Division of Clinical Pharmacy[1], Keio University Faculty of Pharmacy; Department of Pharmacy[2], Keio University Hospital; Department of Clinical Pharmacy[3], Keio University School of Medicine, Tokyo, Japan

Comprehensive analysis of responses from ChatGPT to consumer inquiries regarding over-the-counter medications

K. Kiyomiya[1,*], T. Aomori[1,2], H. Ohtani[1,2,3]

Received September 20, 2023, accepted November 3, 2023

**Corresponding author: Keisuke Kiyomiya, Division of Clinical Pharmacy, Keio University Faculty of Pharmacy, 1-5-30 Shibakoen, Minato-ku, Tokyo 105-8512, Japan*
keisuke.kiyomiya@adst.keio.ac.jp

Pharmazie 79: 24-28 (2024) *doi: 10.1691/ph.2024.3628*

Background: The use of generative artificial intelligence (AI) applications such as ChatGPT is becoming increasingly popular. In Japan, consumers can purchase most over-the-counter (OTC) drugs without having to consult a pharmacist, so they may ask generative AI applications which OTC drugs they should purchase. This study aimed to systematically evaluate responses from ChatGPT to consumer inquiries about various OTC drugs. *Methods:* We selected 22 popular OTC drugs and 12 typical consumer characteristics, including physical and disease conditions and concomitant medications. We input a total of 264 questions (*i.e.*, all combinations of drugs and characteristics) to ChatGPT in Japanese, asking whether it is safe for consumers with each characteristic to take these OTC drugs. We used the generic name for 10 of the 22 drugs and the brand name for the remaining 12. Responses were evaluated based on the following three criteria: 1) coherence between the question and response, 2) scientific correctness, and 3) appropriateness of the instructed actions. When we received a response that satisfied all three criteria, we input the exact same question on a different day to assess reproducibility. *Results:* The proportions of ChatGPT's answers that satisfied criteria 1, 2, and 3 were 79.5%, 54.5%, and 49.6%, respectively. However, the proportion of responses that satisfied all three criteria was only 20.8% (55/264); 61.8% (34/55) of these responses were reproduced when the same question was input again on a different day. Compared with questions using generic names, those using brand names resulted in lower coherence and scientific correctness. Among the 12 characteristics, the appropriateness of the instructed actions tended to be lower in responses to questions about driving and concomitant medications. *Conclusions:* Our study revealed that ChatGPT was less accurate in its responses and less consistent in its instructed actions compared with the package inserts. Our findings suggest that Japanese consumers should not consult ChatGPT regarding OTC medications, especially when using brand names.

1. Introduction

The use of generative artificial intelligence (AI) has become increasingly popular in recent years. Among the various generative AI applications, ChatGPT (OpenAI, Inc.), which was publicly released in November 2022, has superior natural language processing capabilities, including article generation, language translation, and question answering (Gilson et al. 2023; Kunze et al. 2023; van Dis et al. 2023). There are already some reports raising concerns about the correctness and ethics of healthcare professionals using ChatGPT in clinical settings (Sallam 2023).

In contrast, no studies have investigated the use of ChatGPT by consumers to collect healthcare information. Many consumers turn to the Internet to collect medical information, and with the rapid increase in the number of smartphone users in recent years, consumers are increasingly seeking health-related information online (Ashkanani et al. 2019; Bergmo et al. 2023; Diaz et al. 2002; Hämeen-Anttila et al. 2018; Mononen et al. 2019).

In Japan, consumers can purchase two categories of drugs at drug stores and community pharmacies without a prescription: over-the-counter (OTC) drugs and pharmaceuticals requiring pharmacist guidance. OTC drugs are classified as class I, class II, or class III drugs. When class I drugs and pharmaceuticals requiring pharmacist guidance are sold, pharmacists are required to counsel the consumer. However, class II and III drugs do not require pharmacist counseling and can be sold by registered salesclerks as well as pharmacists (Nakai and Tanaka 2015). Moreover, it has been legal to sell OTC drugs via the Internet since 2009 (2014 for class I and II drugs) (Nakai and Tanaka 2015; Nomura et al. 2016).

Although an online survey of 40,000 people in Japan conducted in 2009 showed that only about 12% of respondents had purchased OTC drugs via the Internet (Kishimoto et al. 2009), opportunities for consumers to purchase OTC drugs without face-to-face counseling are expected to increase year by year.

