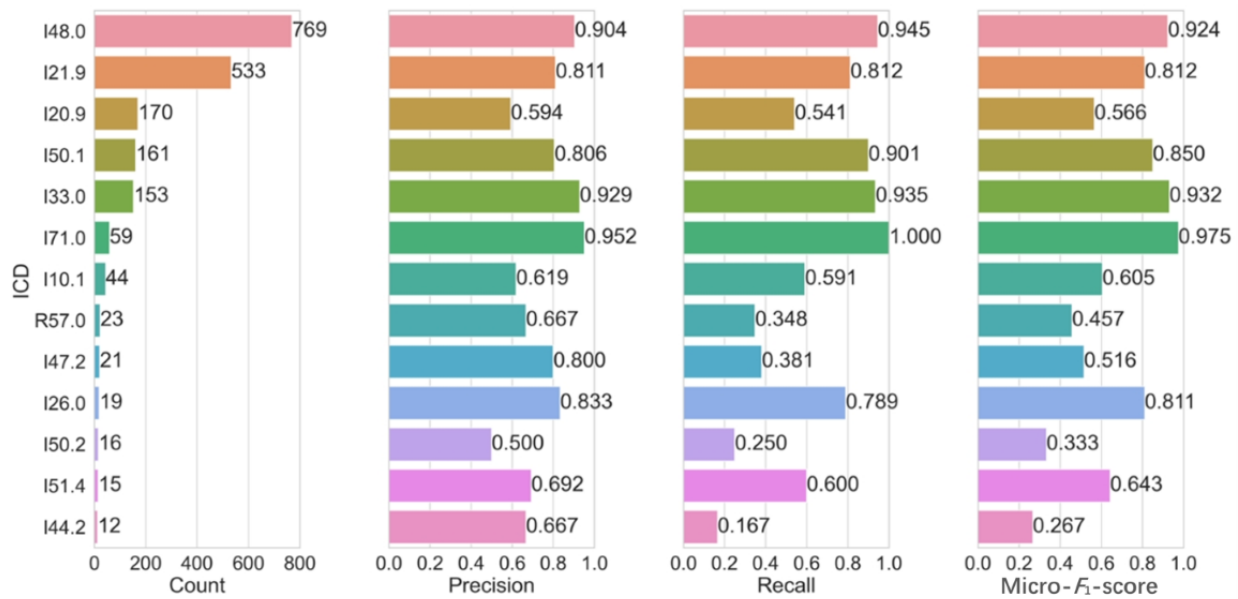
For the mixed template, the patient’s bradycardia requires management through the implantation of a permanent pacemaker, indicating that bradycardia is a major medical concern. By applying soft verbalizers, we can guide the correct diagnosis by emphasizing both the reason for the pacemaker

implantation and the underlying cause of bradycardia: “The patient can be diagnosed with third-degree atrioventricular block.”

Performance of MDR-BERT With Fine-Tuning and Prompt Learning

We evaluated the performance of the MDR-BERT pipeline, incorporating both fine-tuning and prompt learning, for each ICD code using precision, recall, and micro- F_1 -score. Figure 7 presents the results for these metrics across the 13 ICD classes.

Figure 7. Precision, recall, and micro- F_1 scores of every ICD code in the MDR-BERT pipeline with fine-tuning and prompt learning. BERT: bidirectional encoder representations from transformers; ICD: International Classification of Diseases; MDR: medical domain refinement.



The pipeline achieved high scores for most ICD codes, although the scores varied depending on the data distribution and sample size for each code. We observed a weak positive correlation between sample size and model performance, suggesting that larger samples enhanced the model’s learning capability. Conversely, smaller samples tended to have lower micro- F_1 -scores, with a trade-off between precision and recall for certain classes. Although our prediction accuracy for ACSs is relatively low, further analysis revealed that in actual clinical settings, ACS was frequently misdiagnosed as cardiac edema (hypertensive emergency) and pulmonary embolism (acute pulmonary embolism). These diseases exhibit similar clinical manifestations and, therefore, require meticulous differential diagnosis to rule out other possibilities. We believe that the overlap of symptoms is a major cause of the difficulty in classifying the model and that inconsistencies in medical histories recorded by physicians further complicate the model’s ability to differentiate similar pathologies. Despite these variations, our pipeline demonstrated satisfactory performance across the different ICD codes.

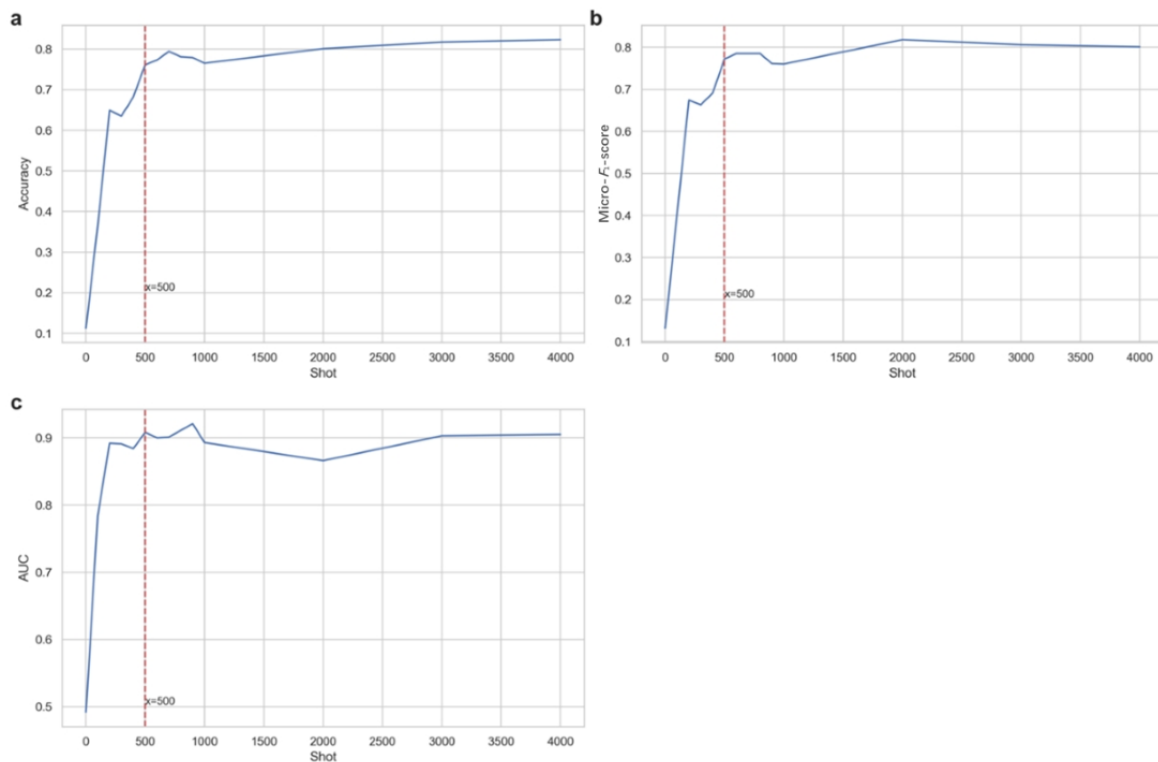
Few-Shot Learning

We conducted few-shot experiments to evaluate the performance of the fine-tuned MDR-BERT with the prompt learning pipeline using different sample sizes from the training set. We randomly

selected samples ranging from 1 to 4000 and evaluated the models on the test set. Figure 8 shows the accuracy, micro- F_1 -score, and macro-AUC scores for each sample size.

The objective of small-sample learning is to develop models that can learn effectively and make accurate predictions with only a small number of samples, such as 500 or fewer. As shown in Figure 8, when the sample size reaches 500, the model’s accuracy, AUC score, and other indicators not only achieve relatively high scores but also reach an inflection point and plateau. At this point, the model produces a relatively satisfactory outcome. This indicates that for the task of ICD coding using medical records, 500 samples may be sufficient for the model to learn the key features needed to distinguish between different diagnoses. It suggests that the model has captured enough information to make effective predictions. Additionally, the workload involved in annotating 500 medical texts is manageable and feasible. This number strikes a balance between the effort required for data preparation and the performance gains achieved by the model. Given the complexity and specialized nature of medical records, annotating 500 examples provides a comprehensive representation of the dataset while staying within practical limits. This makes it a reasonable and efficient choice for training the model to achieve satisfactory performance in ICD coding tasks.

Figure 8. Few shots experiments on MDR-BERT with fine-tuning and prompt learning. AUC: area under the receiver operating characteristic curve; BERT: bidirectional encoder representations from transformers; MDR: medical domain refinement.



Discussion

Principal Findings

An automated *ICD* coding system for long free-text data is a fundamental platform for clinical research and practice, including clinical trials and pharmaco-economic management. In this study, we developed a framework based on Key-BERT, a continuously trained and tunable PLM, combined with task-specific prompt learning. We collected a total of 584,969 clinical notes from admission records and discharge summaries in the cardiovascular departments of 8 medical centers.

We used most of the data to continue pretraining a medical corpus and used an independent set of 9337 discharge records with 13 *ICD* codes for CVDs in the *ICD* classification subtask. Although the MDR-BERT model has some limitations, such as restricted generalization capacity and constraints on the length of context it can effectively process, it is important to note that medical texts often have a consistent structure and are generally less dependent on extensive contextual information. Given these characteristics of medical literature, our model is designed to avoid the errors commonly associated with the inherent limitations of BERT's methodology. The structured nature of medical documents enables the MDR-BERT model to function effectively within its designed parameters, mitigating potential issues that could arise from the broader weaknesses of the BERT framework when applied to more contextually complex or varied text types. To remove irrelevant information and limit the input token size, we filtered and truncated all the data for the *ICD* task into keyword-based segments using Key-BERT. The data were then stratified and split into training, validation, and test sets, with the test set used independently for final evaluation.

This study primarily focused on transformer-based algorithms, which have been widely applied and shown superior performance in large-scale medical long free-text tasks [4,11,16,17]. These algorithms can leverage PLMs that capture the semantic and syntactic information of natural language from extensive corpora, leading to significant performance improvements through multicenter datasets.

We compared 6 pipelines for the classification downstream task: BERT with fine-tuning, XLNet with fine-tuning, RoBERTa with fine-tuning, frozen BERT with prompt learning, frozen MDR-BERT with prompt learning, and tunable MDR-BERT with prompt learning. The prompt learning setup included 3 types of templates and 2 types of verbalizers. Among these pipelines, MDR-BERT with fine-tuning and prompt learning achieved the best performance on the test set, attaining a micro- F_1 -score of 0.838, a macro-AUC of 0.958, and an accuracy of 0.838.

Compared with the pretraining models of RoBERTa and XLNet, our model achieved superior performance in terms of final accuracy and micro- F_1 -score. This improvement was primarily due to the targeted optimization of the methods and the medical data we selected, which substantially enhanced the model's performance. Although RoBERTa and XLNet have larger pretraining corpora compared with BERT, our approach benefited more from using a continuation training corpus built from real electronic health records. This specialized data, tailored to our specific requirements, provided a greater enhancement to the model than more general pretraining data. This is why MDR-BERT performs comparably to, or better than, these alternatives in our settings. The favorable outcome of this pipeline can be attributed to the use of a large-scale

corpus-based PLM and the task-specific enhancements from the combination of fine-tuning and prompt learning [16,20,22-25]. Fine-tuning acts as a model adapter, aligning the model distribution with the task distribution and addressing domain shift and task mismatch issues inherent in PLMs. Prompt learning, with its compact prefix representation and sparse attention mechanism, augments the training data with diverse and natural examples. This augmentation helps mitigate data scarcity and label noise issues in small-sized datasets for downstream tasks.

The combination of fine-tuning and prompt learning acts as a regularization term that balances model complexity with data quality, ultimately enhancing overall performance. This integrated approach highlights the potential of leveraging advanced transformer-based models and customized learning strategies to improve automated medical coding and other clinical tasks.

Among the different prompt learning setups, the mixed template and soft verbalizer achieved the best performance. The soft template method outperformed the manual templates method, which can be attributed to the greater semantic and syntactic information, broader search space, and reduced trial-and-error process associated with the soft template method, making it more effective and less time-consuming [23,24].

The mixed template method is a hybrid approach that combines the advantages of both soft and manual templates. It uses a manual template as a base prompt to provide human-readable instructions and natural language labels, while a soft template serves as an auxiliary prompt to provide tunable embeddings that can adapt to specific downstream tasks. This way, the manual template leverages existing knowledge, while the soft template enhances expressiveness and flexibility.

For the verbalizer, the self-adaptive type had a significantly greater impact on overall performance compared with traditional manual vectors. The soft verbalizer adjusts to the optimal label space for each task and the scale of the pretrained model, rather than being limited by a fixed set of tokens [22,24]. This enhances the accuracy and robustness of the predictions, as well as the diversity and naturalness of the labels. Additionally, by tuning the verbalizer alongside other continuous prompts, it retains the benefits of prompt tuning over fine-tuning, eliminating the need to maintain a separate copy of model parameters for each task during inference.

To explore the influence of sample size on the performance of our pipeline, we conducted few-shot experiments with a range from 1 to 4000 shots. The results showed unsatisfactory evaluation metrics for small-scale shots, but performance improved rapidly and stabilized at around 500 shots. This suggests that for mid-sized language models, such as BERT, the semantic understanding and representation capabilities may not be strong enough. Therefore, tuning the parameters of the PLM with an appropriate sample size is necessary to achieve better performance on specific tasks.

Our research confirms that *ICD* classification tasks can be effectively accomplished by continuously optimizing the BERT model. Although this study used cardiology data for training,

our model development strategy is not limited to this specific dataset; substituting the training data with data from other departments would also yield the expected outcomes. Therefore, our model demonstrates remarkable generalization capability. We firmly believe that the model we have developed, combined with the expertise of professional physicians, can effectively address the challenges of *ICD* classification for various diseases.

Limitations

Despite the reasonable performance of our pipeline, this study has certain limitations. First, we trained both the corpus part and the classification task of the framework solely in the cardiovascular department. As a result, the conclusions of this paper may not be generalizable to other medical fields. Second, the *ICD* classification subtask only involved 13 CVD codes, which is not comprehensive enough for clinical practice. Future research could expand to explore the automatic encoding of additional critical heart diseases or even extend to the entire clinical field. This could potentially enhance the applicability and effectiveness of the proposed approach for a broader range of clinical tasks. Third, our model aims to establish an automated analysis system using medical text. However, medical data are inherently multimodal, and modality augmentation can lead to improvements in accuracy. In this context, models such as label alignment for multimodal prompt learning [37] and multimodal equivalent transformer [38] are designed to handle multimodal data, demonstrating the greater potential for future advancements.

Conclusions

We proposed a real-time framework for *ICD* coding from long medical field-related text to *ICD* labels, eliminating the need for semistructured preprocessing. This framework incorporates Key-BERT, a continuously trained and tunable PLM, and task-specific prompt learning with mixed templates and soft verbalizers. We evaluated our model on a multicenter cardiovascular dataset and applied it to predict 13 *ICD* codes for CVDs, achieving high performance. Our model also demonstrated transferability and generalization across different centers.

Furthermore, we conducted few-shot experiments to investigate the impact of data size on model performance. The results showed that while the framework was effective on smaller datasets, a certain sample size was necessary to achieve a relatively stable performance level. This study serves as a benchmark for exploring the feasibility and performance of prompt learning in the subtask of large language models or PLMs. Using a multicenter dataset, the approach demonstrated robust performance across hospitals, highlighting its potential for broad deployment.

Few-shot learning experiments demonstrated feasibility with small-scale datasets, enabling applications for local training on single centers or various single-disease databases. The real-time model identifies *ICD* codes directly, accelerating automated coding compared with semiautomatic approaches that require segment preprocessing. This is particularly impactful for clinical decision support systems that rely on real-time *ICD* coding data.

Overall, the prompt learning paradigm achieved cutting-edge ICD assignment accuracy while offering deployability, few-shot learning capacity, and low latency—advantages that are highly beneficial for health care applications. This automated ICD coding pipeline could be further implemented in various clinical applications, such as clinical decision support systems, cohort studies, and disease early warning and diagnosis systems.

Acknowledgments

This work was supported by the National Key R&D Program of China (grant 2021ZD0140408) and the Independent Research Project of Medical Engineering Laboratory of Chinese PLA General Hospital (grant 2022SYSZZKY23).

Conflicts of Interest

None declared.

References

1. Steindel SJ. International Classification of Diseases, 10th Edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc* 2010;17(3):274-282 [FREE Full text] [doi: [10.1136/jamia.2009.001230](https://doi.org/10.1136/jamia.2009.001230)] [Medline: [20442144](https://pubmed.ncbi.nlm.nih.gov/20442144/)]
2. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005 Oct;40(5 Pt 2):1620-1639 [FREE Full text] [doi: [10.1111/j.1475-6773.2005.00444.x](https://doi.org/10.1111/j.1475-6773.2005.00444.x)] [Medline: [16178999](https://pubmed.ncbi.nlm.nih.gov/16178999/)]
3. Kusnoor S, Blasingame MN, Williams AM, DesAutels SJ, Su J, Giuse NB. A narrative review of the impact of the transition to ICD-10 and ICD-10-CM/PCS. *JAMIA Open* 2020 Apr;3(1):126-131 [FREE Full text] [doi: [10.1093/jamiaopen/ooz066](https://doi.org/10.1093/jamiaopen/ooz066)] [Medline: [32607494](https://pubmed.ncbi.nlm.nih.gov/32607494/)]
4. Chen P, Chen K, Liao W, Lai F, He T, Lin S, et al. Automatic International Classification of Diseases coding system: deep contextualized language model with rule-based approaches. *JMIR Med Inform* 2022 Jun 29;10(6):e37557 [FREE Full text] [doi: [10.2196/37557](https://doi.org/10.2196/37557)] [Medline: [35767353](https://pubmed.ncbi.nlm.nih.gov/35767353/)]
5. Upadhyaya SG, Murphree DH, Ngufor CG, Knight AM, Cronk DJ, Cima RR, et al. Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clin Proc Innov Qual Outcomes* 2017 Jul;1(1):100-110 [FREE Full text] [doi: [10.1016/j.mayocpiqo.2017.04.005](https://doi.org/10.1016/j.mayocpiqo.2017.04.005)] [Medline: [30225406](https://pubmed.ncbi.nlm.nih.gov/30225406/)]
6. Diao X, Huo Y, Zhao S, Yuan J, Cui M, Wang Y, et al. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *Int J Med Inform* 2021 Sep;153:104543 [FREE Full text] [doi: [10.1016/j.ijmedinf.2021.104543](https://doi.org/10.1016/j.ijmedinf.2021.104543)] [Medline: [34391016](https://pubmed.ncbi.nlm.nih.gov/34391016/)]
7. Maheshwari S, Agarwal A, Shukla A, Tiwari R. A comprehensive evaluation for the prediction of mortality in intensive care units with LSTM networks: patients with cardiovascular disease. *Biomed Tech (Berl)* 2020 Aug 27;65(4):435-446. [doi: [10.1515/bmt-2018-0206](https://doi.org/10.1515/bmt-2018-0206)] [Medline: [31846424](https://pubmed.ncbi.nlm.nih.gov/31846424/)]
8. Bao W, Lin H, Zhang Y, Wang J, Zhang S. Medical code prediction via capsule networks and ICD knowledge. *BMC Med Inform Decis Mak* 2021 Jul 30;21(Suppl 2):55 [FREE Full text] [doi: [10.1186/s12911-021-01426-9](https://doi.org/10.1186/s12911-021-01426-9)] [Medline: [34330264](https://pubmed.ncbi.nlm.nih.gov/34330264/)]
9. Kreuzthaler M, Pfeifer B, Kramer D, Schulz S. Secondary use of clinical problem list entries for neural network-based disease code assignment. *Stud Health Technol Inform* 2023 May;18(302):788-792. [doi: [10.3233/shiti230267](https://doi.org/10.3233/shiti230267)]
10. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. *arXiv* [FREE Full text] Preprint posted online on March 22, 2018 [doi: [10.48550/arXiv.1802.05365](https://doi.org/10.48550/arXiv.1802.05365)]
11. Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, et al. Unified language model pre-training for natural language understanding and generation. *arXiv* [FREE Full text] Preprint posted online on October 15, 2019 [doi: [10.48550/arXiv.1905.03197](https://doi.org/10.48550/arXiv.1905.03197)]
12. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* [FREE Full text] Preprint posted online on October 29, 2019 [doi: [10.48550/arXiv.1910.13461](https://doi.org/10.48550/arXiv.1910.13461)]
13. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* [FREE Full text] Preprint posted online on May 24, 2019 [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
14. Coutinho I, Martins B. Transformer-based models for ICD-10 coding of death certificates with Portuguese text. *J Biomed Inform* 2022 Dec;136:104232 [FREE Full text] [doi: [10.1016/j.jbi.2022.104232](https://doi.org/10.1016/j.jbi.2022.104232)] [Medline: [36307020](https://pubmed.ncbi.nlm.nih.gov/36307020/)]
15. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, et al. RadBERT: adapting transformer-based language models to radiology. *Radiol Artif Intell* 2022 Jul;4(4):e210258 [FREE Full text] [doi: [10.1148/ryai.210258](https://doi.org/10.1148/ryai.210258)] [Medline: [35923376](https://pubmed.ncbi.nlm.nih.gov/35923376/)]
16. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv* [FREE Full text] Preprint posted online on July 26, 2019 [doi: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692)]
17. Elbattah M, Arnaud E, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. 2021 Presented at: 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Scale-IT-up; February 11–13, 2021; Virtual Event. [doi: [10.5220/0010414508250832](https://doi.org/10.5220/0010414508250832)]

