JMIR Publications
The Leading eHealth Publisher

ChatGPT versus Human Researchers—Efficacy in Identifying Relevant Studies on m-health Interventions for Improving Medication Adherence in Ischemic Stroke Patients during Systematic Reviews: A Comparative Analysis

Suebsarn Ruksakulpiwat, Lalipat Phianhasin, Chitchanok Benjasirisan, Kedong Ding, Anuoluwapo Ajibade, Ayanesh Kumar, Cassie Stewart

Submitted to: JMIR mHealth and uHealth
on: August 17, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

ChatGPT versus Human Researchers—Efficacy in Identifying Relevant Studies on m-health Interventions for Improving Medication Adherence in Ischemic Stroke Patients during Systematic Reviews: A Comparative Analysis

Suebsarn Ruksakulpiwat[1] RN, MMed, PhD; Lalipat Phianhasin[1] RN, MS, AGPCNP-BC; Chitchanok Benjasirisan[1] RN, MS; Kedong Ding[2] AM; Anuoluwapo Ajibade[3] BS; Ayanesh Kumar[4] MS; Cassie Stewart[5]

[1]Department of Medical Nursing, Faculty of Nursing, Mahidol University Bangkok TH
[2]Jack, Joseph and Morton Mandel School of Applied Social Sciences, Case Western Reserve University Cleveland US
[3]College of Art and Science, Department of Anthropology, Case Western Reserve University Cleveland US
[4]School of Medicine, Case Western Reserve University Cleveland US
[5]Frances Payne Bolton School of Nursing, Case Western Reserve University Cleveland US

Corresponding Author:
Suebsarn Ruksakulpiwat RN, MMed, PhD
Department of Medical Nursing, Faculty of Nursing, Mahidol University
2 Wang Lang Road, Siriraj, Bangkok Noi
Bangkok
TH

Abstract

Background: ChatGPT emerged as a potential tool for researchers, aiding in various aspects of research. One such application was the identification of relevant studies in systematic reviews. However, a comprehensive comparison of the efficacy of relevant study identification between human researchers and ChatGPT has yet to be determined.

Objective: To compare the efficacy of ChatGPT and human researchers in identifying relevant studies on medication adherence improvement using m-health interventions in ischemic stroke patients during systematic reviews.

Methods: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses were used as a guideline for this study. Four electronic databases, including CINAHL Plus with Full Text, Web of Science, and PubMed/Medline, were searched to identify articles published from inception until 2023 using search terms based on Medical Subject Headings (MeSH) generated by human researchers versus ChatGPT. The authors independently screened the titles, abstracts, and full text of the studies identified through separate searches conducted by human researchers and ChatGPT. The comparison encompassed several aspects, including the ability to retrieve relevant studies, accuracy, efficiency, limitations, and challenges associated with each method.

Results: Six articles based on search terms generated by human researchers were included in the final analysis. While, ten studies were included based on search terms generated by ChatGPT, with 60% (n = 6) of them overlapping. The precision in accurately identifying relevant studies was higher in human researchers (0.86) than in ChatGPT (0.77). However, when considering the time required for both humans and ChatGPT to identify relevant studies, ChatGPT significantly outperformed human researchers as it took less time to identify relevant studies.

Conclusions: Our comparative analysis highlighted the strengths and limitations of both approaches. Ultimately, the choice between human researchers and ChatGPT depended on the specific requirements and objectives of each review, but the collaborative synergy of both approaches held the potential to advance evidence-based research and decision-making in the healthcare field. Clinical Trial: Not applicable.

(JMIR Preprints 17/08/2023:51526)
DOI: https://doi.org/10.2196/preprints.51526

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

ChatGPT versus Human Researchers—Efficacy in Identifying Relevant Studies on m-health Interventions for Improving Medication Adherence in Ischemic Stroke Patients during Systematic Reviews: A Comparative Analysis

Suebsarn Ruksakulpiwat[1*] Email: suebsarn25@gmail.com
Lalipat Phianhasin[1] Email: lalipat.phi@gmail.com
Chitchanok Benjasirisan[1] Email: chitchanok.ben@gmail.com
Kedong Ding[2] Email: kxd352@case.edu
Anuoluwapo D. Ajibade[3] Email: ada79@case.edu
Ayanesh Kumar[4] Email: axk1670@case.edu
Cassie Stewart[5] Email: cms376@case.edu

[1]Department of Medical Nursing, Faculty of Nursing, Mahidol University, Bangkok, Thailand
[2]Jack, Joseph and Morton Mandel School of Applied Social Sciences, Case Western Reserve University, Cleveland, Ohio, USA
[3]College of Art and Science, Department of Anthropology, Case Western Reserve University, Cleveland, OH, USA
[4]School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA
[5]Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, Ohio, USA

*Corresponding author: Suebsarn Ruksakulpiwat, RN, MMed, Ph.D.
Department of Medical Nursing, Faculty of Nursing, Mahidol University, Bangkok, 10700, Thailand.
Phone: (+66)984782692
Email address: suebsarn25@gmail.com

ChatGPT versus Human Researchers—Efficacy in Identifying Relevant Studies on m-health Interventions for Improving Medication Adherence in Ischemic Stroke Patients during Systematic Reviews: A Comparative Analysis

Abstract

Background: ChatGPT emerged as a potential tool for researchers, aiding in various aspects of research. One such application was the identification of relevant studies in systematic reviews. However, a comprehensive comparison of the efficacy of relevant study identification between human researchers and ChatGPT has yet to be determined.

Objective: To compare the efficacy of ChatGPT and human researchers in identifying relevant studies on medication adherence improvement using m-health interventions in ischemic stroke patients during systematic reviews.

Methods: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses were used as a guideline for this study. Four electronic databases, including CINAHL Plus with Full Text, Web of Science, and PubMed/Medline, were searched to identify articles published from inception until 2023 using search terms based on Medical Subject Headings (MeSH) generated by human researchers versus ChatGPT. The authors independently screened the titles, abstracts, and full text of the studies identified through separate searches conducted by human researchers and ChatGPT. The comparison encompassed several aspects, including the ability to retrieve relevant studies, accuracy, efficiency, limitations, and challenges associated with each method.

Results: Six articles (100%) identified through search terms generated by human researchers were included in the final analysis. Among these studies, the majority (66.7%, n = 4) reported improvements in medication adherence after the intervention. However, two of the included studies (33.3%) did not clearly state whether medication adherence improved after the intervention. Ten studies (100%) were included based on search terms generated by ChatGPT, and 60% (n = 6) of them overlapped with studies identified by human researchers. Regarding the impact of m-Health interventions on medication adherence, the majority of included studies based on search terms generated by ChatGPT (n = 8, 80%) reported improvements in medication adherence after the intervention. However, two studies (20%) did not clearly state whether medication adherence improved after the intervention. The precision in accurately identifying relevant studies was higher in human researchers (0.86) than in ChatGPT (0.77). This is consistent with the percentage of relevance, where human researchers (9.8%) demonstrated a higher percentage of relevance than ChatGPT (3%). However, when considering the time required for both humans and ChatGPT to identify relevant studies, ChatGPT significantly outperformed human researchers as it took less time to identify relevant studies.

Conclusions: Our comparative analysis highlighted the strengths and limitations of both approaches. Ultimately, the choice between human researchers and ChatGPT depended on the specific requirements and objectives of each review, but the collaborative synergy of both approaches held the potential to advance evidence-based research and decision-making in the healthcare field.