Therefore, we need to consider the increasing likelihood that consumers will consult generative AI applications such as ChatGPT for advice on medication. This study aimed to systematically evaluate responses from ChatGPT to consumer inquiries about various OTC drugs and to clarify the issues associated with using ChatGPT.

2. Investigations and results

2.1. Setting of ChatGPT, drugs, and questions

In this study, we used ChatGPT (ver. 3.5) with the "chat history & training" function deactivated. We selected 22 popular medications that are sold as OTC drugs in Japan. We used the generic name for 10 of the 22 drugs, and the brand name for the remaining 12 (Table 1). Next, we chose 12 consumer characteristics, including consumer background, medical conditions, and concomitant medications (Table 2). We input a total of 264 questions (*i.e.*, all combinations of drugs and characteristics) to ChatGPT in Japanese,

Table 1: OTC drug list

Brand name	Generic name	Form
-	Ibuprofen	Tablet
-	Acetaminophen	Tablet
-	Loxoprofen sodium hydrate	Tablet
-	Famotidine	Tablet
-	Fexofenadine hydrochloride	Tablet
-	Cetirizine hydrochloride	Tablet
-	Anhydrous caffeine	Tablet
-	Bisacodyl	Tablet
-	Nicotine	Gum
-	Kakkonto (Pueraria root, ephedra herb, jujube, cinnamon bark, peony root, glycyrrhiza, and ginger)	Extract granule
Drewell	Diphenhydramine hydrochloride	Tablet
Aneton	Codeine phosphate hydrate, dl-methylephedrine, chlorpheniramine maleate, anhydrous caffeine, and senega	Tablet
Bufferin EX	Loxoprofen sodium hydrate, dried aluminum hydroxide gel	Tablet
Ohta's Isan	Cinnamon bark, fennel, nutmeg, clove, citrus unshiu peel, gentian, powdered picrasma wood, sodium bicarbonate, precipitated calcium carbonate, magnesium carbonate, synthetic aluminum silicate, and biodiastase	Powder
Panchillon a	Lipase AP12, prozyme 6, biodiastase 2000, magnesium aluminosilicate, sodium bicarbonate, Scopolia extract, aldioxa, powdered glycyrrhiza, powdered turmeric, powdered cinnamon bark, powdered magnolia bark, powdered clove, and powdered ginseng	Powder
Tiovita Gold	Royal jelly, ginseng fluid extract, epimedium herb fluid extract, vitamins B1, B2, and B6, nicotinamide, and anhydrous caffeine	Liquid
Loxonin S tape	Loxoprofen sodium hydrate	Tape
Vantelin Kowa EX	Indomethacin, l-menthol	Gel
Aftach A	Triamcinolone acetonide	Adhesive tablet
Nazal spray	Naphazoline hydrochloride, chlorpheniramine maleate, and benzalkonium chloride	Spray
Preserace ointment	Hydrocortisone acetate, tetrahydrozoline hydrochloride, lidocaine, chlorpheniramine maleate, l-menthol, allantoin, tocopherol acetate, and chlorhexidine hydrochloride	Ointment
Lamsil DX	Terbinafine hydrochloride, crotamiton, glycyrrhetinic acid, l-menthol, and urea	Ointment

The underline indicates the name input to ChatGPT.

Table 2: List of consumer characteristics

Characteristics	
Consumer background	Pregnancy
	Elderly
	Driving
Medical conditions	Glaucoma
	Gastric ulcer
	Hypertension
	Hemodialysis
	History of asthma
Concomitant medications	Antihistamine
	Motion sickness drug
	Cough suppressant
	Pain reliever

asking whether it is safe for consumers with each characteristic to take these OTC drugs. Questions were asked in the order listed in Table 1. The question format was designed to maintain consistency throughout the study by using the following basic structure: "I am/have <CHARACTERISTIC>. Can I take <DRUG NAME>?"

2.2. Evaluation criteria

The obtained responses were evaluated based on the following three criteria: 1) coherence between the question and response, 2) scientific correctness, and 3) appropriateness of the instructed actions. The first two criteria were assessed with "yes" or "no" responses. The instructed actions were categorized into three groups: "allowed" (no special precautions or consultation required), "requires consultation (with a medical professional)," and "contraindicated." If the instructed actions from ChatGPT and the recommendations in the package insert fell into the same category, it was considered "consistent"; otherwise, it was considered "inconsistent."