18. Sezgin E, Hussain S, Rust S, Huang Y. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. *JMIR Form Res* 2023 Mar 07;7:e43014 [FREE Full text] [doi: [10.2196/43014](https://doi.org/10.2196/43014)] [Medline: [36881467](https://pubmed.ncbi.nlm.nih.gov/36881467/)]
19. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: generalized autoregressive pretraining for language understanding. 2019 Sep Presented at: NIPS'19: 33rd International Conference on Neural Information Processing Systems; December 8-14, 2019; Vancouver, BC, Canada p. 5753-5763.
20. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. *AI Open* 2024;5:208-215. [doi: [10.1016/j.aiopen.2023.08.012](https://doi.org/10.1016/j.aiopen.2023.08.012)]
21. Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference. arXiv [FREE Full text] Preprint posted online on January 25, 2021 [doi: [10.48550/arXiv.2001.07676](https://doi.org/10.48550/arXiv.2001.07676)]
22. Taylor N, Zhang Y, Joyce DW, Gao Z, Kormilitzin A, Nevado-Holgado A. Clinical prompt learning with frozen language models. *IEEE Trans Neural Netw Learning Syst* 2024 Nov;35(11):16453-16463. [doi: [10.1109/tnnls.2023.3294633](https://doi.org/10.1109/tnnls.2023.3294633)]
23. Li X, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv [FREE Full text] Preprint posted online on January 21, 2021 [doi: [10.48550/arXiv.2101.00190](https://doi.org/10.48550/arXiv.2101.00190)]
24. Liu X, Ji K, Fu Y, Du Z, Yang Z, Tang J. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv [FREE Full text] Preprint posted online on March 20, 2022 [doi: [10.48550/arXiv.2110.07602](https://doi.org/10.48550/arXiv.2110.07602)]
25. Ding N, Hu S, Zhao W, Chen Y, Liu Z, Zheng H, et al. OpenPrompt: an open-source framework for prompt-learning. arXiv [FREE Full text] Preprint posted online on November 3, 2021 [doi: [10.48550/arXiv.2111.01998](https://doi.org/10.48550/arXiv.2111.01998)]
26. Jiang Z, Xu FF, Araki J, Neubig G. How can we know what language models know? arXiv May 3:1 [FREE Full text] Preprint posted online on May 3, 2020 [doi: [10.48550/arXiv.1911.12543](https://doi.org/10.48550/arXiv.1911.12543)]
27. Haviv A, Berant J, Globerson A. BERTese: learning to speak to BERT. arXiv [FREE Full text] Preprint posted online on March 11, 2021 [doi: [10.48550/arXiv.2103.05327](https://doi.org/10.48550/arXiv.2103.05327)]
28. Wallace E, Feng S, Kandpai N, Gardner M, Singh S. Universal adversarial triggers for attacking and analyzing NLP. arXiv [FREE Full text] Preprint posted online on January 3, 2021 [doi: [10.48550/arXiv.1908.07125](https://doi.org/10.48550/arXiv.1908.07125)]
29. Cui L, Wu Y, Liu J, Yang S, Zhang Y. Template-based named entity recognition using BART. arXiv [FREE Full text] Preprint posted online on June 3, 2021 [doi: [10.48550/arXiv.2106.01760](https://doi.org/10.48550/arXiv.2106.01760)]
30. Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. arXiv [FREE Full text] Preprint posted online on June 2, 2021 [doi: [10.48550/arXiv.2012.15723](https://doi.org/10.48550/arXiv.2012.15723)]
31. Shin R, Lin CH, Thomson S, Chen C, Roy S, Platanios EA, et al. Constrained language models yield few-shot semantic parsers. arXiv [FREE Full text] Preprint posted online on November 16, 2021 [doi: [10.48550/arXiv.2104.08768](https://doi.org/10.48550/arXiv.2104.08768)]
32. Schick T, Schmid H, Schütze H. Automatically identifying words that can serve as labels for few-shot text classification. arXiv [FREE Full text] Preprint posted online on October 26, 2020 [doi: [10.48550/arXiv.2010.13641](https://doi.org/10.48550/arXiv.2010.13641)]
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. arXiv [FREE Full text] Preprint posted online on August 2, 2023 [doi: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762)]
34. Sharma P, Li Y. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. Preprints [FREE Full text] Preprint posted online on August 6, 2019 [doi: [10.20944/preprints201908.0073.v1](https://doi.org/10.20944/preprints201908.0073.v1)]
35. Zhuang Y, Zhang J, Li X, Liu C, Yu Y, Dong W, et al. ICD_promptLearning. GitHub. 2024. URL: https://github.com/PLA301dbgroup2/ICD_promptLearning [accessed 2024-12-03]
36. Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. *Patterns (N Y)* 2023 Apr 14;4(4):100729 [FREE Full text] [doi: [10.1016/j.patter.2023.100729](https://doi.org/10.1016/j.patter.2023.100729)] [Medline: [37123444](https://pubmed.ncbi.nlm.nih.gov/37123444/)]
37. Gao J, Ruan J, Xiang S, Yu Z, Ji K, Xie M, et al. LAMM: label alignment for multi-modal prompt learning. 2024 Mar 24 Presented at: The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24); February 20-27, 2024; Vancouver, BC, Canada p. 1815-1823 URL: <https://ojs.aaai.org/index.php/AAAI/article/view/27950/27920> [doi: [10.1609/aaai.v38i3.27950](https://doi.org/10.1609/aaai.v38i3.27950)]
38. Xiang S, Gao J, Guan M, Ruan J, Zhou C, Liu T, et al. Learning robust visual-semantic embedding for generalizable person re-identification. arXiv [FREE Full text] Preprint posted online on April 19, 2023 [doi: [10.48550/arXiv.2304.09498](https://doi.org/10.48550/arXiv.2304.09498)]

Abbreviations

ACS: acute coronary syndrome

AUC: area under the receiver operating characteristic curve

BERT: bidirectional encoder representations from transformers

BGE-M3: BAAI general embedding multilinguality, multigranularity, and multifunctionality

CVD: cardiovascular disease

ICD: International Classification of Diseases

MDR: medical domain refinement

MLM: masked language modeling

NLP: natural language processing

PLM: pretrained language model

RoBERTa: robustly optimized BERT pretraining approach

Transformer-XL: Transformer-Extra-Long

XLNet: extreme language network

Edited by A Castonguay; submitted 11.06.24; peer-reviewed by S Mao, M Elbattah, S Xiang; comments to author 27.08.24; revised version received 16.10.24; accepted 19.11.24; published 06.01.25.

Please cite as:

Zhuang Y, Zhang J, Li X, Liu C, Yu Y, Dong W, He K

Autonomous International Classification of Diseases Coding Using Pretrained Language Models and Advanced Prompt Learning Techniques: Evaluation of an Automated Analysis System Using Medical Text

JMIR Med Inform 2025;13:e63020

URL: <https://medinform.jmir.org/2025/1/e63020>

doi: [10.2196/63020](https://doi.org/10.2196/63020)

PMID:

©Yan Zhuang, Junyan Zhang, Xiuxing Li, Chao Liu, Yue Yu, Wei Dong, Kunlun He. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 06.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

The Transformative Potential of Large Language Models in Mining Electronic Health Records Data: Content Analysis

Amadeo Jesus Wals Zurita¹, MD; Hector Miras del Rio¹, MP; Nerea Ugarte Ruiz de Aguirre¹, MD; Cristina Nebrera Navarro¹, MD; Maria Rubio Jimenez¹, MD; David Muñoz Carmona¹, MD, PhD; Carlos Miguez Sanchez¹, MD

Servicio Oncología Radioterápica, Hospital Universitario Virgen Macarena, Andalusian Health Service, Seville, Spain

Corresponding Author:

Amadeo Jesus Wals Zurita, MD
Servicio Oncología Radioterápica
Hospital Universitario Virgen Macarena
Andalusian Health Service
Avenida Dr. Fedriani s/n
Seville, 41009
Spain
Phone: 34 954712932
Email: amadeoj.wals.sspa@juntadeandalucia.es

Abstract

Background: In this study, we evaluate the accuracy, efficiency, and cost-effectiveness of large language models in extracting and structuring information from free-text clinical reports, particularly in identifying and classifying patient comorbidities within oncology electronic health records. We specifically compare the performance of gpt-3.5-turbo-1106 and gpt-4-1106-preview models against that of specialized human evaluators.

Objective: We specifically compare the performance of gpt-3.5-turbo-1106 and gpt-4-1106-preview models against that of specialized human evaluators.

Methods: We implemented a script using the OpenAI application programming interface to extract structured information in JavaScript object notation format from comorbidities reported in 250 personal history reports. These reports were manually reviewed in batches of 50 by 5 specialists in radiation oncology. We compared the results using metrics such as sensitivity, specificity, precision, accuracy, F-value, κ index, and the McNemar test, in addition to examining the common causes of errors in both humans and generative pretrained transformer (GPT) models.

Results: The GPT-3.5 model exhibited slightly lower performance compared to physicians across all metrics, though the differences were not statistically significant (McNemar test, $P=.79$). GPT-4 demonstrated clear superiority in several key metrics (McNemar test, $P<.001$). Notably, it achieved a sensitivity of 96.8%, compared to 88.2% for GPT-3.5 and 88.8% for physicians. However, physicians marginally outperformed GPT-4 in precision (97.7% vs 96.8%). GPT-4 showed greater consistency, replicating the exact same results in 76% of the reports across 10 repeated analyses, compared to 59% for GPT-3.5, indicating more stable and reliable performance. Physicians were more likely to miss explicit comorbidities, while the GPT models more frequently inferred nonexplicit comorbidities, sometimes correctly, though this also resulted in more false positives.

Conclusions: This study demonstrates that, with well-designed prompts, the large language models examined can match or even surpass medical specialists in extracting information from complex clinical reports. Their superior efficiency in time and costs, along with easy integration with databases, makes them a valuable tool for large-scale data mining and real-world evidence generation.

(*JMIR Med Inform* 2025;13:e58457) doi:[10.2196/58457](https://doi.org/10.2196/58457)

KEYWORDS

electronic health record; EHR; oncology; radiotherapy; data mining; ChatGPT; large language models; LLMs

Introduction

Real-world data (RWD) holds immense potential for advancing health care by providing a comprehensive view of patient health, disease progression, and treatment outcomes [1]. However, RWD presents significant challenges due to its diverse sources and formats, such as electronic health records, medical imaging, and laboratory results, each with different standards and terminologies. Much of this data is unstructured, like free-text clinical notes, which are difficult to process and analyze. Additionally, missing information is common, leading to gaps that hinder accurate analysis. Advanced methodologies and technologies are needed to effectively extract, standardize, and analyze RWD, ensuring its potential to improve health care outcomes is fully realized.

Extracting information from clinical texts has traditionally relied on manual methods, where trained health care professionals review and annotate clinical notes to identify relevant information such as diagnoses, treatments, and patient outcomes. This manual process is not only time-consuming and labor-intensive but also prone to human error, leading to inconsistencies and inaccuracies. Additionally, statistical and rule-based approaches have been used, which depend on predefined patterns and keywords to extract information. However, these methods often fall short in handling the complexity and variability inherent in natural language, resulting in incomplete or inaccurate data extraction.

The rise of artificial intelligence, driven by advances in computing power, has propelled the development of natural language processing (NLP). NLP algorithms can automatically structure information from unstructured clinical texts, facilitating analysis and integration with other clinical data [2-5]. Earlier NLP systems often relied on rule-based systems and simpler machine learning models, implying limitations such as the need for extensive customization, deep computer science knowledge, significant computational resources, and large volumes of high-quality labeled data. These challenges hinder their widespread adoption and optimal performance across different applications.

Transformer models, a deep learning architecture introduced in the paper "Attention is All You Need" by Vaswani et al [6], have revolutionized the field of NLP, establishing themselves as the foundation upon which modern large language models (LLMs) have been developed. LLMs, such as OpenAI's generative pre-trained transformers (GPTs), are models trained on vast amounts of text to learn complex linguistic patterns. This enables them to generate text, understand context, perform translations, and carry out other tasks with unprecedented accuracy and fluency. Thanks to this capability, users can interact with these models, instructing them to tackle various problems without the need for additional training.

The GPT-3 model, released in 2020, and its successor, GPT-4 [7], introduced in 2023, represent significant advancements in the ability to understand and generate coherent text. The progression from GPT-3 through GPT-3.5 to GPT-4 marks a significant evolution in OpenAI's language model capabilities. GPT-4 offers enhanced understanding and generation of text due to its larger training dataset and more refined architecture, resulting in responses that are more accurate, contextually aware, and nuanced compared to its predecessors. This latest version also demonstrates improved performance on a broader array of tasks, including complex reasoning and problem-solving. Additionally, it is multimodal, capable of processing not only text but also images and audio. However, it is important to note that these models are not specifically designed for medical diagnostic purposes.

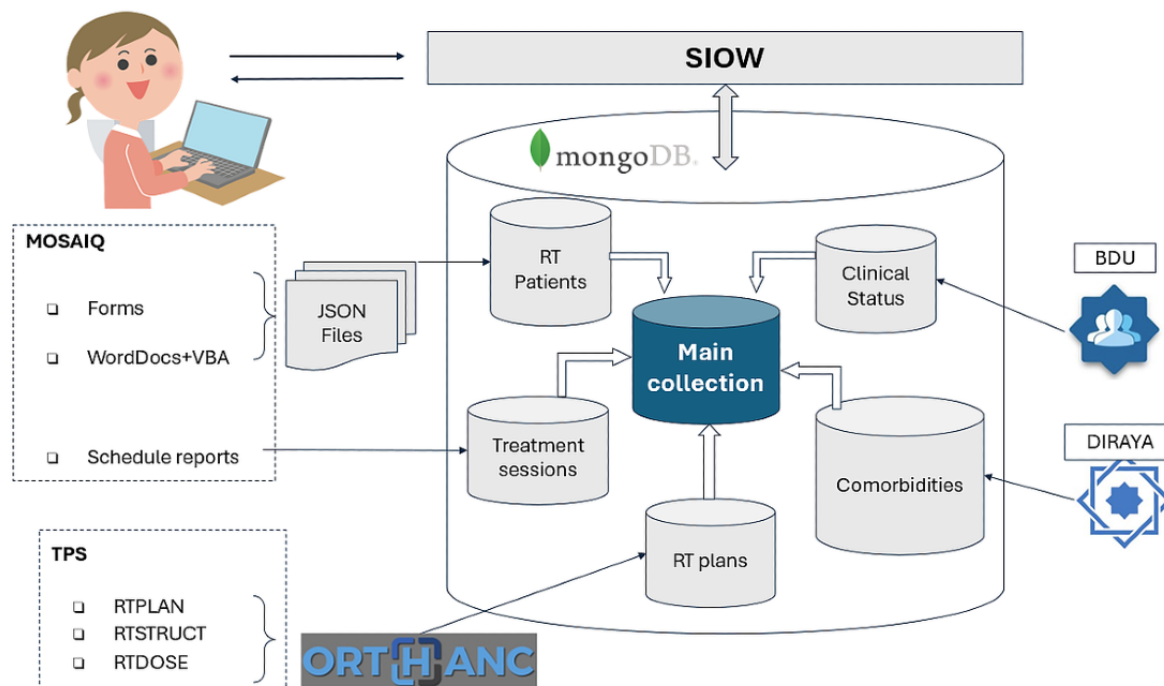
Currently, there are numerous LLMs available, such as LLaMA, Mistral, Claude, or BioBERT. However, in the medical field, the ChatGPT models have been the most extensively studied [8], demonstrating strong capabilities in various applications, including interpreting clinical guidelines and enhancing evidence-based medicine [9], or table summarization in clinical study reports [10]. Despite their potential, concerns about the applicability of these general-purpose models in the medical domain persist [11,12], particularly due to their lack of transparency in training data, which remains largely unknown. Therefore, it is essential to evaluate their performance for each specific application.

In the context of extracting and structuring information from free-text clinical reports, studies have shown promising results with OpenAI models. For instance, Fink et al [13] demonstrated the effectiveness of these models in extracting data from computed tomography reports related to lung cancer, where they outperformed traditional NLP models in classifying disease progression.

Focusing on the significance of appropriate instructions (prompts), studies such as that of Choi et al [14] highlighted that the gpt-3.5-turbo model exhibited an accuracy rate of 87.7% in extracting information from pathology and ultrasound reports of breast cancer patients. Additionally, the LLM methods demonstrated superior efficiency in terms of time and costs compared to manual approaches.

In 2018, the Department of Radiation Oncology at Hospital Universitario Virgen Macarena initiated the implementation of the Mosaiq system, transitioning toward a paperless workflow and centralizing all radiation therapy treatment data within the application. As detailed by Bertolet et al [15], this data was automatically exported to JSON files via Word documents and Visual Basic for Applications code. Figure 1 depicts a diagram illustrating the flow and organization of the described data.

Figure 1. Representative diagram of the Web Oncological Information System (SIOW). It illustrates the integration of data from MOSAIQ and TPS into the MongoDB database and its subsequent management through SIOW, including the collection of administrative data from the Users Data Base (BDU) and clinical data from the electronic health record system DIRAYA. JSON: JavaScript object notation; RT: radiotherapy.



Motivated by the capabilities of LLMs, we aimed to investigate their potential application in extracting and structuring information from clinical reports. Our overarching objective is to integrate LLM-based tools into our information system, enhancing the richness of our real-world datasets. Specifically, in this study, we assess the capability of the GPT-3.5 turbo and GPT-4 models as tools for data mining applied to the identification and classification of comorbidities and relevant lifestyle risk factors in oncological texts. We compare their performance against that of specialized human evaluators to gauge their efficacy and suitability for clinical use.

Methods

OpenAI Models

The application programming interface (API) of OpenAI [16] allows interaction with their advanced LLMs, facilitating various language processing tasks such as generating automatic textual responses, conducting sentiment analysis, and summarizing texts. In our study, we leveraged the *chat completions API* function of the API to extract structured information from unstructured clinical reports.

OpenAI offers a comprehensive library of natural language processing models. Each model features unique characteristics in terms of size, language comprehension ability, speed, and cost. In our study, we have used 2 models from the library: *gpt-3.5-turbo-1106* and *gpt-4-1106-preview*, with the latter being the most advanced model available at the time the study was conducted. While the GPT-3.5 model is a faster and more economical option for general tasks, GPT-4 stands out for its higher accuracy, contextual understanding, and ability to handle more complex and specific applications.

For this study, we used clinical reports in Spanish, exclusively interacting with OpenAI's LLMs in this language. Although LLMs typically exhibit superior performance in English [17], owing to the predominance of this language in training data, recent comparisons indicate notable effectiveness in other languages, including Spanish. The GPT-4 technical report [7] highlights this multilingual capability, demonstrating that performance in Spanish closely approaches that of English, with a minimal difference of only 1.5 percentage points in the MMLU evaluation [18].