Trial Registration: Not applicable

Keywords: ChatGPT; Systematic Reviews; Medication Adherence; m-Health; Ischemic Stroke

Introduction

Artificial intelligence (AI) is the field of computer science that studies and develops systems that can perform tasks, typically requiring human intelligence, such as reasoning, learning, decision-making, natural language processing, computer vision, and speech recognition [1]. AI is a rapidly evolving field with applications in various domains, for example, healthcare, education, business, and entertainment [2]. One of the subfields of AI is natural language processing (NLP), which deals with analyzing and generating natural language texts [3]. Chatbots, a type of NLP system, can interact with humans using natural language, either through text or speech. Chatbots can be used for various purposes, including customer service, entertainment, education, and information retrieval [3]. However, developing chatbots that can engage in natural and coherent conversations with humans is a challenging task that requires advanced NLP techniques and large-scale data.

One of the recent advances in NLP is the development of generative pre-trained transformer (GPT) models, which are neural network models that can generate natural language texts based on a given input or context [4]. GPT models are trained on large corpora of text from various sources, such as books, websites, news articles, and social media posts [4]. GPT models have been used to create chatbots that can generate realistic and diverse responses to human queries or messages [4]. While ChatGPT models have been developed by various research groups and companies (i.e., OpenAI, Google, Facebook, and Microsoft), it was first introduced by OpenAI in 2019 [5]. Since then, ChatGPT has been improved and refined by researchers and developers, who have applied it to various tasks and scenarios, such as customer service, education, entertainment, and social media [5]. Chat GPT models aim to provide engaging, informative, and coherent dialogues with users across different domains and tasks [4].

ChatGPT was applied in the medical field in various ways. For instance, in medical practice, it has the ability to help streamline the clinical workflow, enhance diagnostics, and predict disease risk and outcome [6]. For medical education, ChatGPT can be useful in tailoring education and enabling powerful self-learning [6]. In terms of medical research, a previous study reported that ChatGPT has the potential to advance understanding, identify new research questions, and improve data analysis and interpretation [7]. Additionally, ChatGPT extends to involve in writing articles through improvement in language and communication of result findings [6]. In particular, in the literature review process, which is time and effort-consuming, ChatGPT has a promising advantage because of its potential ability to analyze large amounts of data, particularly in scientific articles [8]. Furthermore, ChatGPT was reported to have the potential to generate effective Boolean queries for systematic review literature searches [9].

Although ChatGPT has several advantages in medical research, it has limitations that could impact the quality of research, particularly in the literature review and search strategies processes. Citation inaccuracies, insufficient references, and references to non-existent sources were reported as current problems [6]. Moreover, ChatGPT has a limited knowledge period based on the datasets used in ChatGPT training, which limits the reliability of the updated source of the literature review [6]. For search strategies in a systematic review, a previous study recommended researchers be concerned about incorrect Medical Subject Headings (MeSH) terms and high variability in query effectiveness upon multiple requests [9]. However, ChatGPT has a high potential to be used in medical research in the future. Therefore, it is imperative to explore and develop in order to improve and use it effectively.

Despite the significant benefits and limitations of using ChatGPT, the evaluation of the quality and performance of ChatGPT models in the review process remains unclear. Therefore, this study aims to compare the efficacy of ChatGPT and human researchers in identifying relevant health-related studies, such as research on medication adherence improvement using m-health interventions in ischemic stroke patients. The review will employ systematic methods to search, select, appraise, and synthesize to address the following questions: 1) How does ChatGPT's performance compare to that of human researchers in terms of accuracy in identifying relevant studies? 2) What challenges and limitations arise from using ChatGPT versus human researchers for identifying relevant studies in systematic reviews? 3) What are the implications of utilizing ChatGPT to enhance the efficiency of systematic reviews? The results of this review will provide crucial insights into the potential of ChatGPT as an innovative tool for conducting systematic reviews.

Objective

To compare the efficacy of using ChatGPT and human researchers in identifying relevant studies on medication adherence improvement using m-health interventions in ischemic stroke patients during systematic reviews.

Methods

Identify relevant studies

In this study, we used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [10] as a guideline to identify the relevant studies. Four electronic databases, including CINAHL Plus with Full Text, Web of Science, PubMed, and Medline, were searched to identify articles published from inception until 2023 on using m-health interventions for improving medication adherence in ischemic stroke patients. We used search terms based on MeSH using

Boolean phrases generated by human researchers and ChatGPT version 3.5 to identify relevant studies. The reference lists of the included studies, generated by human researchers and ChatGPT, were separately stored and screened in EndNote. A PRISMA flow diagram was created to present the results of the search and screening process.

Study selection

The authors independently screened the titles and abstracts of the studies identified through separate searches conducted by human researchers and ChatGPT to determine their relevance. Subsequently, the full text of the selected articles was also assessed to ensure they met the predetermined inclusion criteria. A consistent set of inclusion criteria was applied to ensure that only studies relevant to the review's objective were included. In contrast, the same exclusion criteria were employed to eliminate literature unrelated to the review (Table 1).

Table 1. Study inclusion and exclusion criteria.

Inclusion	Exclusion
• Studies that aimed to utilize m-health interventions for improving medication adherence. • Studies that primarily included adults with ischemic stroke or transient ischemic attack (TIA) aged 18 years or older (the study included other stroke types, such as hemorrhagic stroke is acceptable but must include ischemic and TIA in the study population) • Studies that were described in the English language. • Studies that were published from inception until 2023.	• Studies that included children or adolescents under 18 years old. • Conference proceedings, abstracts, review articles, protocols, dissertations, letters to the editor, brief reports, or statement papers. • Studies that involved animal samples.

Data extraction

A separate summary table for data extraction will be presented, consisting of the following data for each study: Reference, Year, Country, Study Design, Sample Size, Target Population, Intervention and Objective, and Main Findings. This table will be used to compare the included studies obtained through the “Identify relevant studies” phase conducted by human researchers versus ChatGPT. The primary outcome of interest is medication adherence among patients with ischemic stroke. Medication adherence can be measured using various methods, such as drug level measurement, pill count, electronic databases, self-report questionnaires, and electronic monitoring systems [11]. The findings from studies that aimed to utilize m-Health interventions for improving medication adherence, but did not measure medical adherence directly, will be evaluated based on how they operationalized medication adherence according to their study design.

Data Analysis

Accuracy

In our study, we will assess the accuracy of both human researchers and ChatGPT in identifying relevant studies from electronic databases by measuring precision. Precision is a performance metric that measures the accuracy of a model's positive predictions. It focuses on the proportion of correctly identified positive instances (true positives) out of all the cases that the model predicted as positive (true positives + false positives) [12]. Precision is calculated using the following formula: Precision = True Positives / (True Positives + False Positives)

A high precision value close to 1 indicates that the model has a low rate of false positives. This means that when the model predicts an instance as positive, it will likely be correct. On the other hand, a low precision value close to 0 indicates that the model has a high rate of false positives. This means that when the model predicts an instance as positive, it often needs to be corrected [12]. In the context of the current study, precision will help evaluate the ability of both human researchers and ChatGPT to accurately identify relevant studies from electronic databases during the systematic review process. We will compare their precision scores to determine which approach yields a higher proportion of true positives and a lower rate of false positives.

Additionally, as the human researcher will still need to conduct screening, eligibility, and included phrases, we will also calculate the percentage of relevance using the formula ((true positives / total studies identified from the search) x 100). This approach will be chosen to ensure a fair assessment, as relying solely on a formula based on true and false positives (precision) might only reflect human variability and accuracy during the screening, eligibility, and inclusion phases.

Results

Search Term

Human Researcher

In the search phrase, we (researchers) used search terms based on MeSH using Boolean operators. The searched topic is related to using m-health interventions for improving medication adherence in ischemic stroke patients as follows: (Ischemic Stroke* OR Cryptogenic Ischemic Stroke* OR Cryptogenic Stroke* OR Cryptogenic Embolism Stroke* OR Wake up Stroke* OR Acute Ischemic Stroke* OR Embolic Stroke* OR Cardioembolic Stroke* OR Cardio-embolic Stroke* OR Thrombotic Stroke* OR Acute Thrombotic Stroke* OR Lacunar Stroke* OR Lacunar Syndrome* OR Lacunar Infarction* OR Lacunar Infarct*) AND (Medication Adherence OR Medication Nonadherence OR Medication Noncompliance OR Medication Persistence OR Medication Compliance OR Medication Non-Compliance) AND (Tele-Referral* OR Virtual Medicine OR Tele Intensive Care OR Tele ICU OR Mobile Health OR mHealth OR Telehealth OR eHealth OR Remote Consultation OR Teleconsultation* OR Telenursing OR Telepathology OR Teleradiology OR Telerehabilitation* OR Remote Rehabilitation* OR Virtual Rehabilitation*).