2.3. Evaluation methods

To assess the reproducibility of ChatGPT's answers, we resubmitted the questions for which we received responses that satisfied all three criteria on a different day (second survey). The first survey was conducted from April 28 to May 8, 2023 and the second survey was conducted from June 6 to June 8, 2023. Two pharmacists independently evaluated the responses from ChatGPT, and the agreement between their assessments was verified using the κ coefficient. When their initial evaluations did not agree, consensus was reached through discussion and reconciliation.

2.4. Response rate of the criteria

The κ coefficient was 0.82, indicating excellent agreement between the two evaluators. Coherent responses were received to 79.5% of

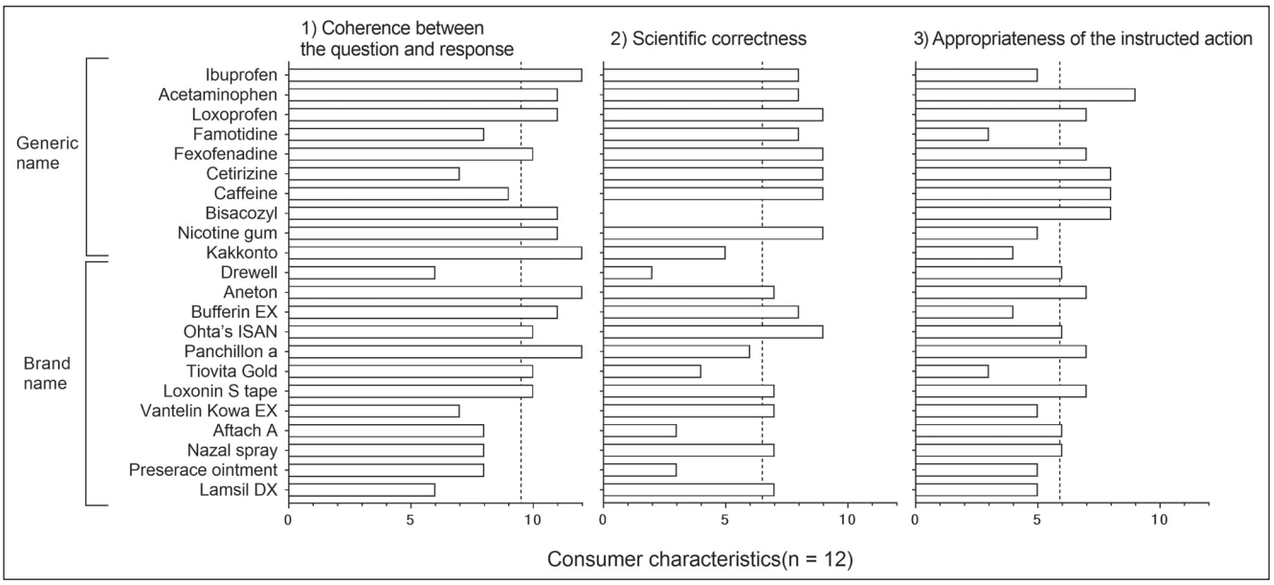

Fig. 1: Number of appropriate responses regarding the use of each drug with the consumer characteristics. The dotted line indicates the average value of all 22 drugs.

Table 3. Comparison of instructed actions between ChatGPT and package inserts

		Package inserts			
		Allowed	Requires consultation†	Contraindicated	Total
ChatGPT	Allowed	24	6	1	31
	Requires consultation†	81	96	28	205
	Contraindicated	6	7	11	24
	Others	2	1	1	4
	Total	113	110	41	264

† Recommended consulting a medical professional

the input questions. Scientifically correct responses were received to 144 of the 264 questions (54.5%). The instructed actions were appropriate in 131 of the 264 responses (49.6%). Only 55 of the 264 responses (20.8%) satisfied all three criteria. After resubmitting these 55 questions to ChatGPT during the second survey period, 61.8% (34/55) of the responses satisfied all three criteria.

2.5. Response rate of 22 drugs and 12 characteristics

Figure 1 shows the number of appropriate answers to the questions regarding the use of each drug for the 12 characteristics. When the questions were input using generic drug names, coherent responses were obtained in 102 of 120 cases (85%). However, when the questions were input using brand names, the rate of coherent responses was significantly lower, at 75% (108/144; $P = 0.048$). Similarly, the scientific accuracy of responses was 61.7% (64/120) to questions using generic names, while responses to questions using brand names had a significantly lower scientific accuracy of 48.6% (70/144; $P = 0.036$). Meanwhile, there was no statistical difference in the appropriateness of instructed actions between responses to the questions using generic names (53.3%) and those using brand names (46.5%).