Prompt Generation

To interact with the LLM models, we first created a prompt that will guide the model through the specific task. The context provided to the model establishes a scenario in which it is asked to assume the role of a specialist in radiation oncology. This setting serves as a reference framework, enabling the model to adopt the appropriate perspective and apply its natural language understanding capabilities in a manner consistent with the medical domain.

Our request is a direct instruction to the model, directing it to process the text of the provided clinical report and return the relevant information in a structured format. Specifically, the model is instructed to use the clinical report provided at the end of the prompt to complete a predefined dictionary in JSON format. This dictionary contains keys related to comorbidities and lifestyle risk factors. The model is tasked with updating the values of these keys with "YES" or "NO" as appropriate. For individuals who are ex-smokers, the model should use "EX" instead. Additionally, the model must identify and add any other relevant comorbidities not classifiable under the provided categories, assigning them to the "Other" key.

The prompt generated for the task is shown in [Textbox 1](#).

The dictionary mentioned in the request is structured with keys labeling the specific comorbidities and lifestyle risk factors we seek to identify. These comorbidities, along with their potential values, are outlined in [Table 1](#).

During a postprocessing phase, we divided the category labeled as “smoker” into 2 distinct categories: “smoker” (representing current smokers) and “ex-smoker.” This division was implemented to ease the subsequent analysis of the results.

Textbox 1. Prompt generated for the task.

- **Context:** “Act as a specialist in radiation oncology.”
- **Request:** “Use the clinical report provided at the end of this prompt to return in JSON format the dictionary [...] with the values 'YES' or 'NO'. For the 'Smoker' field: 'YES' if they smoke, 'NO' if they have never smoked, 'EX' if they are an ex-smoker. For the 'Other' field, return a list of comorbidities found that cannot be classified in any of the categories of the keys of the provided dictionary, or empty if there are no other comorbidities. Return only the dictionary with the updated values, DO NOT ADD OR MODIFY KEYS. Clinical report: [text of the clinical report]”

It is important to highlight that the prompt does not provide context or additional instructions regarding how the specified comorbidities of interest should be interpreted.

The development of this prompt was achieved through an iterative process applied to a group of 50 reports that were specifically reserved for this purpose. The methodology included the following steps ([Textbox 2](#)):

Table 1. List of the labels, possible values, and description of the comorbidities and lifestyle risk factors considered in this study.

| Label | Values | Description |
|-----------------|-----------------|--|
| Diabetes | Yes or no | Elevated blood glucose levels |
| HBP | Yes or no | High blood pressure |
| Smoker | Yes or no or ex | Smoking habit. |
| Dyslipidemia | Yes or no | Lipid metabolism disorder |
| Liver disease | Yes or no | Liver disease |
| COPD | Yes or no | Chronic obstructive pulmonary disease |
| Depression | Yes or no | Mood disorder |
| Kidney disease | Yes or no | Kidney disease |
| Fentanyl | Yes or no | Use of WHO step 3 analgesics (opioids) |
| Heart disease | Yes or no | Heart disease |
| Hyperthyroidism | Yes or no | Thyroid disease with increased thyroxine |
| Hypothyroidism | Yes or no | Thyroid disease with decreased thyroxine |
| Dependent | Yes or no | Patient in need of continuous care |
| Other | Text list | Other past comorbidities detected not listed above |

Textbox 2. Prompt development methodology.

- **Prompt definition:** Establishing the parameters and structure of the prompt to guide the model's responses.
- **Information extraction:** The developed prompt was applied to 50 reports using the *gpt-4-1106-preview* model.
- **Verification of structure:** It was ensured that the model's responses adhered to the requested structure, with previous steps being repeated in case of deviations.
- **Accuracy evaluation:** A specialist physician (AW) verified the accuracy of the model's responses. This process was repeated until the accuracy met or exceeded that of a manual analysis performed by the same physician.

Python Script

The Python script developed uses the OpenAI API to automatically structure textual clinical information. All the code developed for this work is publicly available in a GitHub repository [19].

Clinical Report Acquisition Procedure

The clinical reports for our study were provided by the hospital's Innovation & Data Analysis department. These reports were delivered in an Excel spreadsheet format, organized into 2 essential columns: one containing the clinical history number of each patient and another with the text of the medical personal history report. The department responsible for data collection

undertook a process of anonymization and randomization of the reports to ensure an unbiased selection.

Sample Selection Criteria

For estimating the sample size, we relied on the proportion of comorbidities (80%) obtained from a prior analysis of a dataset of 5257 personal history reports from patients treated in our service between May 2018 and October 2022.

The comorbidities selected for the study were chosen based on prior knowledge of prevalences in the general population and those presented by our patients according to the aforementioned analysis. We also considered those that could most significantly impact the clinical outcome of oncological treatments.

With these considerations, we conducted a preliminary calculation that established the need to include 250 clinical reports (see below in the statistical analysis section). Based on this calculation, we selected the first 250 patients from the provided list who had a nonempty personal history report. Before proceeding with the analysis, we verified that our script was capable of correctly interpreting an empty report as equivalent to the absence of comorbidities, thereby avoiding biases in the study results.

Ethical Considerations

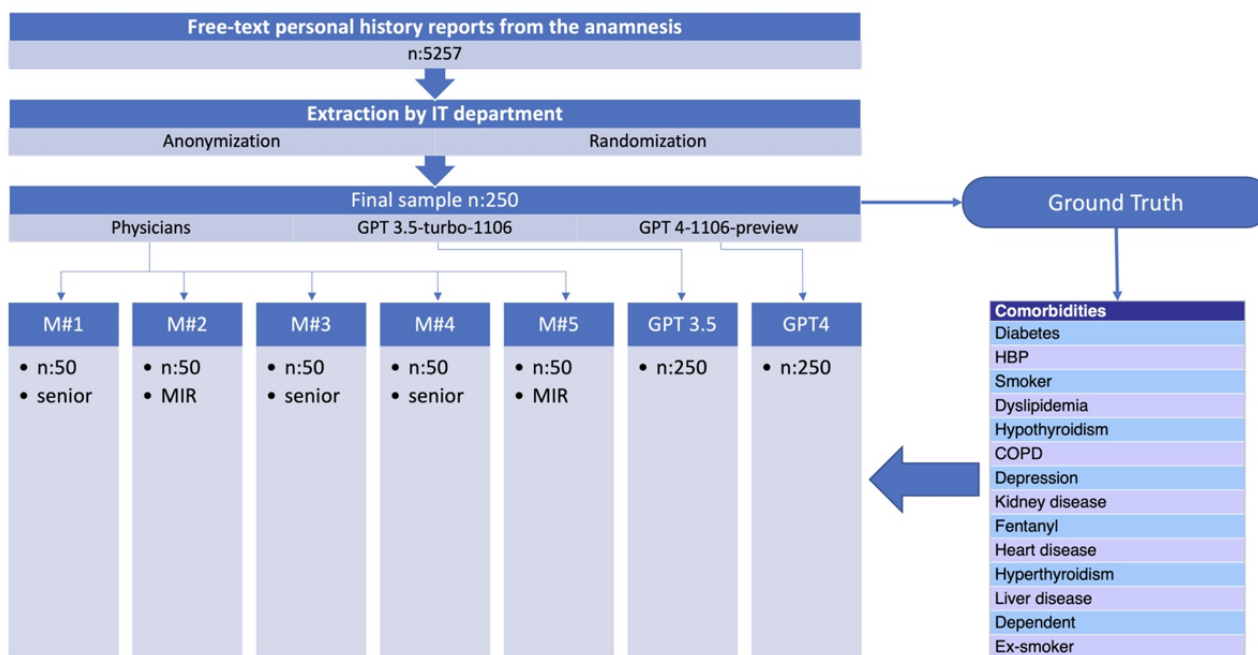
The text processed by the selected LLMs is strictly confined to personal history reports. These reports were stripped of any information that could lead to patient identification, ensuring confidentiality and anonymity. The model’s interpretation of the texts focuses solely on identifying and structuring data relevant to the study without compromising individual privacy.

The study’s design, synthesized in Figure 2, and methodology were previously communicated to and reviewed by the hospital’s ethics committee. The research received the necessary approval, confirming that it adheres to the ethical standards required for patient data research.

This retrospective study adheres to the guidelines outlined in the *seventeenth additional provision, specifically Health Data Processing, Section d) of the Organic Law 3/2018, dated December 5, on Personal Data Protection and Guarantee of Digital Rights*. This law governs the use of pseudoanonymized personal data for health research purposes. The study was granted an exemption from requiring informed consent due to its exclusive use of nonidentifiable data.

On January 18, 2024, the Ethics Committee of the University Hospitals Virgen Macarena and Virgen del Rocío issued a favorable opinion for our study, under the reference EC_IA_V1 (version 1-Dec-2023).

Figure 2. Flowchart of the study design. COPD: chronic obstructive pulmonary disease; HBP: high blood pressure.



Data Extraction

For the manual data extraction, the 250 patient clinical reports were divided into 5 groups, each consisting of 50 reports. These groups were randomly assigned to 5 physicians, including 3 specialists in radiation oncology with more than 15 years of experience and 2 medical residents in the same specialty, one in their first year and the other in their fourth year.

To ensure uniform and accurate data collection, the physicians were provided with a specially designed template for this task. The template features a table where the first column contains the full texts of the clinical reports. The subsequent columns of the table are labeled with the comorbidities of interest. The cells corresponding to each comorbidity only allow the selection of predefined values, as stipulated in Table 1. This restriction ensures consistent annotation and reduces the possibility of errors or variations in the entries.

For the automatic analysis, the 250 clinical reports in the sample were analyzed using our script with the *gpt-3.5-turbo-1106* and *gpt-4-1106-preview* models. To maintain a consistent structure in the study, these reports were organized into the same 5 groups of 50 reports that were assigned to the physicians. The results were recorded in a document that mirrored the structure of the template used in the manual extraction. This uniformity in documentation facilitates a direct comparison of results between manual and automatic extraction methods.

Establishing the Ground Truth

To assess the comparative accuracy and effectiveness of the LLMs used in this study against the evaluations performed by physicians, it is crucial to establish a reference dataset containing the ground truth. To construct this reference dataset, we first compared the results obtained from the physicians and the *gpt-4-1106-preview* model across all 250 reports, identifying and recording any discrepancies between the 2 sources. The radiation oncologist expert AW, with more than 30 years of experience, reviewed several times the whole set of reports, with a particular focus on these discrepancies. For each report where discrepancies in the results were found, physician AW assessed both responses (from the physician and the AI) and determined which one was correct.

It is important to note that the ground truth in this study is based solely on the information explicitly reported in the clinical texts. This means that some patients may have unreported comorbidities, or conversely, conditions may be mentioned that are not actually present. This limitation reflects a common challenge when working with RWD. However, for the purposes of this study, these potential discrepancies are irrelevant, as our primary focus is on evaluating the models' ability to accurately interpret and extract information from the provided texts.

Assessing Reproducibility in Results

The nondeterministic nature of LLMs, such as GPT-3.5 and GPT-4, means they can generate different responses to identical requests [7]. This phenomenon, coupled with the potential for

periodic retraining of the models, significantly impacts the reproducibility of results. Therefore, it is crucial to consider the need for rigorous quality control for algorithms that use LLMs, especially to assess the impact of any changes in the models.

A well-defined and explicit prompt can increase the reproducibility of responses [14]. However, variability remains a possibility, particularly in situations where the information is ambiguous or the prompt is not clear or specific enough.

To measure the consistency of our automatic extraction method, we repeated the analysis of the 250 clinical reports 10 times over 10 consecutive days. This approach allows us to observe the stability of the model responses to the same input.

Statistical Analysis

To ensure the statistical validity of the study, a significance level of 5% (alpha error) and a power of 80% (beta error of 20%) were established. Additionally, a 5% error margin was applied for 95% confidence intervals. With these considerations in mind, it was determined that the sample size (n) should include 245 patient records. To adjust the sample to a practical number, it was rounded up, resulting in a final sample size of 250.

For a comprehensive analysis, we consolidated the results from the 250 reports into a single category named "Physicians," representing the aggregated findings of the 5 doctors involved in the study. Subsequently, we compared this category and the results from the GPT-3.5 and GPT-4 models with the reference dataset, considered as the ground truth. In this process, a confusion matrix was created for each report and comorbidity, from which several key statistical estimators were derived.

To assess the agreement, we used the κ index. The McNemar test was used to determine if there were significant differences in the proportions of discordance between the classifications. We chose the F-score as a measure of balance between precision and sensitivity, which is crucial in a classification model. The calculated metrics are presented in Table 2.

Table 2. Metrics used in the study with their descriptions.

| Metric | Description |
|-------------|---|
| TP | True positives |
| TN | True negatives |
| FP | False positives |
| FN | False negatives |
| Sensitivity | $TP/(TP+FN)$ |
| Specificity | $TN/(FP+TN)$ |
| Precision | $TP/(TP+FP)$ |
| Prevalence | $(TP+FN)/(TP+TN+FP+FN)$ |
| Accuracy | $(TP+TN)/(TP+TN+FP+FN)$ |
| Kappa | $(Pobs - Pesp)/(1 - Pesp)$ |
| F-score | $(2 \times \text{precision} \times \text{sensitivity}) / (\text{precision} + \text{sensitivity})$ |
| McNemar | Exact <i>P</i> value from McNemar test (binomial distribution) |

For some of these metrics, we calculated their CI using the bootstrapping method [20]. This approach starts from the frequencies of true positives, true negatives, false positives, and false negatives to generate 1000 resamples. With these resamples, we recalculated the metrics to obtain a distribution that allows us to calculate the 95% CI.

Additionally, a detailed analysis was conducted on the groups of 50 reports assigned to each physician. This analysis focused on measuring the variability in evaluations among different physicians. For each patient and comorbidity, Cohen κ index was calculated in comparison with the ground truth for the results of each physician.

Textbox 3. Nature of the detected errors.

- **Differences in criteria:** Variations in the interpretation of the relevance of reported pathologies.
- **Incorrect interpretation:** Misunderstandings caused by confusing wording.
- **Incorrect inference:** Erroneous deductions when the comorbidity is not explicitly mentioned.
- **Ambiguous text:** Textual ambiguity that allows for multiple interpretations.
- **Error or hallucination:** Unjustified errors, attributed to human distractions or AI hallucinations.
- **Error in ground truth:** Corrections made upon review that validate the evaluator's interpretation.
- **Explicit omission:** Overlooking direct mentions of comorbidities.
- **Omission by context:** Failure to notice comorbidities deducible from the context or medication.
- **Unrecognized acronyms:** Inability to interpret specific medical acronyms.

The reproducibility of the GPT-3.5 and GPT-4 models was assessed by quantifying the number of different responses for each patient and comorbidity across the 10 repeated analyses conducted on successive days.

Analysis of Discrepant Results

A detailed analysis of discrepancies between the evaluators' results and the established Ground Truth was conducted by the same physician who defined the reference dataset. This analysis covered each report with discrepancies in the identification of comorbidities, identifying the probable causes of each deviation.

Discrepancies were classified according to the nature of the detected errors (Textbox 3).

Results

Cost and Time Analysis

Table 3 details the cost and total time invested in analyzing the 250 reports using the GPT-3.5 and GPT-4 models. Given that both the models and their associated costs can fluctuate over time, it is important to note that the reported results are specific to the usage period from January to February 2024. It is noted

that GPT-4, being a larger and more complex LLM compared to GPT-3.5, incurs longer processing times and a cost approximately 10 times higher. Extrapolating the costs to the entire set of 7500 patients currently registered in our database, processing with GPT-4 would require about 24 hours and would cost approximately 76 dollars. On the other hand, using GPT-3.5 would reduce the processing time to about 9 hours, with a significantly lower cost of around 7 dollars.

Table 3. Execution times and costs in dollars for the analysis of the 250 reports with each of the models used (usage period of the models: between January and February 2024).

| Model | N report | Time (hour) | Cost (US \$) |
|--------------------|----------|-------------|--------------|
| gpt-3.5-turbo-1106 | 250 | 0.31 | 0.23 |
| gpt-4-1106-preview | 250 | 0.79 | 2.53 |

Prevalences

The analysis of our Ground Truth sample reveals a wide range of prevalences in comorbidities and lifestyle risk factors among oncological patients. These are detailed in Table 4, where both the number of cases and the prevalence for each comorbidity are reported. The most common conditions include high blood

pressure and dyslipidemia, present in almost half and a third of the cases, respectively. On the other hand, conditions like hyperthyroidism and liver disease show relatively low prevalence. Categories related to smoking are also highly frequent, accounting for almost 50% of the cases. Interestingly, the proportion of ex-smokers significantly exceeds that of current smokers.

Table 4. Number of reports, out of the total 250 in the sample, that indicate each comorbidity and the corresponding prevalence.

| Condition | Cases, n | Prevalence |
|-------------------|----------|------------|
| Diabetes | 64 | 25.6% |
| HBP ^a | 116 | 46.4% |
| Smoker | 37 | 14.8% |
| Dyslipidemia | 77 | 30.8% |
| Hypothyroidism | 21 | 8.4% |
| COPD ^b | 17 | 6.8% |
| Depression | 25 | 10.0% |
| Kidney disease | 39 | 15.6% |
| Fentanyl | 19 | 7.6% |
| Heart disease | 43 | 17.2% |
| Hyperthyroidism | 1 | 0.4% |
| Liver disease | 13 | 5.2% |
| Dependent | 12 | 4.8% |
| Ex-smoker | 85 | 34.0% |

^aHBP: high blood pressure.

^bCOPD: chronic obstructive pulmonary disease.

Evaluation Metrics

Table 5 displays the values of true positives, false positives, true negatives, and false negatives, detailed by comorbidity, derived from the comparison with the ground truth dataset.

Figure 3 illustrates the performance of the physicians, GPT-3.5, and GPT-4 classifiers, broken down by comorbidity, across various metrics. The “Total” category, which consolidates the

results for all studied comorbidities, enables direct comparison between the 3 evaluators on each assessed metric (**Textbox 4**).

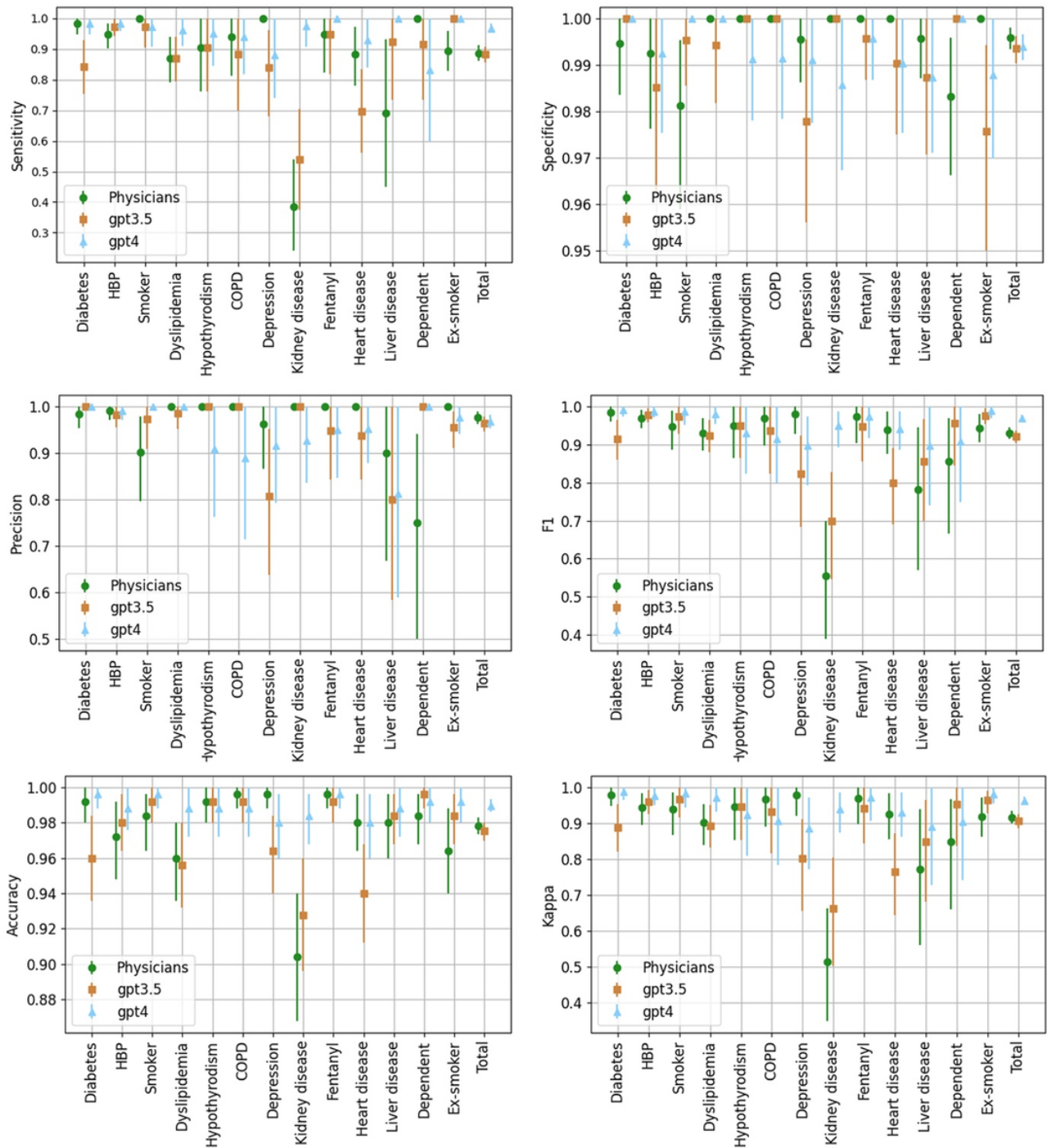
The application of McNemar's test to the “Total” category, comparing Physicians with GPT-3.5 and Physicians with GPT-4, yielded *P* values of .79 and 10^{-6} , respectively. This confirms that the performance differences between the physicians and the GPT-3.5 model are not statistically significant, while the differences between the physicians and GPT-4 are significant.