ChatGPT

To compare with the search by human researchers, we asked ChatGPT [13] on June 23, 2023 at 1.30 PM EST to provide a search term for conducting a systematic review of the same topic as follows, *"Hello ChatGPT, we are researchers and currently conduct a systematic review titled: Using m-health interventions for improving medication adherence in ischemic stroke patients. Can you provide Medical Subject Headings (MeSH) search terms and combine them using Boolean operators for a search process?"* The following search terms are the result from ChatGPT, which we used in the search phrase and compare the result from human researchers: (Mobile Applications OR Cell Phone OR Smartphone OR Telemedicine OR Text Messaging OR Internet) AND (Medication Adherence OR Patient Compliance OR Medication Systems, Intelligent) AND (Stroke OR Ischemic Attack, Transient OR Cerebrovascular Disorders). The search term (generated by human and ChatGPT) was adjusted according to the database searching requirement before searching but the original keyword has not change.

Search Results

We compared the ability of humans and ChatGPT to retrieve all relevant studies. A higher recall indicates a better ability to capture all the relevant literature. Figure 1 shows the flow chart diagram of included studies selection based on search terms generated by human researchers. An initial literature search yielded 61 articles, including 30 from PubMed and MEDLINE, 21 from Web of Science, and 10 from CINAHL Plus Full Text. No additional records were found through other sources. After deduplication (n = 7 studies), the researchers screened 54 references, of which 47 were excluded based on the inclusion and exclusion criteria following the title and abstract screening phase. This left 7 articles for full-text screening, during which one were excluded as it was not include any m-Health-related intervention. Therefore, 6 articles were included in the final analysis. Please note that, human researcher has conducted identification, screening, eligibility, and included phrases.

Figure 2 shows the flow chart diagram of included studies selection based on search terms generated by ChatGPT. An initial literature search yielded 334 articles, including 146 from PubMed and MEDLINE, 130 from Web of Science, and 58 from CINAHL Plus Full Text. No additional records were found through other sources. After deduplication (n = 104 studies), the researchers screened 230 references, of which 217 were excluded based on the inclusion and exclusion criteria following the title and abstract screening phase. This left 13 articles for full-text screening, during which 3 were excluded as they were irrelevant intervention (n = 1), non-English publication (n = 1), and being letter to editor (n = 1). Finally, 10 articles were included in the final analysis. Please note that, ChatGPT has been used only identification phase. The human researcher conducted screening, eligibility and included phrases.

Figure 1. The flow chart diagram displays the selection method of qualified studies searched by a human researcher.

Figure 2. The flow chart diagram displays the selection method of qualified studies searched by ChatGPT.

Description of Included Studies

Included studies from human searched

The analysis included six studies obtained from the human search (Table 2). The majority of these studies were published in 2020 (n = 3, 50%). Among the countries where the studies were conducted, three were from China (50%), while one study each (16.7%) originated from Belgium, the Republic of Korea, and Sweden. In terms of study design, the majority were cohort studies (n = 3, 50%), followed by two randomized control trials (RCTs) (33.3%) and one non-RCT (16.7%). The sample sizes varied, with three studies having a sample size ranging from 1 to 300 (50%) and the other three studies having a sample size of more than 300 (50%). Regarding the impact of m-Health interventions on medication adherence, the majority of included studies (n = 4, 66.7%) reported improvements in medication adherence after the intervention [14-17]. However, in two of the included studies (33.3%), it was not clearly stated whether medication adherence improved after the intervention [18, 19].

Included studies from ChatGPT searched

Ten studies were obtained from the ChatGPT search, out of which six studies (60%) overlapped with the human searches (Table 2). The majority of these studies were published in 2020 (n = 4, 40%). Among the countries where the studies were conducted, five were from China (50%), while one study each (10%) originated from Belgium, the Republic of Korea, Sweden, the USA, and Pakistan. In terms of study design, the majority were randomized control trials (RCTs) (n = 6, 60%), with three being cohort studies (n = 3, 30%), and one being a non-RCT (n = 1, 10%). The sample sizes varied, with seven studies having a sample size ranging from 1 to 300 (70%), and the other three studies having a sample size of more than 300 (30%). Regarding the impact of m-Health interventions on medication adherence, the majority of included studies (n = 8, 80%) reported improvements in medication adherence after the intervention [14-17, 20-23]. However, in two of the included studies (20%), it was not clearly stated whether medication adherence improved after the intervention [18, 19].

Table 2. Summary Table.

Included studies from human searched					Included studies from ChatGPT searched				
Ref	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Finding**	**Ref**	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Findi**
No 1. [18]	- Belgium - A pilot study of a prospective, multicenter, interventional cohort study	- 147 - Ischemic stroke	**Intervention:** A nurse-led self-management program (educational session during hospitalization, tips and tricks concerning a healthy lifestyle through the customized platform: websites). **Objective:** Using a personal coach and digital platform to Improve cardiovascular risk factor control in patients after ischemic stroke.	**Medication Adherence** ● Medication adherence report of 96 % **(NA)** **Other Main Findings** • Quality of life improved ($p < 0.001$). • Reduced in the 10-year risk of fatal cardiovascular disease ($p < 0.001$).	**No 1. [18]**	**Overlap with human searched**			
No 2. [14]	- Republic of Korea - Prospective, Nonrandomized, Interventional Study	- 99 (62% Ischemic stroke and 38% Hemorrhagic stroke)	**Intervention:** A 12-week Smartphone-Based Management System Intervention (regular blood pressure, blood glucose, physical activity measurements, stroke education, an exercise program, a medication program, feedback on reviewing of records by clinicians). **Objective:** To develop a smartphone-based mHealth system and to evaluate its effects on health behavior management and risk factor control in stroke patients.	**Medication Adherence** ● Medication compliance from app record was better at visit 2-3 (60.9%) than at visit 1-2 (47.8%) ($P < 0.001$) **(Y)** **Other Main Findings** • Awareness of stroke, depression, and blood pressure was enhanced when using the smartphone-based mHealth system ($p < 0.001$).	**No 2. [14]**	**Overlap with human searched**			

Included studies from human searched					Included studies from ChatGPT searched				
Ref	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Finding**	**Ref**	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Findi**
No 3. [15]	- China - A secondary data analysis from a retrospective cohort study	- 188 (65 patients paired with 123 controls) - Ischemic stroke and transient ischemic attack (TIA)	**Intervention:** Patients using mobile applications offered adherence promotion strategies **Objective:** To evaluate the effectiveness of secondary stroke prevention mobile application among stroke/TIA patients through medical adherence and stroke awareness	**Medication Adherence** ● Over 93.8% of patients in the mobile application group were adherent to their medications compared with 82.9% in the control group (p = .036) **(Y)** **Other Main Findings** ● The intervention group was more likely to be aware of stroke warning signs (p=.003) and when to seek medical attention compared to the control group (p = 0.016).	**No 3. [15]**	**Overlap with human searched**			
No 4. [19]	- Sweden - A randomized control trials	- Initial 871 (usual care = 438, intervention = 433); final 660 (intervention = 320, control = 340) for analysis - Ischemic stroke, hemorrhagic stroke, and transient ischemic attack (TIA)	**Intervention:** Nurse-led, telephone-based counseling and an assessment of pharmacological treatment (a physician was consulted to assess and adjust the medical treatment when the participants did not achieve the set target for LDL-C and/or blood pressure). **Objection:** To evaluate whether the intervention improved blood pressure values and LDL-C levels	**Medication Adherence** ● A larger proportion of the intervention group reached the treatment goal for BP **(NA)** **Other Main Findings** ● The mean systolic and diastolic BP values in the intervention group were 6.1 and 3.4 mmHg (p < 0.001) lower	**No 4. [19]**	**Overlap with human searched**			