ChatGPT misinterpreted bisacodyl as an antihypertensive agent in all cases, and responses to bisacodyl-related questions had the lowest scientific accuracy. Several misinterpretations of brand names were observed in the responses from ChatGPT; for example, Drewell® (active ingredient: diphenhydramine hydrochloride) was misinterpreted as containing diclofenac or as a proton pump inhibitor. Regarding the instructed actions, the major inconsistency between ChatGPT and package insert came from ChatGPT's tendency to recommend consulting a healthcare professional regarding the issues asked about, even if use of the drug under the given characteristic was allowed (not prohibited) according to the package inserts.

Figure 2 shows the number of appropriate responses to the questions about the use of the 22 OTC drugs under the 12 characteristics. Coherent responses were obtained to 62 of 66 (93.9%) questions regarding the consumer's background, 70 of 110 (63.6%) questions regarding medical conditions, and 78 of 88 (88.6%) questions regarding concomitant medications.

Similarly, scientifically correct responses were obtained to 42 of 66 (63.6%) questions regarding the consumer's background, 54 of 110 (49.1%) questions regarding medical conditions, and 48 of 88 (54.5%) questions regarding concomitant medications. Instructed actions consistent with package inserts were obtained to 48 (54.5%), 70 (63.6%), and 25 (28.4%) questions regarding the consumer's background, medical conditions, and concomitant medications, respectively. Regarding the instructed actions, the responses to questions involving the characteristics "car driving" and "concomitant medications" were less consistent with package inserts.

2.6. Comparison between ChatGPT's instructions and package inserts

Table 3 compares instructed actions between ChatGPT and package inserts. Among the 113 cases in which the package insert allowed the OTC drug to be used, ChatGPT recommended consulting with a medical professional in 81 cases (71.6%). In contrast, among 41 cases in which the drug was contraindicated by the package

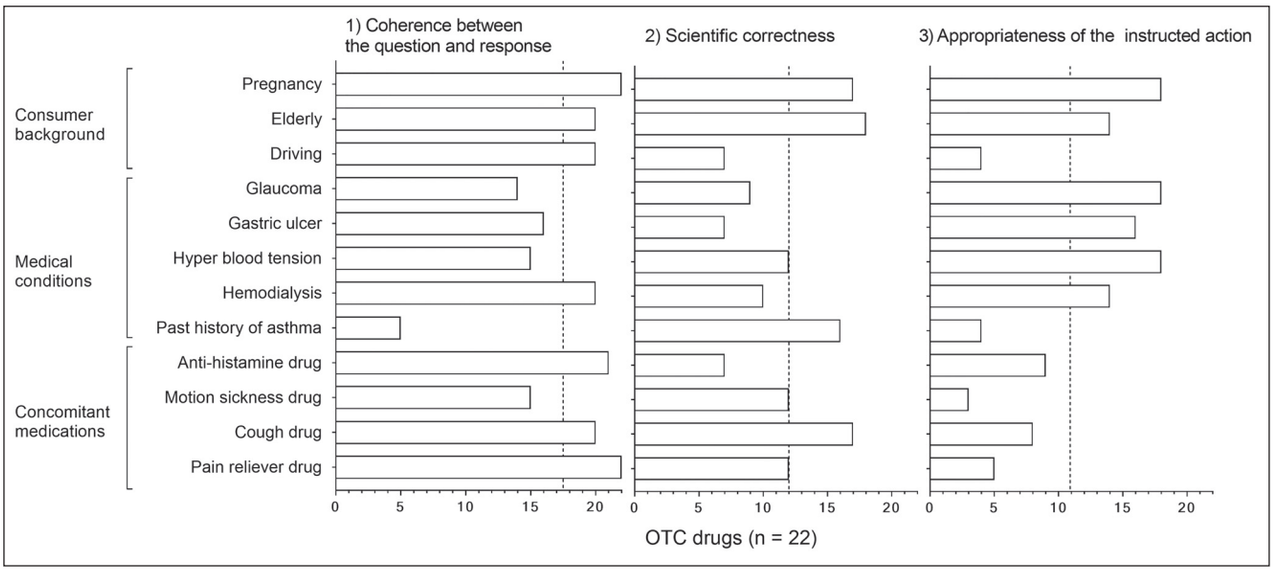

Fig. 2: Number of appropriate responses regarding the use of the OTC drugs under each consumer characteristic. The dotted line indicates the average value of all 12 consumer characteristics.

inserts, ChatGPT recommended consulting with a medical professional in 28 cases (68.3%). Indeed, ChatGPT showed a high tendency to recommend consultation (77.7%; 205 of the 264 combinations). In addition, there was one case in which ChatGPT allowed taking a drug under characteristics contraindicated by the package insert. ChatGPT responded that "Drewell can be taken concomitantly with antihistamines." Furthermore, ChatGPT allowed taking the drug in 6 cases where the package inserts recommended consulting a medical professional. Three of these 6 cases were concomitant administration of nicotine gum and other drugs.