Table 5. Tables displaying the results for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each comorbidity, obtained by each of the evaluators (Physicians, GPT-3.5, and GPT-4).

| | Physicians | | | | GPT-3.5 | | | | GPT-4 | | | |
|-------------------|-----------------|-----------------|-----------------|-----------------|------------|-------------|-----------|-----------|------------|-------------|-----------|-----------|
| | TP ^a | TN ^b | FP ^c | FN ^d | TP | TN | FP | FN | TP | TN | FP | FN |
| Diabetes | 63 | 185 | 1 | 1 | 54 | 186 | 0 | 10 | 63 | 186 | 0 | 1 |
| HBP ^e | 110 | 133 | 1 | 6 | 113 | 132 | 2 | 3 | 114 | 133 | 1 | 2 |
| Smoker | 37 | 209 | 4 | 0 | 36 | 212 | 1 | 1 | 36 | 213 | 0 | 1 |
| Dyslipidemia | 67 | 173 | 0 | 10 | 67 | 172 | 1 | 10 | 74 | 173 | 0 | 3 |
| Hypothyroidism | 19 | 229 | 0 | 2 | 19 | 229 | 0 | 2 | 20 | 227 | 2 | 1 |
| COPD ^f | 16 | 233 | 0 | 1 | 15 | 233 | 0 | 2 | 16 | 231 | 2 | 1 |
| Depression | 25 | 224 | 1 | 0 | 21 | 220 | 5 | 4 | 22 | 223 | 2 | 3 |
| Kidney disease | 15 | 211 | 0 | 24 | 21 | 211 | 0 | 18 | 38 | 208 | 3 | 1 |
| Fentanyl | 18 | 231 | 0 | 1 | 18 | 230 | 1 | 1 | 19 | 230 | 1 | 0 |
| Heart disease | 38 | 207 | 0 | 5 | 30 | 205 | 2 | 13 | 40 | 205 | 2 | 3 |
| Hyperthyroidism | 0 | 249 | 0 | 1 | 0 | 249 | 0 | 1 | 1 | 249 | 0 | 0 |
| Liver disease | 9 | 236 | 1 | 4 | 12 | 234 | 3 | 1 | 13 | 234 | 3 | 0 |
| Dependent | 12 | 234 | 4 | 0 | 11 | 238 | 0 | 1 | 10 | 238 | 0 | 2 |
| Ex-smoker | 76 | 165 | 0 | 9 | 85 | 161 | 4 | 0 | 85 | 163 | 2 | 0 |
| <i>Total</i> | <i>505</i> | <i>2919</i> | <i>12</i> | <i>64</i> | <i>502</i> | <i>2912</i> | <i>19</i> | <i>67</i> | <i>551</i> | <i>2913</i> | <i>18</i> | <i>18</i> |

^aTP: true positive.^bTN: true negative.^cFP: false positive.^dFN: false negative.^eHBP: high blood pressure.^fCOPD: chronic obstructive pulmonary disease.

Figure 3. Statistical metrics comparison between 3 evaluators (Physicians, GPT-3.5, and GPT-4) for individual comorbidities and overall totals. Asymmetric error bars indicate the 95% confidence interval. GPT: generative pre-trained transformer. HBP: hypertension or high blood pressure; COPD: chronic obstructive pulmonary disease.



Textbox 4. Summary of the metrics evaluated.

- **Sensitivity:** The GPT-4 model (96.8%) outperforms both GPT-3.5 (88.2%) and the physicians (88.8%) in most categories, showing notable effectiveness in detecting comorbidities. Although GPT-3.5 presents slightly lower results than the physicians, the difference is not statistically significant, as indicated by the overlap of the 95% confidence intervals shown in [Figure 3](#).
- **Specificity:** All evaluators achieve high specificity values, which is expected given the low prevalences of the studied comorbidities and the relative ease of identifying the absence of a comorbidity in texts. The physicians (99.6%) excel in this metric, often achieving perfection, while both models (99.4%) score slightly lower due to a higher rate of false positives.
- **Precision:** The physicians get the highest score (97.7% vs 96.4% and 96.8%) assessing the proportion of correct positive identifications, possibly also influenced due to the models generating a higher number of false positives.
- **F-score:** Representing the harmonic mean between precision and sensitivity, the F-score is particularly relevant in asymmetric samples like in our study. The GPT-4 model achieves the highest score (96.8%) on this indicator, surpassing both GPT-3.5 (92.1%) and the physicians (93%).
- **Accuracy (Agreement):** In the proportion of correct identifications, GPT-4 shows superior performance (99%), while GPT-3.5 (97.5%) and the physicians (97.8%) achieve similar results.
- **Cohen κ index:** This index, measuring agreement adjusted for chance, reveals that GPT-4 reaches the highest scores (0.962), demonstrating greater consistency compared to the ground truth. The GPT-3.5 score of 0.907, while marginally lower, does not significantly differ from the physicians' score of 0.917.

Variability Among Physicians' Performance

[Table 6](#) displays the Cohen κ index values obtained in the detection of various comorbidities for each of the 5 physician evaluators. It is important to note that each physician analyzed a different group of 50 reports.

Overall, there was considerable similarity in the physicians' responses, except when the comorbidity to be detected was a broader concept, as in the case of "kidney disease" ($\kappa=0.51$) or

"liver disease" ($\kappa=0.77$). It is important to note that no further instructions or explanations were provided beyond finding the comorbidity in the presented text. Therefore, some physicians considered that renal lithiasis was not a relevant "kidney disease" and reserved this category for conditions describing an alteration in renal function (such as chronic renal failure, for example).

Interestingly, the senior physicians scored lower than the medical residents in the overall calculation for the κ index.

Table 6. Concordance values for each comorbidity, calculated using Cohen κ index for each medical evaluator. The "Total" categories summarize the aggregated concordance across all comorbidities and medical evaluators. A dash indicates that the κ index could not be computed because the comorbidity was not present in the corresponding set of reports.

| | M1 senior | M2 resident | M3 senior | M4 senior | M5 resident | Total human evaluators |
|-------------------|-----------|-------------|-----------|-----------|-------------|------------------------|
| Diabetes | 1.00 | 0.95 | 1.00 | 1.00 | 0.95 | 0.98 |
| HBP ^a | 1.00 | 0.96 | 0.83 | 0.96 | 0.96 | 0.94 |
| Smoker | 1.00 | 1.00 | 0.88 | 0.86 | 0.93 | 0.94 |
| Dyslipidemia | 0.91 | 1.00 | 0.75 | 0.77 | 1.00 | 0.90 |
| Hypothyroidism | 0.66 | 1.00 | 0.90 | 1.00 | 1.00 | 0.95 |
| COPD ^b | 1.00 | 1.00 | 1.00 | 0.66 | 1.00 | 0.97 |
| Depression | 0.93 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 |
| Kidney disease | 0.52 | 0.70 | 0.45 | 0.56 | 0.26 | 0.51 |
| Fentanyl | 1.00 | 1.00 | 0.85 | 1.00 | 1.00 | 0.97 |
| Heart disease | 0.95 | 1.00 | 0.91 | 0.79 | 1.00 | 0.93 |
| Hyperthyroidism | — | — | — | — | 0.00 | 0.00 |
| Liver disease | — | 1.00 | 0.63 | 0.65 | 1.00 | 0.77 |
| Dependent | 0.66 | 0.66 | 0.91 | 0.88 | — | 0.85 |
| Ex-smoker | 0.95 | 1.00 | 0.87 | 0.76 | 1.00 | 0.92 |
| Total | 0.95 | 1.00 | 0.87 | 0.76 | 1.00 | 0.92 |

^aHBP: high blood pressure.

^bCOPD: chronic obstructive pulmonary disease.

Reproducibility of Models' Responses

In our reproducibility study, each report was analyzed 10 times by the GPT-3.5 and GPT-4 models. For each comorbidity, we counted the number of different responses generated in these repeated analyses, as well as the total number of variations for each report.

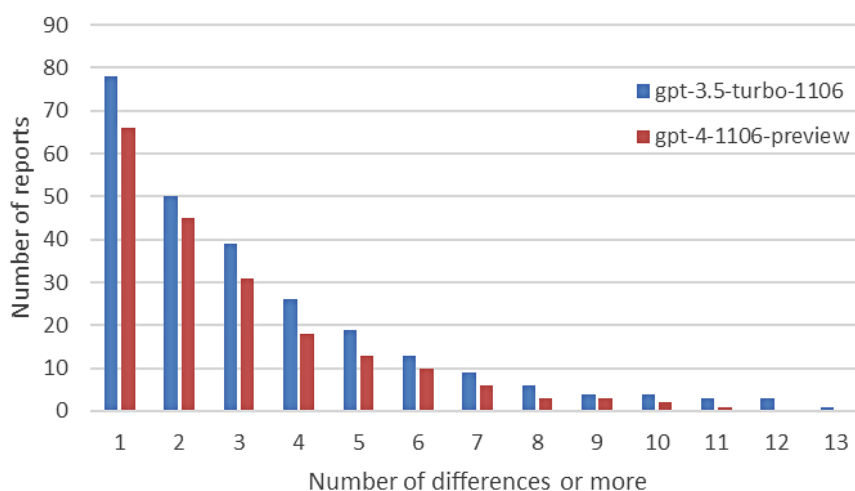
Figure 4 presents a histogram illustrating the number of reports that generated at least the specified number of different responses. This histogram reveals that, in all instances, the GPT-4 model exhibited fewer differences in responses compared to GPT-3.5, suggesting greater consistency and reliability in its results.

Furthermore, it was found that 73.6% of the reports analyzed with GPT-4 reproduced the same result across all comorbidities during the 10 analyses, compared to 59.2% for GPT-3.5. This notable difference in reproducibility underscores the superiority

of GPT-4 in maintaining consistency in its responses across multiple executions.

Variability in responses often stems from ambiguous text, where LLMs may assign values inconsistently. For example, a report describing a patient as an "active smoker (1 month since quitting, 1 pack/day since age 14-16)" resulted in GPT-3.5 identifying the patient as a smoker in 6 out of 10 analyses, while GPT-4 made only 1 error across 10 analyses. However, in the same report, regarding the comorbidity of COPD, GPT-4 shows a split: in 5 instances, it identifies it as present and in 5 as absent. The physician reviewing the results and establishing the ground truth determined the absence of COPD, as it is not explicitly mentioned in the report. Nonetheless, the mention of "mild pulmonary emphysema areas" and the patient's prolonged smoking history could lead GPT-4 to infer the presence of COPD.

Figure 4. The number of reports for each model, in which at least the number of differences indicated on the x-axis was obtained in the 10 analyses.



Discrepancy Analysis

Multimedia Appendices 1 and 2 display the distribution of discrepant results categorized by the causes determined through a detailed manual analysis of the reports.

A notable discrepancy arose in the "kidney disease" category due to differences in criteria. Some physicians and GPT-3.5 did not deem certain renal pathologies, such as renal lithiasis, as relevant comorbidities in the context of oncology treatment, unlike GPT-4, which aligned its results more closely with the ground truth.

In analyzing cases interpreted as hallucinations, it was found that this phenomenon occurred exclusively in 1 response from GPT-4 and in 6 from GPT-3.5, particularly in the smoker and ex-smoker categories, possibly due to the use of the label "toxic habits," even when referring to other habits like alcoholism.

The models, especially GPT-4, tend to infer comorbidities from the context or reported medication more frequently than physicians, who exhibit a more conservative approach. This tendency leads to more false positives by the models, particularly when the medication does not imply the presence of comorbidity.

GPT-3.5 exhibited difficulties in interpreting common medical acronyms such as "DM" for diabetes or "AF" for atrial fibrillation, whereas GPT-4 demonstrated a superior ability to recognize and correctly interpret most of these acronyms.

Interestingly, GPT-4 displayed some false positives when encountering comorbidity labels followed by ":" without additional information, a misinterpretation not common in humans but observed in AI, particularly in GPT-4 more than in GPT-3.5.

Human evaluators showed a greater tendency to overlook comorbidities explicitly reported, likely due to distraction or fatigue.

Only 3 errors were identified in the determination of the ground truth, underscoring the reliability of the review process.

Finally, we identified a category of discrepancies exclusive to the models, related to structural or formatting errors. This includes situations where the models' responses do not follow the guidelines specified in the prompt, resulting in outputs that do not meet the expected JSON format or that incorrectly alter and introduce comorbidity labels. Given that these incidents were limited, affecting less than 10 cases, it was decided to

manually correct these formatting errors for inclusion in the subsequent analysis.

Discussion

Principal Findings

Our study categorizes observers as "Physicians," "GPT-3.5," and "GPT-4," reflecting the synergy between specific models (*gpt-3.5-turbo-1106* and *gpt-4-1106-preview*) and the prompts designed for this research. The effectiveness of GPT models in generating responses is inherently linked to the quality and structure of the prompts [14,21,22], indicating that results may vary significantly with prompt redefinition. Similarly, physician performance is influenced not only by clinical competence but also by the clarity of instructions and the quality of the materials provided. Offering more detailed and specific guidelines, along with access to additional sources within the electronic health records, could potentially improve the accuracy of their responses.

It is important to emphasize that even if LLMs demonstrate superiority in the specific task of processing large volumes of reports to extract information, this should not be extrapolated to other tasks, such as decision-making. In such cases, these tools should always be used as support tools, requiring ongoing physician oversight and intervention.

Based on the results obtained, we can conclude that the GPT-4 model is notably better at identifying present comorbidities, with fewer false negatives, while physicians exhibit slightly higher precision in their diagnoses, resulting in fewer false positives. The GPT-3.5 model generally performs slightly below the physicians, though the differences found are not statistically significant. These results are consistent with findings from other studies, such as Hoppe et al [23], which highlight the potential of ChatGPT models to enhance diagnostic accuracy in emergency medical settings. In their study, GPT-4 also outperformed both resident physicians and GPT-3.5 in diagnostic accuracy.

The superior sensitivity of GPT-4 in our study is particularly noteworthy, demonstrating its advanced ability to accurately identify reported comorbidities, even when not directly evident in the text. However, both GPT-3.5 and GPT-4 generate a comparable number of false positives, which is significantly higher than those recorded by physicians. Physicians' false positives typically result from specific circumstances such as ambiguity in clinical reports, variations in interpretation among professionals, and occasional errors in the template filling process.

In contrast, false positives from the GPT models seem to stem from a less conservative approach in determining comorbidity presence based on inferred context. These cases are also more likely to produce less reproducible responses due to the

nondeterministic nature of LLMs. In these instances, physicians adopted a more conservative criterion to establish the ground truth, considering an unreported comorbidity only when the medication or context necessarily implied it. Whether this conservative approach is preferable to the criteria used by GPT models requires an analysis of complete medical histories to confirm or refute the presence of the comorbidity.

Discrepancies arising from variations in criteria interpretation could be mitigated by using prompts with clearer instructions on interpreting different comorbidities. This underscores the importance of refining prompts to enhance the consistency and accuracy of LLM-generated responses in clinical contexts.

Despite the remarkable capacity of current LLMs as potential tools for data mining in clinical reports, questions arise regarding the practical utility of this RWD for research and the generation of real-world evidence [24]. The variability, subjectivity, and lack of structure in these reports can compromise the quality and reliability of extracted data, affecting its applicability in clinical research contexts. Therefore, while LLMs represent a promising innovation to address the limitations of unstructured data, implementing more structured clinical recording practices could provide a more sustainable and reliable solution for generating real-world clinical evidence. This duality emphasizes the need for a balanced approach that integrates advanced AI technology with robust clinical data management practices.

Future research should concentrate on refining prompt design and expanding the applications of LLMs across various medical fields. Additionally, exploring the performance of new open-source LLMs that can be run locally is essential, as this approach helps to avoid data protection and privacy issues associated with transmitting clinical data outside of the local infrastructure.

Conclusions

This study has shown that, with carefully designed prompts, the OpenAI LLMs examined demonstrate competence comparable to, and in some cases superior to, that of medical specialists in interpreting and extracting relevant information from clinical reports, even when dealing with complex and ambiguously written texts. Considering their superior efficiency in terms of time and costs, along with their seamless integration with databases and other applications, these models emerge as a preferable option for data mining and structuring information in large collections of clinical reports. This highlights the potential of LLMs to enhance RWD usage by efficiently extracting structured information from extensive volumes of clinical texts, which is crucial for generating high-quality real-world evidence. Nonetheless, continuous evaluation of these models is essential to enhance their accuracy and applicability, while also emphasizing the importance of advancing toward more structured clinical records.

Acknowledgments

Data were pseudonymized by Andalusian Health Service technicians according to the GDPR (General Data Protection Regulation) regulation ensuring the technical and functional separation between the research team and those who perform the pseudonymization.