Included studies from human searched					Included studies from ChatGPT searched				
Ref	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Finding**	**Ref**	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Findi**
			at 36-month follow-up compared to usual care, to evaluate whether a larger proportion of the intervention group reached set treatment targets, and to investigate trends in the effects of the intervention.	than in the control group ● The mean LDL-C level was 2.2 mmol/L in the intervention group, which was 0.3 mmol/L ($p < 0.001$) lower than in controls.					
No 5. [16]	- China - A community-based, two-arm cluster-randomized controlled trial with blinded outcome assessment	- 1,299 (Intervention = 637, Control = 662) - 87.1% Ischemic stroke, 12.6% Hemorrhagic stroke, 0.3% Not specified	**Intervention:** The intervention includes provider-side, supported with the SINEMA app, a smartphone application for tracking patient profiles, follow-up, visits, training, and performance indicators, and a voice message system, emphasizing medication adherence and physical activity for the patients for 12 months. **Objective:** To determine whether a primary care-based integrated mobile health intervention (SINEMA) could improve stroke management in rural China	**Medication Adherence** ● The intervention group had Improved diastolic blood pressure ($p < 0.001$), health-related quality of life ($p = 0.008$), physical activity level ($p < 0.001$), adherence to statin ($p = 0.003$), anti-hypertensive medication ($p = 0.039$), and performance in "Timed up and go" test ($p = 0.022$) **(Y)** **Other Main Findings** ● The intervention group had a -2.8 mmHg mean difference in systolic blood	**No 5. [16]**	**Overlap with human searched**			

Included studies from human searched					Included studies from ChatGPT searched				
Ref	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Finding**	**Ref**	**Country and Study Design**	**Total Sample Size, Target Population**	**Intervention and Objective**	**Main Findi**
				pressure compared to the control group (-7.1 vs. -4.3 mmHg, p = 0.005).					
No 6. [17]	- China - A cohort study	- 468 (intervention= 101, traditional = 157 for analysis) - Ischemic stroke (IS), transient ischemic attack (TIA)	**Intervention:** a physician-assisted, WeChat-based improvement service and follow-up self-monitoring platform for medication, blood glucose, and blood pressure. **Objective:** to evaluate the WeChat-based service for ischemic stroke secondary prevention designed to improve treatment adherence of discharge patients	**Medication Adherence** ● At 1-year follow-up, the intervention group showed a tendency for better compliance (3.0%) than the traditional group (7.0%) **(Y)** ● After two years, living in a community-based population was a positive predictor of adherence (OR = 2.373, p = 0.045), while having a prior TIA was a negative predictor of adherence (OR = 0.122, p = 0.04). **Other Main Findings** ● A lower rate of recurrent events (11.9%) was observed in the intervention group after one year, compared to the	**No 6. [17]**	**Overlap with human searched**			

Included studies from human searched					Included studies from ChatGPT searched				
Ref	Country and Study Design	Total Sample Size, Target Population	Intervention and Objective	Main Finding	Ref	Country and Study Design	Total Sample Size, Target Population	Intervention and Objective	Main Findi
				traditional group (13.4%).					
					No 7. [20]	- USA - A randomized, parallel-group, 12-week study	- 28 - Adults with recently diagnosed ischemic stroke receiving any anticoagulation.	**Intervention:** Patients were randomized to daily monitoring by the artificial intelligence platform (intervention) or to no daily monitoring (control). **Objective:** Evaluated the use of an artificial intelligence platform on mobile devices in measuring and increasing medication adherence in stroke patients on anticoagulation therapy.	**Medication Adherence** ● Real-time monitorin has potential increase adherence and cha behavior, particularl in patients direct anticoagul therapy **(Y** ● Mean (S cumulativ adherence based on count 97.2% (4.4%) the interventio group 90.6% (5.8%) the con group. ● Plasma d concentrat n lev indicated that adherence was 10 (15 of and 50% of 12) in interventio and con groups, respective
					No 8. [21]	- Pakistan - A parallel group, assessor-blinded, randomized, controlled, superiority trial	- 162 - Adult participants on multiple medications with access to a cell phone and stroke at least 4	**Intervention:** The intervention group in addition to usual care, received reminder SMS for 2 months that contained	**Medication Adherence** ● A sh interventio of customize SMS improve medicatio adherence

Included studies from human searched					Included studies from ChatGPT searched				
Ref	Country and Study Design	Total Sample Size, Target Population	Intervention and Objective	Main Finding	Ref	Country and Study Design	Total Sample Size, Target Population	Intervention and Objective	Main Findi
							weeks from onset (Onset as defined by last seen normal)	a) Personalized, prescription-tailored daily medication reminder(s) and b) Twice weekly health information SMS. **Objective:** designed a randomized controlled trial to test the effectiveness of SMS on improving medication adherence in stroke survivors in Pakistan.	and af stroke factors diastolic blood pressure stroke survivors with complex medicatio regimens living resource- poor ar **(Y)** • After months, mean medicatio score 7.4 (95 CI: 7.2–7 in interventio group 6.7 (95 CI: 6.4–7. in the con group.
					No 9. [22]	- China - Randomized Assessor-Blind Controlled Trial	- 174 (151 for analysis, which 75 in the control and 76 in the intervention group) - Ischemic stroke	**Intervention:** Comprehensive Reminder System based on Health Belief Model (CRS-HBM). **Objective:** To determine the impact of CRS-HBM on 5 factors (Health behavior, medication adherence, blood pressure, disability, stroke recurrence) on hypertensive patients who were discharged after a stroke, for an interval of 6 months	**Medication Adherence** • Overall, significan improvem and upward tr were no for interventio group health behaviors, medicatio adherence blood pressure, disability **Other M Findings** • No significan correlation with str recurrence was

Included studies from human searched					Included studies from ChatGPT searched				
Ref	Country and Study Design	Total Sample Size, Target Population	Intervention and Objective	Main Finding	Ref	Country and Study Design	Total Sample Size, Target Population	Intervention and Objective	Main Findi
								post-discharge.	establishe due limited incidents during trial.
					No 10. **[23]**	- China - Randomized, parallel-grouped, assessor-blinded experimental design	- 174 (158 for analysis, which 78 in the control and 80 in the intervention group) - Hospitalized hypertensive ischemic stroke patients	**Intervention:** In-person and telephone education on health beliefs and weekly automated short-message services (Data taken at baseline and three months after discharge). **Objective:** Decrease systolic blood pressure, increase blood pressure control rate, and improve health behaviors, including physical activity, low-salt diet, nutrition, and medication adherence.	**Medication Adherence** ● Improved systolic blood pressure (0.001) improved health behaviors, including physical activity, nutrition, low-salt c and medicatio adherence **(Y)** • Medicatio adherence positively correlates with implemen on of interventio **Other M Findings** • No improvem in smok and alco health behaviors

Note. “Y” = Medication adherence improved after intervention, “NA” = Medication adherence was not directly reported or did not clearly state whether it improved or not.

Accuracy

In our study, we employed precision as a metric to assess the accuracy of both human researchers and ChatGPT in identifying relevant studies from electronic databases during the systematic review process. By comparing their precision scores, we aimed to determine which approach yielded a higher proportion of true positives (correctly identified relevant studies) and a lower rate of false positives (incorrectly identified irrelevant studies). The precision calculation formula used was as follows: Precision = True Positives / (True Positives + False Positives).

Moreover, the human researcher conducted identification, screening, eligibility, and included phases, as illustrated in Figure 1. In contrast, ChatGPT was utilized only during the identification phase, and the human researcher conducted screening, eligibility, and included phrases, as depicted in Figure 2. Therefore, we also calculated the percentage of relevance using the formula ((true positives / total studies identified from the search) x 100). This approach was chosen to ensure a fair assessment, as relying solely on a formula based on true and false positives might only reflect human variability and accuracy during the screening, eligibility, and inclusion phases.

For human researchers, the precision in accurately identifying relevant studies from electronic databases was calculated as 6/(6+1) = 0.86, where 6 is included studies in the review (True Positive), 1 (False Positive) is a study that was incorrectly identified as relevant for inclusion in the review (not include any m-Health related intervention) (Figure 1). This means that out of the studies deemed relevant by human researchers, 86% were indeed appropriate for inclusion in the review, while 14% were falsely identified as relevant. The percentage of relevance for the human researcher was calculated as follows: (true positives / total studies identified from the search) x 100 = (6/61) x 100 = 9.8%.

Regarding ChatGPT, its precision in accurately identifying relevant studies from electronic databases was calculated as 10/(10+3) = 0.77, where 10 is included studies in the review (True Positive), 3 (False Positive) is a study that was incorrectly identified as relevant for inclusion in the review (irrelevant intervention, non-English publication, and is a letter to the editor) (Figure 2). This indicates that out of the studies identified by ChatGPT as potentially relevant, 77% were indeed relevant and suitable for inclusion in the review, while 23% were mistakenly identified as relevant. The percentage of relevance for ChatGPT was calculated as follows: (true positives / total studies identified from the search) x 100 = (10/334) x 100 = 3%.

According to our findings, the precision of human researchers was higher (Precision = 0.86) compared to ChatGPT (Precision = 0.77). This is consistent with the percentage of relevance, where human researchers (9.8%) demonstrated a higher percentage of relevance than ChatGPT (3%). These results indicate that human researchers were more effective in identifying relevant studies during the systematic review process. However, it is noteworthy that despite the lower precision and percentage of relevance, ChatGPT's initial search yielded a significantly larger number of studies (n = 334) compared to human researchers (n = 61), and ultimately resulted in more studies included in the final analysis (n = 10 for ChatGPT versus n = 6 for human researchers). This suggests that ChatGPT's performance was more efficient in terms of study retrieval and inclusion, even though there was a 60% overlap in the studies included between both approaches (Table 2).

Efficiency

As reported above in the accuracy section, human researchers demonstrated higher precision in identifying relevant studies compared to ChatGPT. However, the efficiency and ability of ChatGPT to retrieve relevant studies could still hold value in the systematic review process. When considering the time required for both humans and ChatGPT to identify relevant studies, from the beginning (search term generation) to the outcome (identification of relevant studies before screening), our study found that ChatGPT significantly outperformed human researchers. ChatGPT took approximately 10 minutes, whereas human researchers spent an hour in the search term identification process using MeSH and Boolean operators before obtaining the relevant study.

In our study, we employed ChatGPT to generate search terms for conducting the systematic review based on our research topic. This significantly reduced the time and effort required for initial study identification. However, it's important to note that ChatGPT's current capabilities are limited to providing search terms, and human researchers are still needed to carry out the titles, abstracts, and full-text screening of the identified studies with refined inclusion and exclusion criteria.

Discussion

According to our findings, the precision of human researchers was higher compared to ChatGPT, indicating that human researchers were more accurate in identifying relevant studies during the systematic review process. Our findings are congruent with a previous study, which reports inaccuracies of using ChatGPT in research that require an in-depth understanding of the literature [24]. Likewise, Zhao et al. (2023) report that the factual accuracy of the ChatGPT cannot be ensured, even though it has massive resources like Microsoft and Google [25]. Additionally, a case study of using ChatGPT to conduct literature searches indicates that ChatGPT does not provide an answer to the queries that researchers ask for [26].

Despite the lower precision of ChatGPT compared to human search, a previous study reports that ChatGPT has more accurate and comprehensive relevance judgments than all other types of NLP [27]. Moreover, our findings show that ChatGPT's initial search yielded a significantly larger number of studies compared to human researchers and ultimately resulted in more studies being included in the final analysis despite its lower precision. This suggests that ChatGPT's performance was more efficient in terms of study retrieval and inclusion, even though there was a 60% overlap in the studies included between both approaches. Similarly, a study of a ChatGPT about the future of scientific publishing reports it as a useful tool to get started [28]. However, a previous study using ChatGPT for retrieval of clinical, radiological information reports that ChatGPT provided only two-thirds of correct responses to questions [29].

Regarding the efficiency issues of using ChatGPT in identifying relevant search terms, the results of this study suggest that ChatGPT can be a useful tool for generating search terms for systematic reviews, as it can save time and effort for human researchers and potentially retrieve more relevant studies. The previous study on the use of ChatGPT Boolean query construction and refinement for systematic review showed that ChatGPT can generate queries with high precision [9]. Therefore, ChatGPT could be a valuable tool, especially for rapid reviews where time is limited and high precision is preferred over high recall [9].

Some may argue that as ChatGPT has lower precision and may generate irrelevant or inaccurate terms, human researchers still need to carefully screen the studies that ChatGPT identified and verify the evidence's quality and validity [30]. Chat GPT should be used with caution and verification and supplemented with other methods and sources to ensure the validity and rigor of the literature search [9]. Furthermore, ChatGPT's performance may vary depending on the research topic, data availability, and input quality. Thus, future studies are needed to evaluate ChatGPT's generalizability and reliability across different domains and contexts.

As mentioned above, using ChatGPT to generate search terms for systematic reviews raises some ethical questions regarding the quality and validity of the research process. While ChatGPT may offer some advantages in terms of efficiency and comprehensiveness, it may also introduce some biases and errors that could affect the reliability and reproducibility of the systematic reviews. For example, ChatGPT may generate search terms that are irrelevant to the research topic or too broad or narrow, resulting in either missing or including studies that do not meet the inclusion criteria [31]. Moreover, ChatGPT may generate search terms that are based on its own internal knowledge and information, which may not reflect the current state of the art or the best available evidence in the field [31]. Therefore, human researchers need to carefully evaluate and validate the search terms generated by ChatGPT and document their rationale and methods for using them. Additionally, human researchers need to disclose the use of ChatGPT as a tool for generating search terms and report its strengths and limitations and any potential ethical implications in their systematic review reports [31]. This would ensure that the systematic review process is transparent, accountable, and trustworthy and that the results are credible and useful for informing decision-making.

As we embark on a comparative analysis between ChatGPT and human researchers in the pursuit of identifying relevant studies within systematic reviews, particularly focused on m-health interventions for improving medication adherence in ischemic stroke patients, it becomes evident that several challenges and limitations underscore the intricate nature of this exploration. These challenges offer insight into the complex interplay between cutting-edge technology and the established domain expertise of human researchers, shaping the landscape in which this study unfolds.

First and foremost, the outcomes of our study are intrinsically linked to the performance of ChatGPT, an AI-driven tool that relies on its current capabilities to generate search terms. As an entity in constant evolution, ChatGPT's performance may undergo shifts over time, potentially influencing the accuracy and efficiency with which it generates relevant search terms. Moreover, replicating the search in subsequent studies is essential due to ChatGPT's intrinsic unpredictability. The lack of such repetition presents challenges in determining whether the observed phenomenon reflects an inherent trait of the model or is simply a random incident.

This dynamic underscores the need to interpret our findings in the context of the tool's state during the study period. Within the realm of medical research, the intricate and evolving nature of terminology poses a formidable challenge. While ChatGPT exhibits language generation prowess, the intricate nuances of medical terminology—constantly adapting and expanding—could potentially pose challenges to its accurate formulation of search terms. The complexity inherent to medical concepts demands a level of contextual understanding that might be challenging for an AI system.