Authors' Contributions

AWZ and HMR contributed to study idea and design; CMS, DMC, MRJ, NURA, and CNN involved in data collection; data analysis and results interpretation were performed by HMR and AWZ; HMR and AWZ contributed to manuscript writing; critical review and editing were carried out by all authors. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Number of false positive (FP) results attributed to each of the considered causes. Diff: differences; Inco: incorrect; Erro: error. [\[PNG File , 27 KB - medinform_v13i1e58457_app1.png \]](#)

Multimedia Appendix 2

Number of false negative (FN) results attributed to each of the considered causes. [\[PNG File , 29 KB - medinform_v13i1e58457_app2.png \]](#)

References

1. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol* 2022 Nov 05;22(1):287 [FREE Full text] [doi: [10.1186/s12874-022-01768-6](https://doi.org/10.1186/s12874-022-01768-6)] [Medline: [36335315](https://pubmed.ncbi.nlm.nih.gov/36335315/)]
2. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. *JAMA Oncol* 2016 Jun 01;2(6):797-804. [doi: [10.1001/jamaoncol.2016.0213](https://doi.org/10.1001/jamaoncol.2016.0213)] [Medline: [27124593](https://pubmed.ncbi.nlm.nih.gov/27124593/)]
3. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. *Cancer Res* 2019 Nov 01;79(21):5463-5470 [FREE Full text] [doi: [10.1158/0008-5472.CAN-19-0579](https://doi.org/10.1158/0008-5472.CAN-19-0579)] [Medline: [31395609](https://pubmed.ncbi.nlm.nih.gov/31395609/)]
4. Adamson B, Waskom M, Blarre A, Kelly J, Krismer K, Nemeth S, et al. Approach to machine learning for extraction of real-world data variables from electronic health records. *Front Pharmacol* 2023;14:1180962 [FREE Full text] [doi: [10.3389/fphar.2023.1180962](https://doi.org/10.3389/fphar.2023.1180962)] [Medline: [37781703](https://pubmed.ncbi.nlm.nih.gov/37781703/)]
5. Waskom ML, Tan K, Wiberg H, Cohen AB, Wittmershaus B, Shapiro W. A hybrid approach to scalable real-world data curation by machine learning and human experts. *medRxiv* 2023. [doi: [10.1101/2023.03.06.23286770](https://doi.org/10.1101/2023.03.06.23286770)]
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need? In *Advances in Neural Information Processing Systems*.: Curran Associates, Inc URL: <https://tinyurl.com/3bh6m3xw> [accessed 2024-02-22]
7. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. *arXiv:2303.08774* 2023:1-100. [doi: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774)]
8. Li L, Zhou J, Gao Z, Hua W, Fan L, Yu H, et al. A scoping review of using Large Language Models (LLMs) to investigate electronic health records (EHRs). *arXiv:2405.03066* 2024:1-45. [doi: [10.48550/arXiv.2405.03066](https://doi.org/10.48550/arXiv.2405.03066)]
9. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med* 2024 Apr 23;7(1):102 [FREE Full text] [doi: [10.1038/s41746-024-01091-y](https://doi.org/10.1038/s41746-024-01091-y)] [Medline: [38654102](https://pubmed.ncbi.nlm.nih.gov/38654102/)]
10. Landman R, Healey SP, Loprinzo V, Kochendoerfer U, Winnier AR, Henstock PV, et al. Using large language models for safety-related table summarization in clinical study reports. *JAMIA Open* 2024 Jul;7(2):ooae043 [FREE Full text] [doi: [10.1093/jamiaopen/ooae043](https://doi.org/10.1093/jamiaopen/ooae043)] [Medline: [38818116](https://pubmed.ncbi.nlm.nih.gov/38818116/)]
11. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res* 2023 Jun 28;25:e48568 [FREE Full text] [doi: [10.2196/48568](https://doi.org/10.2196/48568)] [Medline: [37379067](https://pubmed.ncbi.nlm.nih.gov/37379067/)]
12. Mumtaz U, Ahmed A, Mumtaz S. LLMs-Healthcare: current applications and challenges of large language models in various medical specialties. *AIH* 2024;1(2):16-28. [doi: [10.36922/aih.2558](https://doi.org/10.36922/aih.2558)]
13. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-Text CT reports on lung cancer. *Radiology* 2023 Sep;308(3):e231362. [doi: [10.1148/radiol.231362](https://doi.org/10.1148/radiol.231362)] [Medline: [37724963](https://pubmed.ncbi.nlm.nih.gov/37724963/)]
14. Choi HS, Song JY, Shin KH, Chang JH, Jang B. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J* 2023 Sep;41(3):209-216 [FREE Full text] [doi: [10.3857/roj.2023.00633](https://doi.org/10.3857/roj.2023.00633)] [Medline: [37793630](https://pubmed.ncbi.nlm.nih.gov/37793630/)]
15. Bertolet A, Wals A, Miras H, Macías J. Organic generation of real-world real-time data for clinical evidence in radiation oncology. *Int J Med Inform* 2020 Dec;144:104301 [FREE Full text] [doi: [10.1016/j.ijmedinf.2020.104301](https://doi.org/10.1016/j.ijmedinf.2020.104301)] [Medline: [33091831](https://pubmed.ncbi.nlm.nih.gov/33091831/)]
16. OpenAI Platform. URL: <https://platform.openai.com> [accessed 2024-02-22]
17. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S. Better to ask in english: cross-lingual evaluation of large language models for healthcare queries. *arXiv:2310.13132* 2023:1-18. [doi: [10.48550/arXiv.2310.13132](https://doi.org/10.48550/arXiv.2310.13132)]
18. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. *arXiv:2009.03300* 2021:1-27. [doi: [10.48550/arXiv.2009.03300](https://doi.org/10.48550/arXiv.2009.03300)]

19. openaiAPIscript_forsharing. GitHub. URL: https://github.com/RFMacarena/openaiAPIscript_forsharing [accessed 2024-12-23]
20. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall/CRC; 1994:456.
21. Zagher J, Naguib M, Bjelogrić M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review and recommendations for better practices. arXiv:2405.01249 2024:1-29. [doi: [10.48550/arXiv.2405.01249](https://doi.org/10.48550/arXiv.2405.01249)]
22. Li J, Chen X, Wang L, Deng XW, Wen H, You M, et al. Are You Asking GPT-4 Medical Questions Properly? - Prompt Engineering in Consistency and Reliability With Evidence-Based Guidelines for ChatGPT-4: A Pilot Study. Durham, North Carolina: Research Square Platform LLC; 2023. [doi: [10.21203/rs.3.rs-3336823/v1](https://doi.org/10.21203/rs.3.rs-3336823/v1)]
23. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT With GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res 2024 Jul 08;26:e56110 [FREE Full text] [doi: [10.2196/56110](https://doi.org/10.2196/56110)] [Medline: [38976865](https://pubmed.ncbi.nlm.nih.gov/38976865/)]
24. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis 2023 Mar;82(3):306-311 [FREE Full text] [doi: [10.1136/ard-2022-222626](https://doi.org/10.1136/ard-2022-222626)] [Medline: [36150748](https://pubmed.ncbi.nlm.nih.gov/36150748/)]

Abbreviations

API: application programming interface

GPT: generative pretrained transformers

LLM: large language model

NLP: natural language processing

RWD: real-world data

Edited by A Castonguay; submitted 16.03.24; peer-reviewed by L Guo, D Hu, S Kresevic; comments to author 26.06.24; revised version received 10.09.24; accepted 21.10.24; published 02.01.25.

Please cite as:

Wals Zurita AJ, Miras del Rio H, Ugarte Ruiz de Aguirre N, Nebrera Navarro C, Rubio Jimenez M, Muñoz Carmona D, Miguez Sanchez C

The Transformative Potential of Large Language Models in Mining Electronic Health Records Data: Content Analysis
JMIR Med Inform 2025;13:e58457

URL: <https://medinform.jmir.org/2025/1/e58457>

doi: [10.2196/58457](https://doi.org/10.2196/58457)

PMID:

©Amadeo Jesus Wals Zurita, Hector Miras del Rio, Nerea Ugarte Ruiz de Aguirre, Cristina Nebrera Navarro, Maria Rubio Jimenez, David Muñoz Carmona, Carlos Míguez Sanchez. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 02.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Development and Evaluation of a Mental Health Chatbot Using ChatGPT 4.0: Mixed Methods User Experience Study With Korean Users

Boyoung Kang¹, ME, MBA; Munpyo Hong¹, PhD

Sungkyunkwan University, Seoul, Republic of Korea

Corresponding Author:

Boyoung Kang, ME, MBA

Sungkyunkwan University

25-2, Sungkyunkwan-Ro, Jongno-gu

Seoul, 03063

Republic of Korea

Phone: 82 027401770

Email: bykang2015@gmail.com

Abstract

Background: Mental health chatbots have emerged as a promising tool for providing accessible and convenient support to individuals in need. Building on our previous research on digital interventions for loneliness and depression among Korean college students, this study addresses the limitations identified and explores more advanced artificial intelligence-driven solutions.

Objective: This study aimed to develop and evaluate the performance of HoMemeTown Dr. CareSam, an advanced cross-lingual chatbot using ChatGPT 4.0 (OpenAI) to provide seamless support in both English and Korean contexts. The chatbot was designed to address the need for more personalized and culturally sensitive mental health support identified in our previous work while providing an accessible and user-friendly interface for Korean young adults.

Methods: We conducted a mixed methods pilot study with 20 Korean young adults aged 18 to 27 (mean 23.3, SD 1.96) years. The HoMemeTown Dr CareSam chatbot was developed using the GPT application programming interface, incorporating features such as a gratitude journal and risk detection. User satisfaction and chatbot performance were evaluated using quantitative surveys and qualitative feedback, with triangulation used to ensure the validity and robustness of findings through cross-verification of data sources. Comparative analyses were conducted with other large language models chatbots and existing digital therapy tools (Woebot [Woebot Health Inc] and Happify [Twill Inc]).

Results: Users generally expressed positive views towards the chatbot, with positivity and support receiving the highest score on a 10-point scale (mean 9.0, SD 1.2), followed by empathy (mean 8.7, SD 1.6) and active listening (mean 8.0, SD 1.8). However, areas for improvement were noted in professionalism (mean 7.0, SD 2.0), complexity of content (mean 7.4, SD 2.0), and personalization (mean 7.4, SD 2.4). The chatbot demonstrated statistically significant performance differences compared with other large language models chatbots ($F=3.27$; $P=.047$), with more pronounced differences compared with Woebot and Happify ($F=12.94$; $P<.001$). Qualitative feedback highlighted the chatbot's strengths in providing empathetic responses and a user-friendly interface, while areas for improvement included response speed and the naturalness of Korean language responses.

Conclusions: The HoMemeTown Dr CareSam chatbot shows potential as a cross-lingual mental health support tool, achieving high user satisfaction and demonstrating comparative advantages over existing digital interventions. However, the study's limited sample size and short-term nature necessitate further research. Future studies should include larger-scale clinical trials, enhanced risk detection features, and integration with existing health care systems to fully realize its potential in supporting mental well-being across different linguistic and cultural contexts.

(*JMIR Med Inform* 2025;13:e63538) doi:[10.2196/63538](https://doi.org/10.2196/63538)

KEYWORDS

mental health chatbot; Dr. CareSam; HoMemeTown; ChatGPT 4.0; large language model; LLM; cross-lingual; pilot testing; cultural sensitivity; localization; Korean students

Introduction

The COVID-19 pandemic has exacerbated the already concerning rates of depression and anxiety among college students worldwide [1,2]. In Korea, the situation is particularly alarming. Recent statistics highlight the severity of mental health issues among college students in this country. According to the “2021 COVID-19 National Mental Health Survey,” individuals in their 20s showed the highest average depression score Patient Health Questionnaire-9 (PHQ-9) of 6.7 and the highest proportion of high-risk depression groups at 30% among all age groups surveyed. This represents a significant increase from 2018 when the average depression score was 2.3. The proportion of the high-risk depression group (PHQ-9 score ≥ 10) increased to 22.8% in 2021, approximately 6 times higher than in 2018.

Our previous study [3] on digital interventions for loneliness and depression among Korean college students highlighted the need for more personalized and culturally sensitive approaches, which this research aims to address. Among the initial 63 applicants in that study, the average PHQ-9 score was 9.23, with 25 (39.7%, 25/63) participants classified as high-risk for depression (PHQ-9 score ≥ 10). At the baseline of our study with 53 participants, 23 (43.4%, 23/53) were categorized as high-risk for depression. These findings underscore the urgent need for effective interventions to improve mental health among college students in Korea.

Despite the increasing severity of mental health issues among college students, the infrastructure and support systems within universities to effectively address these problems remain inadequate. In Korea, the annual mental health budget for college students is limited, with most of the resources focused on counseling services and little emphasis on preventive approaches (Counseling Council for University Students). Furthermore, due to the lack of professional counseling personnel and resources, access to services is limited, and many students are unable to receive timely and appropriate help [4,5].

Social stigma and prejudice against mental health issues also act as significant barriers for college students seeking help. Many students, despite experiencing psychological difficulties, tend to avoid help-seeking behaviors due to negative perceptions about psychiatric treatment or counseling [6]. This can exacerbate symptoms and prolong problems. In particular, those with low mental health literacy are less likely to recognize their condition or understand the need for professional intervention. In fact, the mental health literacy score of college students who participated in our previous study [3] averaged only 2.57 out of 5 points, highlighting the urgent need for educational intervention in this area.

In this context, digital technology-based mental health management solutions are gaining attention as a new alternative. Digital intervention services that overcome spatial and temporal constraints and ensure anonymity can contribute to improving accessibility and participation. Considering the high digital literacy of college students, these methods can be more familiar and acceptable to them [7]. Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have paved the way for the development of sophisticated chatbots

that can engage in human-like conversations and provide personalized support [2]. Large language models (LLMs), such as ChatGPT, have revolutionized the field of conversational AI. These models are trained on vast amounts of text data, enabling them to generate human-like responses and understand context. In mental health support, LLMs can be fine-tuned to provide empathetic responses, recognize emotional cues, and offer personalized support, making them potentially powerful tools for accessible mental health interventions.

Building upon the findings of our previous study [3], which explored the effectiveness of digital interventions for loneliness and depression among college students, this research aims to address the limitations identified in earlier digital interventions and develop a more effective and user-friendly mental health support tool. Specifically, this study focuses on the development and evaluation of an LLM-based chatbot prototype, named HoMemeTown, designed to provide personalized mental health support. The HoMemeTown chatbot, powered by ChatGPT 4.0, offers several unique features, that are (1) cross-lingual capability in English and Korean, ensuring cultural sensitivity, (2) a built-in gratitude journaling feature to promote positive thinking, (3) emotion recognition and empathetic response generation, and (4) risk detection algorithms to identify potential mental health crises.

These features aim to provide a comprehensive, user-friendly mental health support tool for young adults.

This pilot study was conducted as part of a larger research project titled “Development of a Youth Mental Health Platform Using Natural Language Processing.” While the overarching project received initial institutional review board approval, we acknowledge that this specific chatbot experiment was added later due to rapid developments in AI technology. Despite this limitation, we maintained rigorous ethical standards throughout our research, including informed consent, data privacy measures, and risk mitigation strategies.

The primary objective of this study is to conduct an initial usability test of the chatbot prototype, providing valuable insights for future, more comprehensive clinical studies. While our sample size is limited, we have used a mixed methods approach, combining quantitative usability metrics with in-depth qualitative feedback. This approach allows us to gain rich insights into user experiences and chatbot performance, even with a smaller participant pool. In addition, we have conducted comparative analyses with existing digital mental health tools to contextualize our findings within the broader landscape of mental health technologies.

As we navigate the rapidly evolving landscape of AI and its potential to revolutionize mental health support, it is crucial to explore innovative solutions that can bridge the gap between technology and human empathy [8]. This pilot study contributes to the growing body of knowledge surrounding the use of AI in mental health and sheds light on the potential of LLM-based chatbots like HoMemeTown to make a positive impact on people’s lives, while also identifying areas for future research and development.

Methods

Study Design and Participants

This pilot study used a mixed methods approach to evaluate the HoMemeTown chatbot's usability and effectiveness in providing mental health support. In total, 20 participants (12 female and 8 male) aged 18-27 (mean 23.3, SD 1.96) years were recruited through university email lists and social media advertisements. The sample size was determined to be appropriate for this pilot study, particularly given that 70% (14/20) of participants had previous experience with mental health chatbots from our previous research [3], providing valuable comparative insights. This continuity in participation enhanced our ability to gather meaningful longitudinal observations about user engagement with mental health technologies.

Participant eligibility criteria were established to ensure appropriate sampling while maintaining ethical considerations. Eligible participants included university students aged 18-27 years with Korean language proficiency and access to digital devices. We excluded individuals with severe mental health conditions requiring immediate professional intervention, as determined through initial screening questionnaires. This exclusion criterion was implemented to ensure participant safety and appropriate levels of support, following established ethical guidelines in digital mental health research [9].

Prototype Development Using the ChatGPT Application Programming Interface

The HoMemeTown chatbot is an innovative web-dependent service designed to support users in cultivating gratitude practice and providing emotional support through engaging, personalized interactions [10]. By leveraging cutting-edge NLP and emotion detection technologies, the HoMemeTown chatbot creates a unique and rewarding user experience that encourages regular engagement and promotes mental well-being [11].

The development process involved several key steps, including server setup, domain acquisition, Secure Sockets Layer certification, screen planning, development method selection, application programming interface (API) integration, front-end and back-end development, database design, performance tuning, and additional feature implementation [12].

The chatbot relies on the GPT API, a general-purpose language model provided by OpenAI, instead of a domain-specific model trained for mental health counseling. The GPT API offers a range of models, such as Davinci, GPT-3.5, and GPT-4, which can be selected based on desired performance and cost considerations. The chatbot relies on the GPT API, a general-purpose language model provided by OpenAI, instead of a domain-specific model trained for mental health counseling. The GPT API offers a range of models, such as Davinci,

GPT-3.5, and GPT-4, which can be selected based on desired performance and cost considerations.

Rate limiting and resource management are (1) maximum 4,000,000 tokens per minute processing capacity, (2) up to 5000 requests per minute, (3) implementation of a request queuing system to prevent rate limit exceedance, and (4) monthly budget monitoring and automated alerts for cost control.

These technical constraints were carefully managed to ensure consistent service delivery while maintaining cost-effectiveness. Regular monitoring of API performance and reliability metrics helped optimize the system's operation throughout the study period.

The desired functionalities of the chatbot, such as its role as a counselor and its ability to detect risks, are implemented through the use of prompts [8]. However, due to the nature of the GPT model, there is no guarantee that the chatbot will always behave exactly as intended, as its responses may vary slightly even with the same prompt [7].

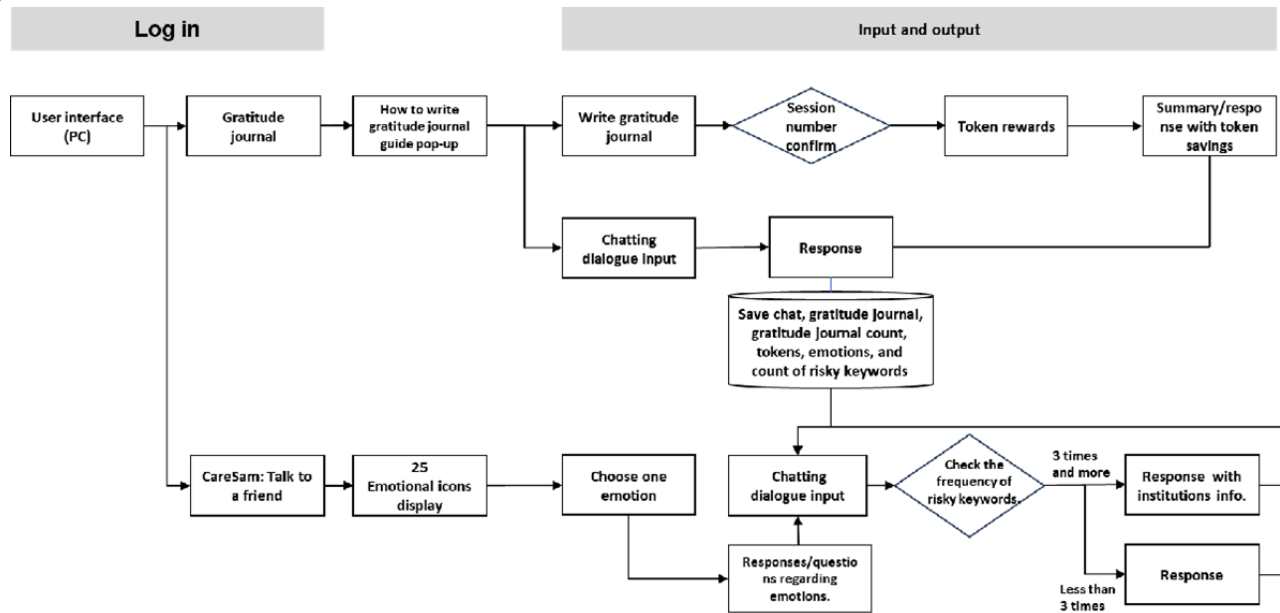
Figure 1 illustrates the service flow architecture of the HoMemeTown chatbot. The architecture depicts the user's journey, starting from the login process through the user interface on their PC. After logging in, users can access the gratitude journal section, where they can find a guide on "How to write gratitude journal" and proceed to write their own entries [13]. The system assigns a unique session number to each journaling session and securely saves the user's journal entries along with metadata such as the gratitude journal count, detected tokens or keywords, expressed emotions, and word count [14].

Users are rewarded with token rewards upon completing a journal entry, and the system generates a personalized response acknowledging their entry [4]. They can then continue their gratitude practice by initiating a new chat through the "CareSam: Talk to a friend" option. This feature allows users to select an emotion from a set of 25 emotional icons and provide more context about their feelings. Cowen and Keltner's [15] research on emotional classification inspired the inclusion of these icons. Recognizing these 25 unique emotions can help users cultivate greater self-awareness and sensitivity toward others, leading to increased empathy, connection, and understanding.