Another pivotal consideration revolves around the potential biases embedded within ChatGPT's training data. Drawing insights from vast datasets, ChatGPT's generated search terms might inadvertently inherit biases present in the underlying data sources. This potential bias, albeit unintentional, introduces an element of caution when relying solely on AI-generated search terms for systematic reviews. A crucial aspect of our study's execution pertains to refining search terms. While ChatGPT serves as a catalyst for initial search term generation, human researchers play a pivotal role in the subsequent validation and fine-tuning of these terms. This collaborative process introduces an additional layer of complexity, as human intervention becomes essential to ensure the relevance and accuracy of the generated search terms. Moreover, the resources available and the access to ChatGPT's capabilities could introduce variability in the study's outcomes. Depending on factors such as subscription tiers or institutional resources, the extent of ChatGPT's contributions and, subsequently, its comparative assessment against human researchers may exhibit nuances that warrant consideration.

The study's defined scope, focused on m-health interventions for medication adherence improvement in ischemic stroke patients, provides a specific lens through which insights are garnered. However, this specificity inherently limits the direct transposability of findings to other medical domains or broader systematic review topics. The nuances of different research contexts might yield distinct results. Language and geographic considerations further amplify the complexity. The study predominantly engaged with English-language studies, potentially omitting valuable research published in other languages or regions. This limitation underscores the need for meticulous attention to language diversity and inclusion in systematic reviews.

Human researcher variability introduces a layer of subjectivity into the study. With multiple researchers contributing to search term generation, variations in expertise and individual approaches could impact the study's outcomes. The potential for differing interpretations and formulations of search terms necessitates careful management. Publication bias, a well-known challenge in research, extends its influence into our study's design. Both ChatGPT and human researchers might inadvertently be swayed by publication bias, where certain types of studies are more likely to be published, potentially influencing the pool of studies considered in this review.

External factors beyond the purview of our study could exert unanticipated influence. Variables such as changes in database availability, updates to search algorithms, or shifts in the research landscape might subtly shape the study's design and outcomes, introducing an element of unpredictability. The study's designated timeframe for data collection and inclusion introduces potential time constraints and selection bias. Studies published after the search period might be inadvertently omitted, potentially impacting the completeness of the review. While the study provides valuable insights within its specific scope, the generalizability of findings to other systematic review topics or research questions requires cautious interpretation. The intricate interplay between technology and human expertise forms the cornerstone of our study, emphasizing the necessity for a balanced and nuanced approach when leveraging ChatGPT for systematic reviews.

The Implications of Using ChatGPT to Improve the Efficiency of Systematic Reviews

The integration of ChatGPT into the systematic review process for identifying relevant studies on m-health interventions holds several noteworthy implications for research methodology, efficiency, and the advancement of evidence-based practices. This section explores the key implications that arise from incorporating ChatGPT as a tool to expedite and enhance the systematic review process.

One of the most immediate and impactful implications of utilizing ChatGPT is its ability to significantly expedite the systematic review process. Traditionally, the generation of search terms for identifying relevant articles is a time-intensive task that requires meticulous crafting and refinement by human researchers. ChatGPT's capacity to swiftly generate search terms offers an innovative solution to this bottleneck, reducing the time invested in this preliminary phase. This acceleration holds the potential to expedite the overall timeline of systematic reviews, enabling researchers to allocate more time to critical appraisal, synthesis, and analysis of selected studies.

The inherent nature of ChatGPT's language generation capabilities allows for a more diverse and expansive range of search terms. By tapping into its capacity to comprehend and generate natural language, researchers can explore a broader spectrum of keyword variations and synonyms. This expanded search scope can lead to the inclusion of studies that might have been overlooked using traditional search methods. As a result, the systematic review process becomes more comprehensive, encompassing a wider array of relevant literature.

ChatGPT's ability to generate novel and contextually relevant search terms introduces a valuable avenue for exploratory research and hypothesis generation. Researchers can leverage ChatGPT to identify emerging trends, novel terminologies, or unconventional associations that may inform the direction of their systematic reviews. This capacity to extract insights from the vast expanse of existing literature can potentially lead to the formulation of innovative research questions and avenues for investigation.

While ChatGPT demonstrates remarkable efficiency in generating search terms, its use necessitates a collaborative approach with human researchers. The synergy between ChatGPT's speed and human researchers' expertise in refining and validating search terms ensures a balanced and accurate outcome. Human researchers play a pivotal role in critically evaluating the generated search terms, refining them to align with the specific objectives of the review, and subsequently verifying the relevance of the identified articles. This collaborative interplay mitigates the risk of introducing erroneous or irrelevant studies into the review process.

In research environments with limited resources, such as time and personnel, ChatGPT offers a solution to address scalability challenges. Its ability to rapidly generate search terms can prove invaluable in scenarios where timely completion of systematic reviews is imperative. Researchers operating within resource-constrained contexts can leverage ChatGPT to conduct preliminary searches efficiently, thus optimizing the allocation of limited resources to subsequent stages of the review.

In summary, the integration of ChatGPT into the systematic review process introduces a transformative approach to enhancing efficiency and enriching the scope of literature exploration. While its speed and breadth of search terms hold the promise of expediting the review timeline and uncovering hidden associations, the collaborative involvement of human researchers remains pivotal for ensuring accuracy, relevance, and the meticulous execution of subsequent review stages. The strategic utilization of ChatGPT in conjunction with traditional research practices paves the way for a new era of evidence synthesis and knowledge advancement in the field of healthcare interventions.

Conclusions

Our study compares the accuracy and efficacy of human researchers and ChatGPT in providing search terms to identify articles during a systematic review on m-health interventions for improving medication adherence in ischemic stroke patients. While human researchers achieved greater precision, ChatGPT's search results exhibited lower accuracy. However, ChatGPT excelled in efficacy, taking less time to generate search terms compared to human researchers, who required more time to identify appropriate search terms. ChatGPT's search also yielded a higher number of articles compared to human researchers. Following exclusions, human researchers were left with six articles, and ChatGPT resulted in ten articles after screening, six of which overlapped with our human researchers' findings. The use of ChatGPT in creating search terms can significantly accelerate the systematic review process, though human researchers are still essential to carry out the selection process and ensure accuracy.

Acknowledgments

No funding was received for this study. We would like to express our gratitude to Mahidol University, Thailand, for covering the Article Publishing Charge.

Ethical Considerations

This study considers non-human research according to the "Self-Assessment form whether an activity is human subject research which requires ethical approval" recommended by Mahidol University Central Institutional Review Board (MU-CIRB). Therefore, ethical approval from the research ethics committee is not required.

Conflicts of Interest

There are no conflicts of interest to declare.

Abbreviations

AI: Artificial intelligence
NLP: Natural Language Processing
GPT: Generative Pre-trained Transformer
MeSH: Medical Subject Headings
PRISMA: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TIA: Transient Ischemic Attack
RCT: Randomized Control Trial

m-Health: Mobile Health

References

1. Chassignol M, Khoroshavin A, Klimova A, Bilyatdinova A. Artificial Intelligence trends in education: a narrative overview. Procedia Computer Science. 2018;136:16-24.
2. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in healthcare: Elsevier; 2020. p. 25-60.
3. Hapke H, Howard C, Lane H. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python: Simon and Schuster; 2019. ISBN: 1638356890.
4. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. Journal of the Association for Information Science and Technology. 2023;74(5):570-81. doi: https://doi.org/10.1002/asi.24750.
5. CNBC. Why tech insiders are so excited about ChatGPT, a chatbot that answers questions and writes essays. CNBC; 2022 [cited 2023]; Available from: https://www.cnbc.com/2022/12/13/chatgpt-is-a-new-ai-chatbot-that-can-answer-questions-and-write-essays.html.
6. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 2023;11(6):887. PMID: doi:10.3390/healthcare11060887.
7. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in Medical Research: Current Status and Future Directions. Journal of Multidisciplinary Healthcare. 2023 2023/12/31;16:1513-20. doi: 10.2147/JMDH.S413470.
8. Dahmen J, Kayaalp ME, Ollivier M, Pareek A, Hirschmann MT, Karlsson J, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surgery, Sports Traumatology, Arthroscopy. 2023 2023/04/01;31(4):1187-9. doi: 10.1007/s00167-023-07355-6.
9. Wang S, Scells H, Koopman B, Zuccon G. Can chatgpt write a good boolean query for systematic review literature search? arXiv preprint arXiv:230203495. 2023.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Reprint—preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Phys Ther. 2009;89(9):873-80.
11. Anghel LA, Farcas AM, Oprean RN. An overview of the common methods used to measure treatment adherence. Medicine and pharmacy reports. 2019;92(2):117.
12. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:201016061. 2020.
13. OpenAI. ChatGPT (3.5) [Large language model]. 2024; Available from: https://chat.openai.com.
14. Kim DY, Kwon H, Nam K-W, Lee Y, Kwon H-M, Chung YS. Remote Management of Poststroke Patients With a Smartphone-Based Management System Integrated in Clinical Care: Prospective, Nonrandomized, Interventional Study. Journal of medical

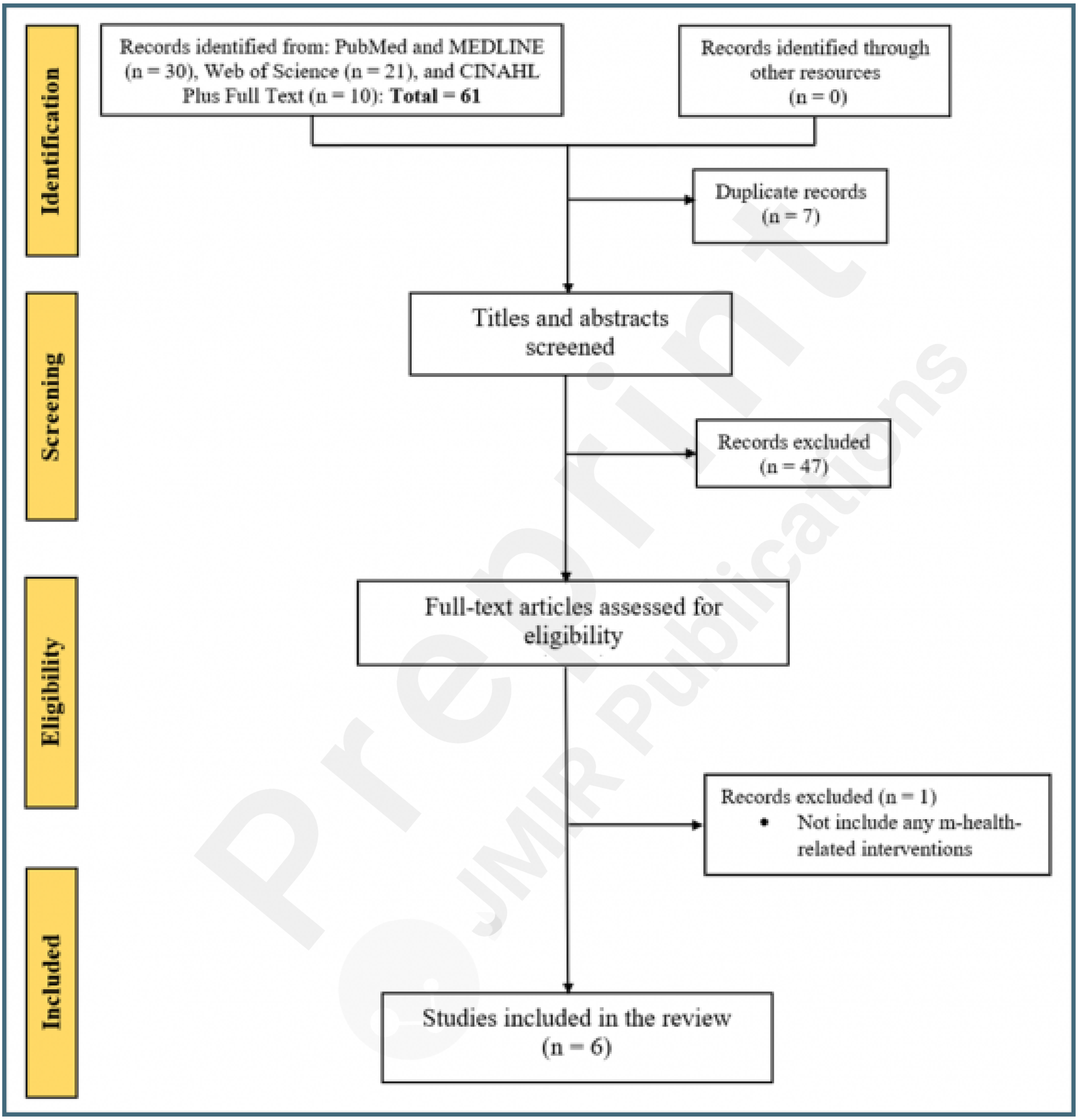
Internet research. 2020;22(2):N.PAG-N.PAG. PMID: 142128575. Language: English. Entry Date: 20200927. Revision Date: 20211220. Publication Type: journal article. doi: 10.2196/15377.
15. Li DM, Lu XY, Yang PF, Zheng J, Hu HH, Zhou Y, et al. Coordinated Patient Care via Mobile Phone-Based Telemedicine in Secondary Stroke Prevention: A Propensity Score-Matched Cohort Study. J Nurs Care Qual. 2023 Jul-Sep 01;38(3):E42-e9. PMID: 36827597. doi: 10.1097/ncq.0000000000000693.
16. Yan LL, Gong E, Gu W, Turner EL, Gallis JA, Zhou Y, et al. Effectiveness of a primary care-based integrated mobile health intervention for stroke management in rural China (SINEMA): A cluster-randomized controlled trial. PLoS medicine. 2021;18(4):1-20. PMID: 150038262. Language: English. Entry Date: In Process. Revision Date: 20211117. Publication Type: journal article. doi: 10.1371/journal.pmed.1003582.
17. Zhang Y, Fan D, Ji H, Qiao S, Li X. Treatment Adherence and Secondary Prevention of Ischemic Stroke Among Discharged Patients Using Mobile Phone- and WeChat-Based Improvement Services: Cohort Study. JMIR Mhealth Uhealth. 2020 Apr 15;8(4):e16496. PMID: 32293574. doi: 10.2196/16496.
18. Kamoen O, Maqueda V, Yperzeele L, Potter H, Cras P, Vanhooren G, et al. Stroke coach: a pilot study of a personal digital coaching program for patients after ischemic stroke. Acta Neurologica Belgica. 2020 Feb;120(1):91-7. PMID: WOS:000495056500001. doi: 10.1007/s13760-019-01218-z.
19. Ögren J, Irewall AL, Söderström L, Mooe T. Long-term, telephone-based follow-up after stroke and TIA improves risk factors: 36-month results from the randomized controlled NAILED stroke risk factor trial. BMC Neurol. 2018 Sep 21;18(1):153. PMID: 30241499. doi: 10.1186/s12883-018-1158-5.
20. Labovitz DL, Shafner L, Reyes Gil M, Virmani PD, Hanina A, Virmani D. Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy. Stroke (00392499). 2017;48(5):1416-9. PMID: 122680566. Language: English. Entry Date: 20170703. Revision Date: 20180504. Publication Type: journal article. doi: 10.1161/STROKEAHA.116.016281.
21. Kamran Kamal A, Shaikh Q, Pasha O, Azam I, Islam M, Ali Memon A, et al. A randomized controlled behavioral intervention trial to improve medication adherence in adult stroke patients with prescription tailored Short Messaging Service (SMS)-SMS4Stroke study. BMC Neurology. 2015;15(1):1-11. PMID: 110575986. Language: English. Entry Date: 20180801. Revision Date: 20180920. Publication Type: journal article. doi: 10.1186/s12883-015-0471-5.
22. Meng-Yao W, Meng-Jie S, Li-Hong W, Miao-Miao M, Zhen W, Li-Li L, et al. Effects of a Comprehensive Reminder System Based on the Health Belief Model for Patients Who Have Had a Stroke on Health Behaviors, Blood Pressure, Disability, and Recurrence From Baseline to 6 Months: A Randomized Controlled Trial. Journal of Cardiovascular Nursing. 2020;35(2):156-64. PMID: 141923688. Language: English. Entry Date: 20200228. Revision Date: 20210311. Publication Type: Article. doi: 10.1097/JCN.0000000000000631.
23. Li-Hong W, Xiao-Pei Z, Li-Ming Y, Heng-Fang R, Shao-Xian C. The Efficacy of a Comprehensive Reminder System to Improve Health Behaviors and Blood Pressure Control in Hypertensive Ischemic Stroke Patients: A Randomized Controlled Trial.