The HoMemeTown chatbot aims to encourage and motivate users to cultivate gratitude practice by providing a seamless user flow, personalized responses, and emotional attunement.

The HoMemeTown chatbot is currently accessible as an open web application. This pilot version is available for public testing, allowing anyone to interact with the chatbot and experience its features firsthand. The chatbot's continued operation demonstrates our commitment to transparency and ongoing exploration of digital mental health interventions. This open access approach enhances research reproducibility and provides opportunities for continuous feedback and improvement.

Figure 1. Service Flow Architecture of the HoMemeTown chatbot.



Technical Implementation and Server Architecture

The HoMemeTown chatbot operates on a cloud infrastructure that prioritizes privacy through a “privacy by design” approach. The key technical feature is that no personal data or chat history is stored, making the system completely stateless between sessions (Textbox 1).

This minimalist architecture was specifically chosen to eliminate privacy concerns by avoiding any form of user data collection or storage. The system operates on a request-response basis, where each interaction is treated as a new session without any historical context or user identification. This approach, while limiting some personalization features, ensures maximum privacy protection for users engaging with mental health support services [16].

Textbox 1. Server architecture.

| |
|--|
| <p>Frontend development</p> <ul style="list-style-type: none"> • React.js framework for responsive user interface • Material-UI component library for consistent design • WebSocket implementation for real-time chat functionality • Client-side-only session management with no persistent storage <p>Backend infrastructure</p> <ul style="list-style-type: none"> • Node.js runtime environment with Express.js framework • Stateless architecture with no database implementation • Direct application programming interface (API) integration with OpenAI’s GPT-4 • Nginx reverse proxy for load balancing <p>OpenAI API integration</p> <ul style="list-style-type: none"> • Implementation of GPT-4 API with custom prompt engineering • No retention of conversation history • Each interaction is processed as a new, independent session • Response token limiting for cost optimization • Regular monitoring of API performance and reliability |
|--|

User-Centered Design and Emotion Monster Selection

The chatbot targets young adults, particularly university students aged 18-27 years. It uses a set of 25 emoticons based on emotion recognition research [15]. These emoticons allow users to

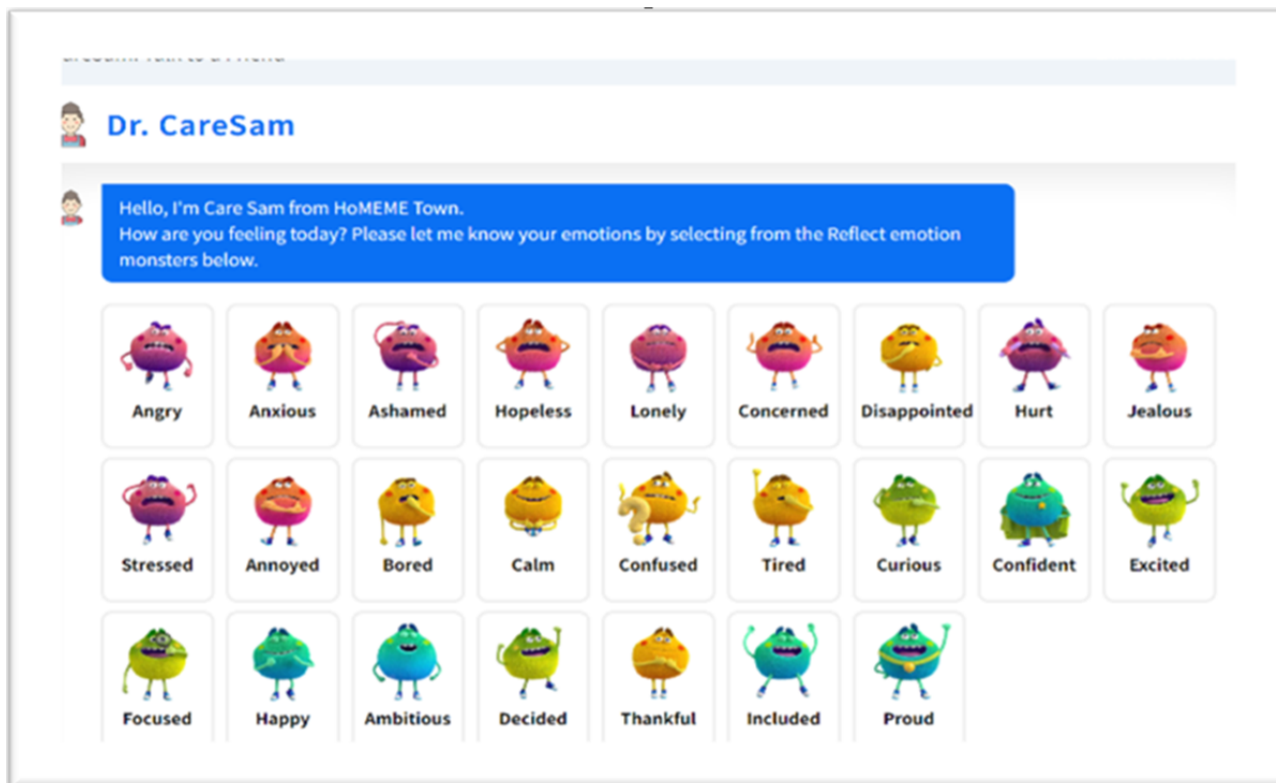
express a wide range of emotions, similar to a broader palette of colors enhancing artistic appreciation. Recognizing this wider range of emotions can lead to greater self-awareness and sensitivity toward others, ultimately fostering empathy, connection, and understanding. When a user selects an emotion

monster, a message appears on the chat screen stating their current mood. However, this message is not directly sent to the API. Instead, it is arbitrarily generated to include the selected emotion keyword and appear natural in the chat context. The actual message sent to the API is a combination of prompts, such as: “My current emotion is” + emotionText + “Acknowledge my feelings and greet me with a message of 50 characters or less.” The GPT engine receives this prompt-based

message and generates a response reacting to the user’s selected emotion.

Figure 2 displays the 25 emotional icons used in the HoMemeTown chatbot. Inspired by Microsoft’s emotion monsters, these icons represent a wide range of human emotions, allowing users to select one that reflects their current mood. This selection facilitates a more personalized and emotionally attuned response from the chatbot.

Figure 2. Emotional icons are used in the HoMemeTown chatbot.



Cross-Lingual Dialogue Development

The HoMemeTown chatbot was developed to support both English and Korean languages to cater to a wider audience. While our primary participants were Korean university students with intermediate to upper-intermediate English proficiency (based on self-reported language skills and academic records), they were able to effectively evaluate both language versions. The English version received higher ratings specifically in terms of written expression quality, though overall satisfaction levels were comparable between the 2 versions. This may be attributed to the more straightforward nature of emotional expression in English compared with the complex honorific system in Korean, as noted in our language-specific challenges.

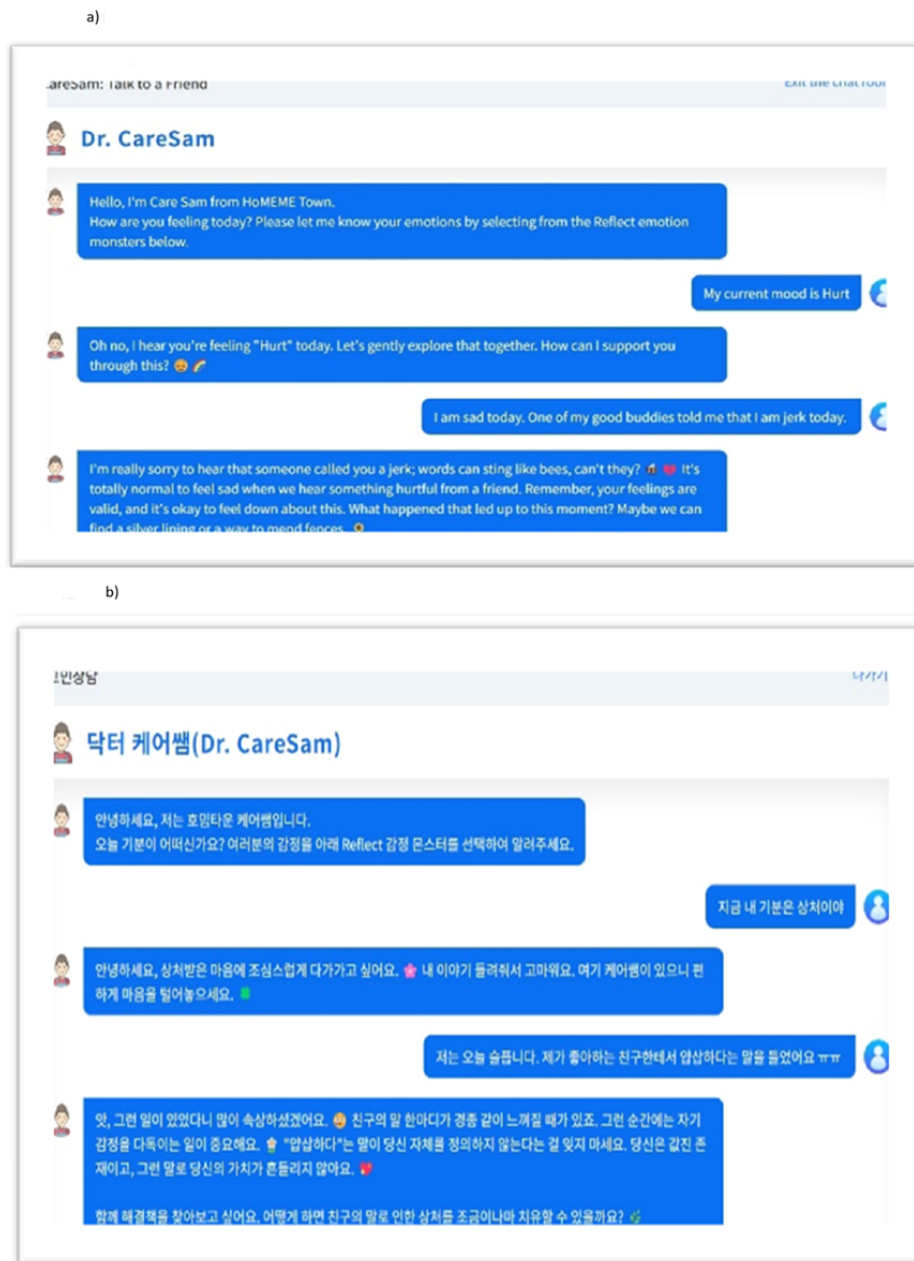
In addition to our main study, we conducted informal preliminary feedback sessions with a native English speaker (not included in the formal participant count of 20). This native speaker reported high satisfaction with the natural flow and cultural appropriateness of the English version. While this

supplementary feedback was exploratory in nature and limited in scope, it provided valuable initial insights into the cross-cultural applicability of our system.

The localization process involved adapting the chatbot’s dialogue to ensure natural conversation flow, accurate language translation, and cultural sensitivity [17]. The HoMemeTown chatbot was developed to support both English and Korean languages to cater to a wider audience. The localization process involved adapting the chatbot’s dialogue to ensure natural conversation flow, accurate language translation, and cultural sensitivity [17].

Figure 3 illustrates the dialogue flow localization process for both English and Korean versions of the HoMemeTown chatbot. This visual representation demonstrates how the chatbot’s responses are adapted to maintain natural conversation flow and cultural appropriateness in each language.

The localization process is dialogue localization, language-specific challenges, and emoji usage.

Figure 3. Dialogue flow localization. A: English version; B: Korean version.

Dialogue Localization

The chatbot's dialogue was adapted for both English and Korean languages, taking into account language structure, emoji usage, and cultural expressions. The localization process aimed to maintain the chatbot's empathetic and supportive tone while ensuring coherence and readability in each language. During the prompt tuning process, efforts were made to make the chatbot's responses more flexible and engaging, particularly in the Korean version. For example, when a user expresses feeling down and asks for something fun to do, the Korean chatbot is tuned to provide humorous and entertaining responses, similar to the English version, instead of giving a rigid, therapist-like response.

Language-Specific Challenges

The Korean localization posed unique challenges due to its agglutinative nature and honorific system. In the Korean version of the HoMEME Town chatbot, Microsoft's emotion adjectives were translated into Korean abstract nouns, such as "상처" (hurt), "외로움" (lonely), "소속감" (inclusive), "실망" (disappointed), and "지루함" (bored). To maintain a consistent rule, these abstract nouns were combined with the verb ending "이야" (이야). However, this approach led to some awkward expressions, such as "상처이야" (hurt) and "자랑이야" (proud), while others, like "실망이야" (disappointed), "외로움이야" (lonely), and "지루함이야" (bored), sounded more natural. This inconsistency in the naturalness of the expressions highlights the complexity of the Korean language and the challenges in developing a chatbot that can generate linguistically accurate and culturally appropriate responses. Although collaborating with native Korean speakers and

linguists helped ensure grammatical accuracy and appropriate honorific usage, further improvements are necessary to fully address the unique challenges posed by the Korean language, such as refining the translations and developing more sophisticated language-specific tuning techniques [17].

Emoji Usage

Emojis were strategically incorporated into both English and Korean dialogues to convey emotions and soften the tone of the conversation. The Korean dialogue used emojis more frequently to align with cultural communication preferences.

Gratitude Journal

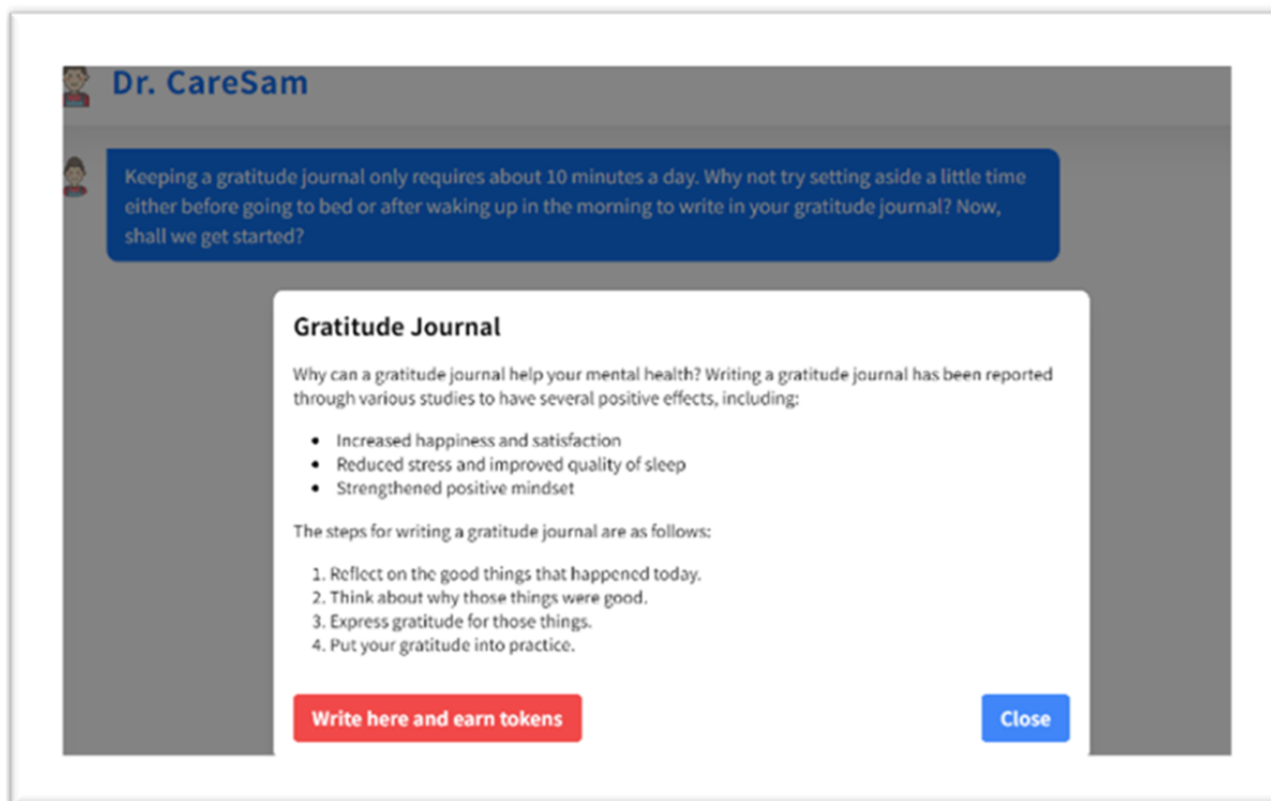
The gratitude journal feature was included in the prototype version based on positive feedback from our previous study [3]. In the previous year's experiment, many students mentioned it as one of the most satisfying and enjoyable aspects of the app. Unlike the previous year's app, where explanations of the effects and simple procedures for writing a gratitude journal were often omitted in Woebot and Happify, the user interface of the

HoMemeTown chatbot's gratitude journal includes a description of its effects and the procedure for using it.

The effects and procedures were based on established research [18], and the benefits of gratitude journaling have been demonstrated in numerous studies [13,14]. Users accumulate tokens for completed gratitude journal entries, viewable in their interface. To facilitate admin tracking of multiple user accounts, an admin page is implemented, allowing administrators to view gratitude journal entries, token balances, and chat frequencies for all user accounts.

Figure 4 illustrates the gratitude journal interface in the HoMemeTown chatbot. The interface includes a description of the benefits of gratitude journaling and a step-by-step guide on how to write a gratitude journal entry. This feature encourages users to reflect on positive experiences, express gratitude, and cultivate a more optimistic mindset. Upon completing a journal entry, users are rewarded with tokens, which are displayed in the top-right corner of the chat interface.

Figure 4. Gratitude journal interface in the HoMemeTown chatbot.



Risk Detection and Response System

The risk detection system in HoMemeTown was developed based on established clinical guidelines [19] and validated screening tools [20], implementing a sophisticated approach to identifying and responding to potential mental health concerns. The system continuously monitors user interactions for primary risk indicators, including expressions of suicidal ideation, severe depression symptoms, and anxiety crisis signals, while also tracking secondary indicators such as sleep disturbance patterns and social withdrawal signs.

When potential risks are detected, the system implements a graduated response protocol that has been carefully designed to provide appropriate levels of support while avoiding unnecessary escalation. For mild risk situations, the system offers empathetic acknowledgment and self-help resources, drawing from evidence-based interventions [21]. In cases of moderate risk, the response includes more direct expressions of concern and specific mental health resources, while severe risk triggers an immediate crisis response protocol with direct connections to professional support services.

To address the challenge of potential false positives in risk detection, we implemented a sophisticated validation system that examines multiple contextual factors before triggering interventions. This system uses NLP techniques to analyze the broader context of user communications, helping to distinguish between casual expressions and genuine indicators of distress. Regular professional review of high-risk cases ensures the ongoing refinement of detection algorithms and response protocols, maintaining a balance between sensitivity and specificity in risk assessment.

The risk detection and response system undergoes continuous evaluation and improvement based on user feedback and system performance metrics. Professional mental health experts regularly review the system's performance, leading to protocol updates that reflect emerging best practices in digital mental health support. This iterative improvement process has been shown to enhance the accuracy and effectiveness of automated mental health support systems [22].













In addition, to mitigate variability and potential errors in LLM responses, we introduced a validation process including semantic consistency checks, medical reference verification, and automatic escalation to human review when necessary, ensuring responses remain clinically appropriate and user safety is maintained.

Our implementation of these technical and clinical safeguards reflects a balanced approach to leveraging AI capabilities while maintaining high standards of user safety and support quality.