Journal of Cardiovascular Nursing. 2018;33(6):509-17. PMID: 132624191. Language: English. Entry Date: 20181031. Revision Date: 20190212. Publication Type: Article. doi: 10.1097/JCN.0000000000000496.
24. Van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023;614(7947):224-6.
25. Zhao R, Li X, Chia YK, Ding B, Bing L. Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. arXiv preprint arXiv:230411076. 2023.
26. McGee RW. Using ChatGPT to Conduct Literature Searches: A Case Study. Journal of Business Ethics. 2023;95(2):165-78.
27. Sun W, Yan L, Ma X, Ren P, Yin D, Ren Z. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. arXiv preprint arXiv:230409542. 2023.
28. Hill-Yardin EL, Hutchinson MR, Laycock R, Spencer SJ. A Chat (GPT) about the future of scientific publishing. Brain Behav Immun. 2023;110:152-4.
29. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. Canadian Association of Radiologists Journal. 2023:08465371231171125.
30. Bozkurt A, Karadeniz A, Baneres D, Guerrero-Roldán AE, Rodríguez ME. Artificial intelligence and reflections from educational landscape: A review of AI Studies in half a century. Sustainability. 2021;13(2):800.
31. Khlaif ZN. Ethical Concerns about Using AI-Generated Text in Scientific Research. Available at SSRN 4387984. 2023.

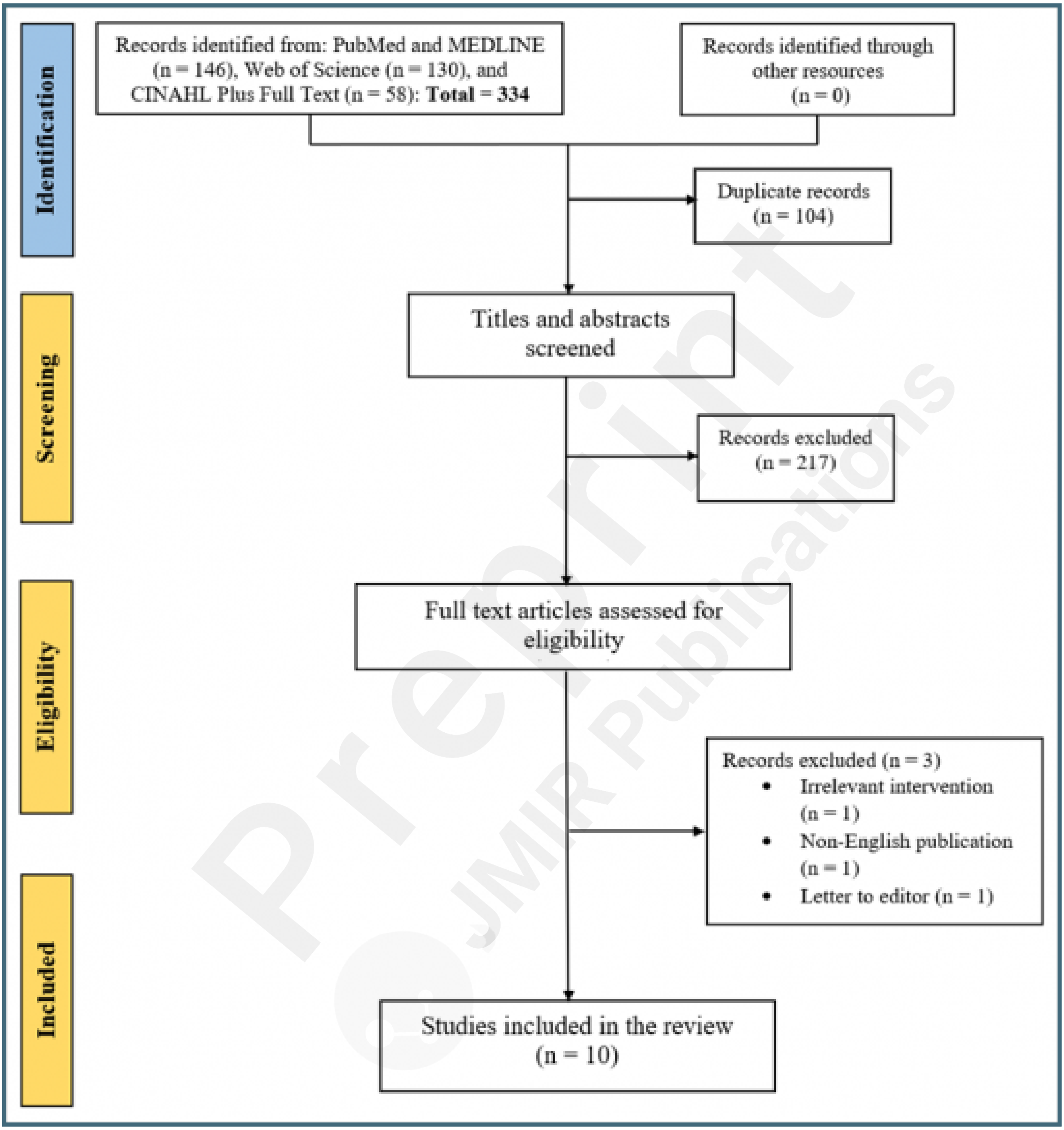
Supplementary Files

Figures

Flow chart diagram displaying the selection method of qualified studies searched by a human researcher.


Identification
Screening
Eligibility
Included
Records identified from: PubMed and MEDLINE (n = 30), Web of Science (n = 21), and CINAHL Plus Full Text (n = 10): **Total = 61**
Records identified through other resources (n = 0)
Duplicate records (n = 7)
Titles and abstracts screened
Records excluded (n = 47)
Full-text articles assessed for eligibility
Records excluded (n = 1)
• Not include any m-health-related interventions
Studies included in the review (n = 6)

Flow chart diagram displaying the selection method of qualified studies searched by ChatGPT.


Identification
Screening
Eligibility
Included
Records identified from: PubMed and MEDLINE (n = 146), Web of Science (n = 130), and CINAHL Plus Full Text (n = 58): **Total = 334**
Records identified through other resources (n = 0)
Duplicate records (n = 104)
Titles and abstracts screened
Records excluded (n = 217)
Full text articles assessed for eligibility
Records excluded (n = 3)
Irrelevant intervention (n = 1)
Non-English publication (n = 1)
Letter to editor (n = 1)
Studies included in the review (n = 10)

CONSORT (or other) checklists

PRISMA checklist.
URL: http://asset.jmir.pub/assets/7147b58d385023c6c7c6c20532e73f49.pdf