Table 1. Comparison of English and Korean chatbot versions.

| Characteristic | English version | Korean version |
|-----------------------------------|---------------------------------|--|
| Conversation style | Indirect, supportive | Direct, information-focused |
| Emotional approach | In-depth emotion exploration | Acknowledge emotions, quick transition |
| Introduction of professional help | Gradual | Immediate |
| Risk assessment method | Subtle progression | Explicitly stated |
| Cultural nuance | Emphasis on individual feelings | Focus on practical solutions |

Table 2. Example risk detection responses with cultural adaptations.

| Risk level | English response | Korean response |
|------------|---|--|
| Mild | "I notice you're having a tough day  Would you like to talk about what's bothering you?  " | "요즘 마음이 무거워 보이네요.  함께 이야기 나누면 도움이 될 수 있어요.  " |
| Moderate | "It sounds like you're going through a difficult time  Have you considered talking to someone who can help? I can suggest some resources if you'd like  " | "힘들어하시는 모습이 느껴져요.  전문가와 상담해보는 건 어떠실까요? 제가 도움이 될 만한 정보를 알려드릴 수 있어요.  " |
| Severe | "I'm very concerned about what you're sharing  There are people available 24/7 who want to support you. Would you like the contact information for immediate help?  " | "많이 걱정되는 이야기네요.  24시간 언제든지 도움을 받으실 수 있는 곳이 있습니다. 지금 바로 연락하실 수 있는 곳을 알려드릴까요?  " |

Ethical Considerations

This study was designed as an initial usability test of the HoMemeTown chatbot prototype, focusing on user experience and potential effectiveness. While it does not constitute a

The system's architecture and protocols were designed to be scalable and adaptable, allowing for continuous improvement based on ongoing research in digital mental health interventions. The HoMemeTown chatbot incorporates a risk detection function to identify potential mental health concerns through user interactions. This feature is based on the *DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition])* criteria [19] and a Korean corpus of psychopathological symptoms [20], allowing for culturally sensitive risk assessment.

The system monitors key symptoms such as depressed mood, changes in appetite or weight, sleep disturbances, fatigue, feelings of worthlessness, cognitive difficulties, and thoughts of death or suicide. When 3 or more symptoms are detected, the chatbot activates a response protocol to encourage professional help-seeking and provide relevant resources. As shown in [Table 1](#), the response strategies are carefully adapted for cultural appropriateness in both English and Korean versions, with particular attention to different cultural norms in discussing mental health concerns.

In [Table 2](#), it shows the differences that reflect the cultural variations in addressing mental health issues, highlighting the importance of culturally sensitive AI development in mental health applications. The use of emojis, while more prevalent in the Korean version, serves to soften the tone and enhance emotional expression in both languages, aligning with digital communication norms among young adults.

full-scale clinical intervention, we adhered to strict ethical guidelines for research involving human participants.

The study was conducted as part of a larger research project titled "Development of a Youth Mental Health Platform Using

Natural Language Processing.” While the overarching project received initial institutional review board approval from Sungkyunkwan University (2023-02-043-004; February 27, 2023, to June 26, 2025), we acknowledge that this specific chatbot experiment was added later due to rapid developments in AI technology. Despite this limitation, we maintained rigorous ethical standards throughout our research, including informed consent, data privacy measures, and risk mitigation strategies.

Before the usability test, all participants were provided with a comprehensive informed consent form. This form detailed the nature and purpose of the study, the procedures involved, potential risks and benefits, and the measures taken to ensure confidentiality and data protection. Participants were required to sign this form to confirm their understanding and voluntary agreement to participate.

Approximately 70% of the participants in this study had previously participated in our earlier experiment [3]. This continuity helped streamline the consent process as these participants were already familiar with the ethical standards and procedures in place for digital mental health research.

Data Security and Privacy

To address privacy concerns and protect sensitive information, the public version of the HoMemeTown chatbot operates without user registration or login requirements, collecting no personal information beyond chat interactions. This approach enhances user privacy but limits personalization features. For the usability test in this study, we implemented stringent data security measures, which are (1) all collected data (chat logs, gratitude journal entries, token rewards) were anonymized, (2) data were stored on secure, encrypted servers with restricted access, and (3) no personally identifiable information was linked to chatbot interactions or survey responses.

We incorporated a risk detection function to identify potential mental health crises and provide appropriate resources when necessary. These measures align with best practices in digital health research, ensuring ethical compliance and participant protection while advancing AI-assisted mental health support. Future developments will explore balancing personalization benefits with privacy protection, possibly through advanced encryption methods or privacy-preserving technologies.

Results

Overview

We conducted a mixed methods study to evaluate the performance and user satisfaction of our HoMemeTown Dr CareSam chatbot. The study design integrated quantitative and qualitative approaches to provide comprehensive insights: 8 quantitative questions (1 overall satisfaction item and 7

components of chatbot performance) and 4 qualitative questions (2 positive aspects and 2 areas for improvement). This mixed methods approach allowed for triangulation of data through cross-verification between quantitative metrics and qualitative user feedback, enhancing the validity and depth of our findings. The results provide multifaceted insights into the chatbot's strengths, areas for improvement, and comparative performance with other LLM and digital therapy chatbots.

Participants

The study included 20 participants aged 18 to 27 (mean 23.3, SD 1.96) years with 60% (12/20) female and 40% (8/20) male. Participants were recruited through university email lists and social media advertisements. All participants provided informed consent.

Quantitative Findings

The usability and satisfaction evaluation of the Dr CareSam counseling chatbot was conducted using a comprehensive survey consisting of 1 overall satisfaction question and 7 quantitative items assessing key components of effective psychological counseling, which consists of empathy, accuracy and usefulness, complex thinking and emotions, active listening and appropriate questions, positivity and support, professionalism, and personalization. Users generally expressed positive views, with positivity and support receiving the highest score on a 10-point scale (mean 9.0, SD 1.2), followed by empathy (mean 8.7, SD 1.6) and active listening (mean 8.0, SD 1.8). These findings align with previous research on the importance of empathy and support in mental health chatbot interactions. However, areas for improvement were noted in professionalism (mean 7.0, SD 2.0), complexity of content (mean 7.4, SD 2.0), and personalization (mean 7.4, SD 2.4), indicating potential avenues for future development to enhance user engagement and satisfaction.

Figure 5 shows the distribution of scores for various usability questions. The boxes represent the IQR, and the whiskers extend to the minimum and maximum values within 1.5 times the IQR. Black dots represent the mean values for each category.

To provide a more comprehensive understanding of the evaluation factors and their theoretical foundations, we present the following in Table 3.

This comprehensive evaluation framework enables a nuanced assessment of the chatbot's performance across key dimensions of effective counseling. The results provide valuable insights into the strengths of the Dr CareSam chatbot, particularly in areas of empathy and support. In addition, the findings highlight opportunities for improvement in professionalism, personalization, and complexity of responses, suggesting potential avenues for future development to enhance user engagement and satisfaction.

Figure 5. Box plot of scores for Dr CareSam usability questions.

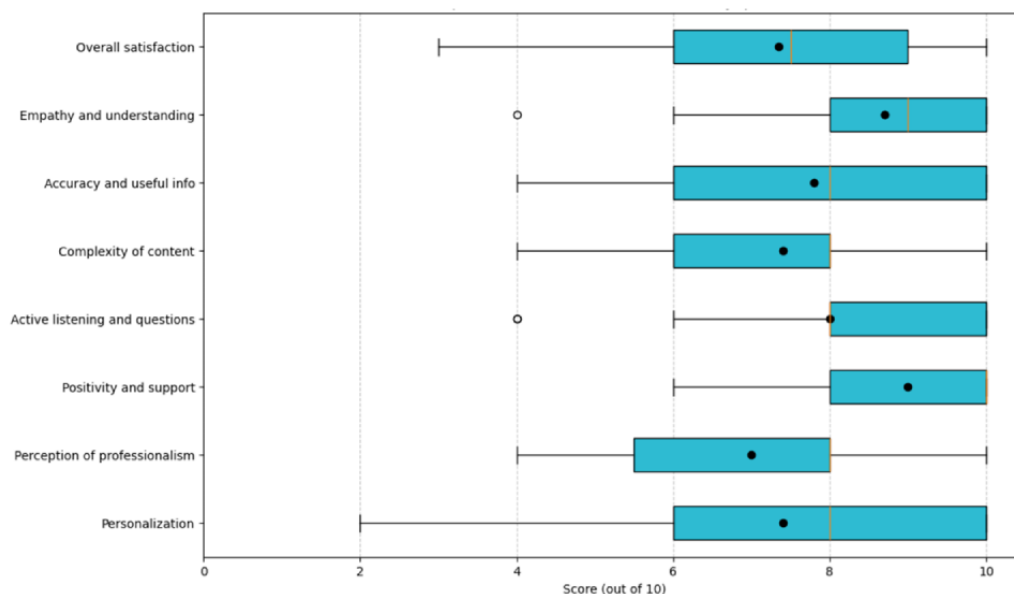


Table 3. Evaluation Factors for Dr CareSam chatbot.

| Evaluation factor | Question | Description | Previous studies |
|--|--|--|--|
| Empathy | Did CareSam’s responses express empathy and understanding of the user’s question? | An empathetic understanding of the client’s emotions and experiences is essential in effective counseling conversations. | Rogers [23], Elliott et al [24] |
| Accuracy and usefulness | Did CareSam’s responses provide accurate and useful information regarding the user’s question? | Providing accurate and actionable information aids in problem-solving and decision-making for the client. | Hepworth [25], Egan [26] |
| Complex thinking and emotions | Did CareSam’s responses include complex thinking and emotions rather than simple knowledge? | Skilled counselors should address cognitive and emotional experiences interactively, facilitating insights into the client’s internal experiences. | Greenberg [27], Gendlin [28] |
| Active listening and appropriate questions | Did CareSam’s responses include not only active listening but also appropriate questions? | Active listening and appropriate questioning techniques promote self-exploration and deeper understanding. | Weger et al [29], Hill [30] |
| Positivity and support | Did CareSam’s responses include positivity and support? | Support and encouragement enhance the client’s self-esteem and motivation for change. | Mearns and Thorne [31], Norcross and Lambert [32] |
| Professionalism | Did CareSam’s responses demonstrate professionalism? | Professionalism increases client trust and adherence to treatment, encompassing theoretical knowledge, clinical experience, and ethical awareness. | Sue and Sue [33]; Ratts et al [34] |
| Personalization | Did CareSam’s responses appear to be customized? | Effective counseling should be tailored to the individual characteristics and needs of the client. | Norcross and Wampold, [35], Beutler and Harwood [36] |

Qualitative Feedback

User feedback presented in Table 4 was collected through structured interviews and open-ended survey responses, focusing on key themes such as response speed, empathy, and personalization. Responses were categorized based on frequency of mention, providing a clear overview of commonly reported strengths and areas for improvement. User feedback confirms the chatbot’s strengths in providing detailed, empathetic responses, a user-friendly interface, and a supportive demeanor. However, areas for improvement include slow response times, Korean text flow issues, and repetitive interactions. Interestingly,

both “quick feedback” and “slow response time” were frequently mentioned, suggesting that while feedback is relevant and timely within the context of a rule-based chatbot, it may not be the fastest possible response.

The qualitative themes align with quantitative satisfaction ratings, illustrating consistent patterns across user experiences. For instance, high ratings for “positivity and support” were reflected in user comments praising Dr CareSam’s empathetic responses. Conversely, lower ratings in “personalization” corresponded with feedback indicating a desire for more tailored interactions.

Table 4. Users' experiences of good and bad points for HoMemeTown chatbots.

| Mention level and category | Positive points | Negative points |
|-----------------------------|--|---|
| Most mentioned | | |
| Response speed | Quick responses and real-time feedback. | Slow response time, lack of prompt responses. |
| Friendly and Positive Tone | Encouraging, supportive responses, provides courage and understanding. | Overemphasis on empathy, repetitive responses without practical advice. |
| Frequently mentioned | | |
| Empathy capability | Expresses understanding and empathy effectively. | Focused too heavily on empathy without specific guidance or solutions. |
| Korean language processing | Supports English and Korean; human-like, natural conversation flow. | Awkward phrasing in Korean, lack of accuracy and appropriateness in Korean responses. |
| Moderately mentioned | | |
| Detailed expression | Uses varied emotions and appropriate emoticons. | Repetitive tone, lack of personalization. |
| Chatbot functionality | Emotional expression, gratitude journaling guidance, diverse responses for different situations. | Limited content, need for features like a reward for engagement, and the option to revisit previous interactions. |
| Less mentioned | | |
| Design and interface | Clean, user-friendly interface with intuitive design elements. | Issues with long text bubbles, lack of auto line breaks, and need for design improvements. |
| Content and information | Provides relevant information and problem-solving advice. | Excessive use of emoticons, insufficient professional advice, and need for mobile accessibility improvement. |
| Least mentioned | | |
| Personalization | Positive and personalized responses, convey a warm and positive energy. | Insufficient personalization, lack of integration with medical or counseling services, missing human touch. |
| Miscellaneous | Easy usability, approachable demeanor. | Inconsistent engagement incentives need for enhanced humanistic features. |

Comparative Analysis

Performance variations among LLM chatbots, including Google Bard and the freely accessible version of ChatGPT, as illustrated in Figure 6, were statistically significant ($F=3.27$; $P=.048$). While the evaluation primarily focused on "Overall Satisfaction" [3], it is important to note that user experience differences may reflect limitations inherent to the free versions available for these comparisons, including ChatGPT 3.0 and Google Bard's publicly accessible iteration.

Satisfaction levels with HoMemeTown's Dr CareSam, compared with Woebot [5] and Happify [37], showed more pronounced

differences ($F=12.94$; $P<.001$), as shown in Figure 7, suggesting unique benefits in mental health support for college students.

In the comparison with Woebot and Happify, the previous evaluation [3] used 5 specific criteria, that are overall satisfaction, ease of use, novelty, effectiveness, and intention to maintain use. However, for this study, "Overall Satisfaction" was selected as the primary comparative metric to simplify and provide a focused assessment of user impressions. Statistical analyses were conducted using ANOVA to determine significant differences, with appropriate post hoc tests applied for multiple comparisons to identify specific group variations.

Figure 6. Comparison of large language models chatbots. LLM: Large language models.

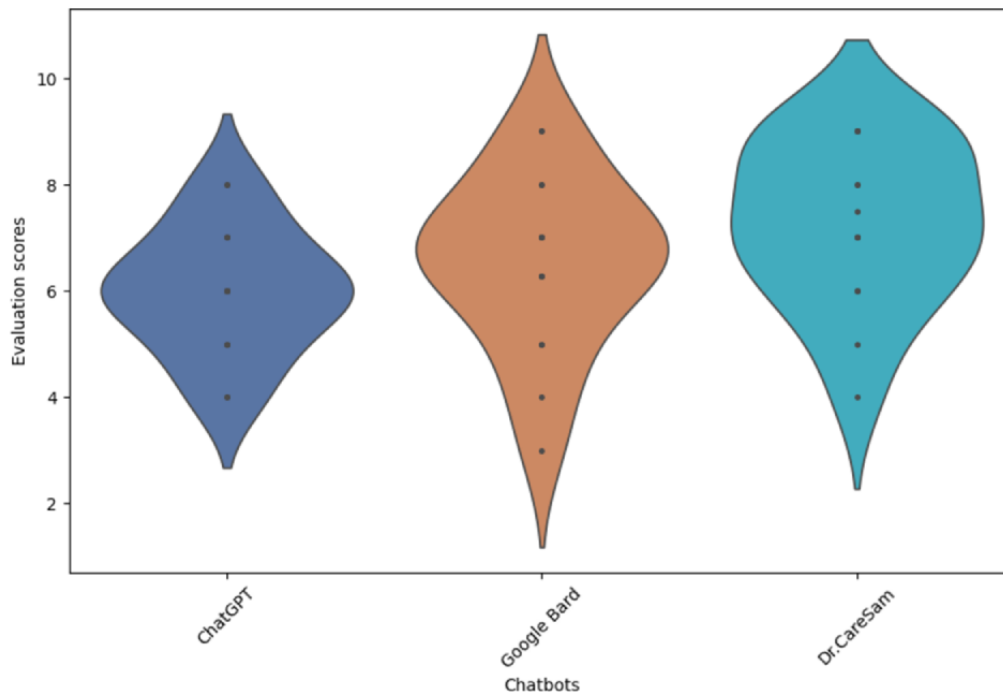
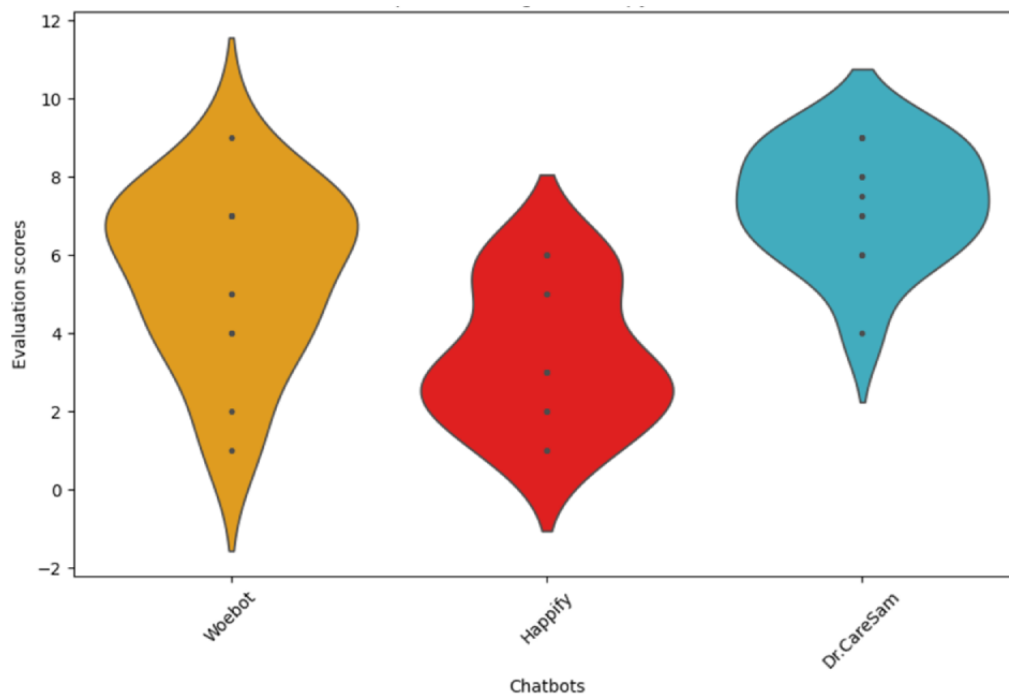


Figure 7. Comparison of digital therapy chatbots.



Discussion

Principal Findings

This study reveals several key insights about the HoMemeTown Dr CareSam chatbot, demonstrating its potential as an innovative tool in digital mental health support. The chatbot’s high performance across multiple dimensions of effective counseling, particularly in providing empathetic responses and a user-friendly interface, aligns with current research emphasizing the importance of these factors in digital mental health

interventions [3,10,11]. The statistically significant performance differences observed between HoMemeTown Dr. CareSam and other chatbots, both LLM-based ($F=3.27$; $P=.047$) and traditional digital therapy tools like Woebot and Happify ($F=12.94$; $P<.001$), suggest that our approach offers unique benefits in mental health support for college students. This could be attributed to our comprehensive evaluation framework based on established counseling principles and the integration of a sophisticated risk detection function. The chatbot’s risk detection capability, grounded in *DSM-5* criteria and a Korean corpus of

psychopathological symptoms, represents a significant advancement in AI-driven mental health support tools, enhancing its potential as a safe and responsible digital intervention.

However, the study also highlighted important challenges and areas for improvement. While the chatbot's bilingual capability is a strength, issues with unnatural expressions and response speed in the Korean version underscore the complexities of cross-cultural adaptation in AI-driven mental health tools. Furthermore, our decision to prioritize user privacy over extensive personalization features reveals a critical challenge in developing ethical AI-driven mental health interventions. This trade-off between personalization and data protection warrants further exploration in the field. Specifically, areas for improvement were noted in professionalism (mean 7.0, SD 2.0), complexity of content (mean 7.4, SD 2.0), and personalization (mean 7.4, SD 2.4), indicating potential avenues for future development to enhance user engagement and satisfaction. Collectively, these findings suggest that HoMemeTown Dr CareSam represents a promising step forward in AI-assisted mental health support, while also illuminating critical areas for future research and development in this rapidly evolving field.

Addressing LLM Variability and Technological Enhancements

To tackle the inherent challenges posed by LLM variability and potential hallucinations [38] in chatbot responses, we developed a comprehensive response validation pipeline. This pipeline includes semantic consistency checking, medical reference validation to prevent the dissemination of inaccurate information, and automatic escalation to human review when responses deviate from predetermined safety parameters. These safeguards are integral to ensuring the chatbot's interactions remain clinically appropriate, fostering user trust and alignment with established mental health support practices.

While these measures provide a critical baseline for reliability, further advancements in the underlying LLM technologies are essential for achieving higher accuracy and contextual nuance in responses. For example, the integration of LangChain [39] technology allows for the systematic management and connection of multiple language models, offering improved contextual understanding and the ability to tailor responses to specific counseling scenarios. This modular approach enhances the flexibility and precision of chatbot interactions, particularly in complex or sensitive exchanges.

In addition, leveraging retrieval-augmented generation [40] techniques further bolsters response precision by drawing upon curated counseling databases and real-world cases. This not only strengthens the relevance of responses but also minimizes the risk of erroneous or hallucinated outputs. Such enhancements highlight the evolving interplay between foundational AI capabilities and domain-specific knowledge, positioning the chatbot as a more robust and dependable digital mental health intervention.

From an ethical standpoint, these advancements underscore the importance of balancing technological innovation with user safety and data integrity. Ensuring consistent oversight, ongoing

evaluation, and refinement based on user feedback is vital to maintaining a high standard of care in digital interventions. As LLM technologies continue to evolve, our approach serves as a model for integrating emerging tools into practical applications, demonstrating how AI can be effectively harnessed to provide compassionate and reliable mental health support while continually adapting to user needs and technological developments.

Comparison With Previous Work

Unlike rule-based chatbots, HoMemeTown Dr CareSam, leveraging LLM technology, was able to provide more flexible and personalized interactions. Our findings align with previous research on Woebot's effectiveness in supporting young adults with depression and anxiety symptoms [5] while extending these benefits through our enhanced risk detection capabilities. Similarly, while Happify has shown promise in addressing loneliness during COVID-19 [37], our system demonstrates additional advantages in providing culturally adapted support for Korean users. This addresses several limitations of existing chatbots highlighted in previous studies, such as rigid response patterns and limited contextual understanding [3,5]. Specifically, our chatbot improved upon these limitations by offering more nuanced and context-appropriate responses, as evidenced by higher user satisfaction scores in empathy and active listening. The chatbot's risk detection function, grounded in *DSM-5* criteria and a Korean corpus of psychopathological symptoms, represents an advancement in AI-driven mental health support tools, offering a level of clinical relevance not typically seen in general-purpose chatbots.

Strengths and Limitations

A key strength of this study is the development of a personalized and empathetic mental health support tool using state-of-the-art LLM technology, with a sophisticated risk detection function. The bilingual support in English and Korean is another significant strength, addressing linguistic diversity and potential cross-cultural applications [17].

This study has several important limitations. The small sample size ($n=20$) limits the generalizability of the results but was chosen to ensure continuity and comparability with previous usability studies for Dr CareSam [3]. While this limited sample size may restrict broader applicability, it allowed for detailed and focused insights, particularly beneficial for pilot studies. Future research will aim to expand the sample size to validate findings and provide a more comprehensive evaluation of the chatbot's effectiveness.

In addition, feedback from a small group of native English speakers, not formally included in the 20-participant sample, revealed potential areas for improving cross-lingual functionality. Though this feedback offered valuable preliminary insights, future studies will involve broader validation with a larger group of native speakers to ensure accurate and culturally appropriate responses.

Another limitation is the potential inconsistency in comparing Dr CareSam, built on the ChatGPT 4.0 API, with user experiences based on the freely available ChatGPT 3.0. Differences in capabilities between these versions may have

influenced user perceptions and performance metrics. Future studies should strive to standardize versions for more direct and valid comparisons.

There are technical limitations associated with relying on the GPT API [7]. While this approach allows for advanced NLP capabilities, it also means that the chatbot's performance is dependent on the underlying model, which may have inherent biases or limitations. Furthermore, the reliance on an external API raises considerations about data privacy and the long-term sustainability of the system.

These limitations highlight the need for large-scale, long-term studies to fully evaluate the chatbot's effectiveness and generalizability. Future research should also explore the development of more specialized models that can be run locally, potentially addressing some of the limitations associated with relying on external APIs.

Clinical Implications

The HoMemeTown Dr CareSam chatbot shows potential as an accessible mental health support tool for young adults, with its risk detection function providing an additional layer of safety. However, it's crucial to clarify that this chatbot cannot replace professional mental health treatment, especially in cases where significant risk is detected. The chatbot's role should be seen as complementary to traditional mental health services, potentially serving as an initial point of contact or a supplementary support tool. It may be particularly useful for providing immediate support during nonclinical hours, for mild to moderate concerns, or for individuals who may be hesitant to seek traditional face-to-face therapy. However, clear guidelines must be established for when and how to transition users from the chatbot to professional human intervention.

Privacy and Personalization Considerations

A key challenge in this study was balancing personalization with user privacy. In our previous study, participants emphasized the importance of personalization features [3]. However, in developing the pilot version of HoMemeTown, we faced a significant dilemma between implementing these desired personalization features and ensuring robust privacy protection.

Ultimately, we made the decision to prioritize privacy by eliminating user registration and personal data collection in the current public version. This choice was driven by the sensitive nature of mental health data and the potential risks associated with data breaches or misuse.

As a result, the chatbot relies solely on the capabilities of the LLM to provide a sense of personalization within individual conversations, without retaining user-specific information across sessions. This approach, while enhancing data security, limits our ability to offer some of the personalized features that participants in our previous study had requested. The chatbot attempts to mimic personalization through its conversational abilities, but it cannot retain or learn from past interactions with specific users. This trade-off highlights a crucial challenge in digital mental health interventions, how to balance user expectations for personalized experiences with the ethical imperative of protecting sensitive personal information. It also underscores the need for transparent communication with users about the capabilities and limitations of AI-driven mental health tools.

Future research should explore advanced technologies like federated learning or differential privacy, which could potentially allow for more personalized features without compromising user privacy. In addition, developing clear guidelines for handling mental health data in AI-powered interventions will be essential. Our experience underscores the need for innovative solutions that balance the benefits of personalization with robust data protection in mental health contexts. As the field evolves, finding this balance will be key to developing effective, trustworthy, and ethically sound AI-powered mental health interventions [8,41].

Future Directions

Based on our findings, we propose the key areas for future research in [Textbox 2](#).

These focused directions align closely with our current work while suggesting meaningful advancements in the field of AI-assisted mental health support within the context of medical informatics.

Textbox 2. Key Points of evaluation criteria for large language model (LLM) chatbots.

Long-term effectiveness

- Conduct large-scale, longitudinal studies to evaluate the long-term impact of our cross-lingual chatbot on mental health outcomes.

Cross-cultural adaptations

- Further refine the chatbot's ability to provide culturally appropriate responses, particularly focusing on improving Korean language naturalness. Improving the naturalness of Korean language responses will involve refining language processing algorithms and collaborating with linguists and native speakers to address issues with awkward phrasing and cultural nuance. Incorporating user feedback to continuously adapt and optimize the chatbot's language output is also essential.

Speech-based user interface

- Develop and evaluate a speech-based user interface to increase usability and accessibility, particularly for users who may find voice interactions more natural or easier than text-based communication. This would involve integrating robust voice recognition and response capabilities to align with user preferences and accessibility needs.

Privacy-preserving personalization

- Explore technologies like federated learning to enhance personalization while maintaining robust data protection.

Risk detection enhancement

- Improve the accuracy and effectiveness of the risk detection function, potentially integrating it with existing mental health screening tools.

Integration with health care systems

- Investigate secure ways to integrate chatbot data with electronic health records, while maintaining user privacy.

Conclusions

This study represents a significant advancement from our previous work [3], addressing the limitations identified and exploring the potential of more sophisticated AI technologies in mental health support. By leveraging ChatGPT 4.0 and incorporating features like cross-lingual support and risk detection, we have developed a more comprehensive and adaptable tool for supporting young adults' mental health needs. This pilot study demonstrates the potential of HoMemeTown Dr CareSam, an LLM-based cross-lingual chatbot with advanced risk detection capabilities, in providing mental health support for young adults. While the chatbot showed promising results

in user satisfaction, empathetic responses, and risk assessment, challenges in professionalism, cross-lingual adaptations, and the need for technical refinements were also identified. Further long-term, large-scale studies are needed to fully evaluate its effectiveness and potential integration with existing health care systems, as we continue to refine this technology to support mental well-being in our increasingly digital world. From a medical informatics perspective, this study contributes to our understanding of how advanced AI technologies can be applied in mental health care, potentially informing the development of more sophisticated, culturally sensitive digital health tools in the future.

Acknowledgments

This research was conducted with funding from the K-medi global talent development project of the Ministry of Health and Welfare of the Republic of Korea (assignment number HI22C2185).

This research was also supported by the Sung Kyun Kwan University and the BK21 FOUR (Graduate School Innovation) funded by the Ministry of Education (South Korea) and the National Research Foundation of Korea.

The authors acknowledge the use of AI language models in the preparation of this manuscript. Claude AI [42] was used for proofreading, editing (removing redundant expressions), and formatting references. ChatGPT [43] assisted in refining the data statistical analysis code. In addition, AI tools were used to check for duplicates during the thematic analysis process. These AI tools were used to enhance the manuscript's readability and provide technical support, while the research content, analysis, and results remain entirely the work of the authors.

Authors' Contributions

BK designed the study, collected and analyzed the data, and drafted the manuscript. MH provided feedback and critically reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. *Verhaltenstherapie* 2019;29(4):266-280. [doi: [10.1159/000501812](https://doi.org/10.1159/000501812)]
2. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. *Int J Med Inform* 2019;132:103978 [FREE Full text] [doi: [10.1016/j.ijmedinf.2019.103978](https://doi.org/10.1016/j.ijmedinf.2019.103978)] [Medline: [31622850](https://pubmed.ncbi.nlm.nih.gov/31622850/)]
3. Kang B, Hong M. Digital interventions for reducing loneliness and depression in Korean college students: mixed methods evaluation. *JMIR Form Res* 2024;8:e58791 [FREE Full text] [doi: [10.2196/58791](https://doi.org/10.2196/58791)] [Medline: [39264705](https://pubmed.ncbi.nlm.nih.gov/39264705/)]
4. Lipson SK, Lattie EG, Eisenberg D. Increased rates of mental health service utilization by U.S. college students: 10-year population-level trends (2007-2017). *Psychiatr Serv* 2019;70(1):60-63 [FREE Full text] [doi: [10.1176/appi.ps.201800332](https://doi.org/10.1176/appi.ps.201800332)] [Medline: [30394183](https://pubmed.ncbi.nlm.nih.gov/30394183/)]
5. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Health* 2017;4(2):e19 [FREE Full text] [doi: [10.2196/mental.7785](https://doi.org/10.2196/mental.7785)] [Medline: [28588005](https://pubmed.ncbi.nlm.nih.gov/28588005/)]
6. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR Ment Health* 2016;3(1):e7 [FREE Full text] [doi: [10.2196/mental.4984](https://doi.org/10.2196/mental.4984)] [Medline: [26932350](https://pubmed.ncbi.nlm.nih.gov/26932350/)]
7. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *arXiv:2005.14165* 2020.
8. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The personalization of conversational agents in health care: systematic review. *J Med Internet Res* 2019;21(11):e15360 [FREE Full text] [doi: [10.2196/15360](https://doi.org/10.2196/15360)] [Medline: [31697237](https://pubmed.ncbi.nlm.nih.gov/31697237/)]
9. Andersson G, Titov N. Advantages and limitations of Internet-based interventions for common mental disorders. *World Psychiatry* 2014;13(1):4-11 [FREE Full text] [doi: [10.1002/wps.20083](https://doi.org/10.1002/wps.20083)] [Medline: [24497236](https://pubmed.ncbi.nlm.nih.gov/24497236/)]
10. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth* 2018;6(11):e12106 [FREE Full text] [doi: [10.2196/12106](https://doi.org/10.2196/12106)] [Medline: [30470676](https://pubmed.ncbi.nlm.nih.gov/30470676/)]
11. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Intern Med* 2016;176(5):619-625 [FREE Full text] [doi: [10.1001/jamainternmed.2016.0400](https://doi.org/10.1001/jamainternmed.2016.0400)] [Medline: [26974260](https://pubmed.ncbi.nlm.nih.gov/26974260/)]
12. Stuttard D, Pinto M. *The Web Application Hacker's Handbook: Finding and Exploiting Security Flaws* 2nd Edition. New Jersey, U.S: Wiley; 2011.
13. Emmons RA, McCullough ME. Counting blessings versus burdens: an experimental investigation of gratitude and subjective well-being in daily life. *J Pers Soc Psychol* 2003;84(2):377-389. [doi: [10.1037/0022-3514.84.2.377](https://doi.org/10.1037/0022-3514.84.2.377)] [Medline: [12585811](https://pubmed.ncbi.nlm.nih.gov/12585811/)]
14. Seligman MEP, Steen TA, Park N, Peterson C. Positive psychology progress: empirical validation of interventions. *Am Psychol* 2005;60(5):410-421. [doi: [10.1037/0003-066X.60.5.410](https://doi.org/10.1037/0003-066X.60.5.410)] [Medline: [16045394](https://pubmed.ncbi.nlm.nih.gov/16045394/)]
15. Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc Natl Acad Sci U S A* 2017;114(38):E7900-E7909 [FREE Full text] [doi: [10.1073/pnas.1702247114](https://doi.org/10.1073/pnas.1702247114)] [Medline: [28874542](https://pubmed.ncbi.nlm.nih.gov/28874542/)]
16. McGraw D, Dempsey JX, Harris L, Goldman J. Privacy as an enabler, not an impediment: building trust into health information exchange. *Health Aff (Millwood)* 2009;28(2):416-427. [doi: [10.1377/hlthaff.28.2.416](https://doi.org/10.1377/hlthaff.28.2.416)] [Medline: [19275998](https://pubmed.ncbi.nlm.nih.gov/19275998/)]
17. Shin Y, Kim UJ, Lee HA, Choi EJ, Park HJ, Ahn HS, Policy Development Committee of NAMOK. Health and mortality in Korean healthcare workers. *J Korean Med Sci* 2022;37(3):e22 [FREE Full text] [doi: [10.3346/jkms.2022.37.e22](https://doi.org/10.3346/jkms.2022.37.e22)] [Medline: [35040297](https://pubmed.ncbi.nlm.nih.gov/35040297/)]
18. Huberman A. Huberman Lab Podcast. URL: <https://www.hubermanlab.com/episode/the-science-of-gratitude-and-how-to-build-a-gratitude-practice> [accessed 2024-12-06]
19. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA: American Psychiatric Publishing; 2013.
20. Disease information and symptoms. National Mental Health Portal. URL: <https://www.mentalhealth.go.kr/portal/disease/diseaseList.do> [accessed 2024-04-09]
21. Karyotaki E, Furukawa TA, Efthimiou O, Riper H, Cuijpers P. Guided or self-guided internet-based cognitive-behavioural therapy (iCBT) for depression? Study protocol of an individual participant data network meta-analysis. *BMJ Open* 2019 Jun 05;9(6):e026820. [doi: [10.1136/bmjopen-2018-026820](https://doi.org/10.1136/bmjopen-2018-026820)]
22. Torous J, Wisniewski H, Bird B, Carpenter E, David G, Elejalde E, et al. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. *J. technol. behav. sci* 2019;4(7597):1-13. [doi: [10.1007/s41347-019-00095-w](https://doi.org/10.1007/s41347-019-00095-w)]

23. Rogers CR. The necessary and sufficient conditions of therapeutic personality change. *Psychotherapy (Chic)* 2007;44(3):240-248. [doi: [10.1037/0033-3204.44.3.240](https://doi.org/10.1037/0033-3204.44.3.240)] [Medline: [22122245](https://pubmed.ncbi.nlm.nih.gov/22122245/)]
24. Elliott R, Bohart AC, Watson JC, Greenberg LS. Empathy. *Psychotherapy (Chic)* 2011;48(1):43-49 [FREE Full text] [doi: [10.1037/a0022187](https://doi.org/10.1037/a0022187)] [Medline: [21401273](https://pubmed.ncbi.nlm.nih.gov/21401273/)]
25. Hepworth DH, Rooney RH, Rooney GD, Strom-Gottfried K, Larsen JA. *Direct social work practice: Theory and skills*. United States: Cengage Learning; 2017.
26. Egan G. *The skilled helper: A problem-management and opportunity-development approach to helping*. United States: Cengage Learning; 2013.
27. Greenberg LS. *Emotion-focused therapy: Coaching clients to work through their feelings*. Washington, DC: American Psychological Association; 2011.
28. Gendlin ET. *Focusing*. New York, U.S: Bantam Books; 1978.
29. Weger H, Castle Bell G, Minei EM, Robinson MC. The relative effectiveness of active listening in initial interactions. *Int. J. List* 2014;28(1):13-31. [doi: [10.1080/10904018.2013.813234](https://doi.org/10.1080/10904018.2013.813234)]
30. Hill CE. *Helping skills: Facilitating exploration, insight, and action*. Washington, DC: American Psychological Association; 2009.
31. Mearns D, Thorne B. *Person-centered counseling in action*. Washington DC: Sage; 2013.
32. Norcross JC, Lambert MJ. *Psychotherapy relationships that work: Evidence-based responsiveness*. Oxford, UK: Oxford University Press; 2018.
33. Sue DW, Sue D. *Counseling the culturally diverse: Theory and practice*. Hoboken, New Jersey: John Wiley & Sons; 2012.
34. Ratts MJ, Singh AA, Nassar - McMillan S, Butler SK, McCullough JR. Multicultural and Social Justice Counseling Competencies: Guidelines for the Counseling Profession. *J Multicult Couns & Deve* 2016;44(1):28-48. [doi: [10.1002/jmcd.12035](https://doi.org/10.1002/jmcd.12035)]
35. Norcross J, Lambert MJ. Psychotherapy relationships that work III. *Psychotherapy (Chic)* 2018;55(4):303-315. [doi: [10.1037/pst0000193](https://doi.org/10.1037/pst0000193)] [Medline: [30335448](https://pubmed.ncbi.nlm.nih.gov/30335448/)]
36. Beutler LE, Harwood TM. *Prescriptive psychotherapy: A practical guide to systematic treatment selection*. Oxford, UK: Oxford University Press; 2000.
37. Boucher EM, McNaughton EC, Harake N, Stafford JL, Parks AC. The impact of a digital intervention (Happify) on loneliness during COVID-19: qualitative focus group. *JMIR Ment Health* 2021;8(2):e26617 [FREE Full text] [doi: [10.2196/26617](https://doi.org/10.2196/26617)] [Medline: [33498011](https://pubmed.ncbi.nlm.nih.gov/33498011/)]
38. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput. Surv* 2023;55(12):1-38. [doi: [10.1145/3571730](https://doi.org/10.1145/3571730)]
39. Pandya K, Holia M. Automating customer service using LangChain: building custom open-source GPT chatbot for organizations. arXiv:2310.05421 2023.
40. Kang M, Kwak JM, Baek J, Hwang SJ. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. arXiv:2305.18846 2023.
41. Kocaballi AB, Quiroz JC, Rezazadegan D, Berkovsky S, Magrabi F, Coiera E, et al. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. *J Med Internet Res* 2020;22(2):e15823 [FREE Full text] [doi: [10.2196/15823](https://doi.org/10.2196/15823)] [Medline: [32039810](https://pubmed.ncbi.nlm.nih.gov/32039810/)]
42. Claude login. URL: <https://claude.ai/login> [accessed 2024-08-28]
43. ChatGPT. URL: <https://openai.com/chatgpt/overview/> [accessed 2024-08-28]

Abbreviations

AI: artificial intelligence

API: application programming interface

DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)

LLM: large language models

NLP: natural language processing

PHQ-9: Patient Health Questionnaire-9

Edited by C Lovis; submitted 23.06.24; peer-reviewed by S Munusamy, O Ng, DW Sung, DY Kim; comments to author 04.11.24; revised version received 12.11.24; accepted 16.11.24; published 03.01.25.

Please cite as:

Kang B, Hong M

Development and Evaluation of a Mental Health Chatbot Using ChatGPT 4.0: Mixed Methods User Experience Study With Korean Users

JMIR Med Inform 2025;13:e63538

URL: <https://medinform.jmir.org/2025/1/e63538>

doi: [10.2196/63538](https://doi.org/10.2196/63538)

PMID:

©Boyoung Kang, Munpyo Hong. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 03.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

<https://www.jmirpublications.com/>