3. Discussion

In this study, we examined the responses from ChatGPT to consumer inquiries about various OTC drugs. Only 20.8% of ChatGPT's responses satisfied the criteria for coherency, scientific accuracy, and appropriateness of the instructed actions and were therefore deemed satisfactory. This rate is comparable to the proportion of responses that satisfied only the latter two criteria (25.8%), suggesting that both the scientific accuracy and appropriateness of the instructed actions are the key factors in ChatGPT generating unsatisfactory responses.

Figure 1 shows that using brand names in questions led to less accurate responses compared with using generic names. Although the data on which ChatGPT was trained in terms of OTC drugs are not known, it seems not to have learned enough about the linkage between brand names and generic names, and/or between brand names and pharmacotherapeutic classes in Japan. Indeed, some responses included an obvious misconnection between the brand name and generic name or pharmacotherapeutic class. This suggests that ChatGPT is an unreliable tool for collecting medical information about the Japanese brand names in the submitted questions.

Among the 133 responses in which the instructed actions were inconsistent with the package inserts, 109 recommended consulting a healthcare professional. In these cases, although ChatGPT did not provide the final decision, the risk of serious outcomes may be low as long as the consumer follows this recommendation. Taken together, in 24 of 264 (9.1%) cases, ChatGPT's responses may harm consumers by leading to adverse events. Among these, there were 7 cases in which ChatGPT allowed taking medicine that is contraindicated or requires consultation with a medical professional. These results are consistent with a previous report indicating that using ChatGPT for medical advice is both beneficial and risky, and should be reviewed by medical experts (Sajjad and Saleem 2023). Because incorrect medical information is already disseminated on the Internet, it is just as risky for consumers to indiscriminately rely on responses from ChatGPT, which are based on the collective knowledge from the Internet, as it is to blindly rely on information from the Internet.

Regarding the status "car driving," only 18.2% (4/22) of the actions instructed by ChatGPT were consistent with the package inserts, mainly because ChatGPT recommended consulting a healthcare professional about the potential risk of drowsiness in 57.9% (11/19) of cases, even though doing so is not required. Similarly, the instructed actions regarding concomitant medication were inconsistent between ChatGPT and the package inserts (matching rate of 28.4%), again because 63.5% (40/63) of the responses from ChatGPT recommended consulting a healthcare professional. In contrast, the instructed actions regarding medical conditions were more consistent with those in the package inserts because the package inserts recommend consultations for patients under the care of a physician or dentist.

Minor changes to the text of a question can sometimes lead to different responses from ChatGPT. As mentioned above, ChatGPT allowed the concomitant administration of Drewell and an antihistamine, even though doing so is contraindicated by the package insert. By making the question more specific (e.g., "I am taking an OTC drug named Drewell, which contains diphenhydramine. Now I have hay fever and would like to take an antihistamine. Will this cause a problem?"), we received a more favorable response: "Duplication of antihistamines may increase the risk of excessive effects and side effects. If you are taking Drewell, which contains diphenhydramine, I recommend that you consult a doctor or pharmacist before taking additional antihistamines." Thus, the more specific the wording we use, the more favorable the responses we can get from ChatGPT.

Responses from ChatGPT are not fully reproducible. Morath et al. (2023) investigated the reproducibility of ChatGPT responses to drug-related questions. They input the same 12 questions four times each (the first time, the following day, the following week, and two weeks later) and received the same response in only 3 of the 12 cases. In the present study, responses that satisfied all three criteria were consistent in 61.8% of cases in the second survey. Therefore, when using ChatGPT for medication counseling or collecting medical information, we should consider the lack of reproducibility.

In this study, we input the questions in Japanese. Given that the intelligibility of responses is reported to be lower for questions

input in French and Arabic compared with those input in English (Seghier 2023), it may be necessary to examine differences in the accuracy of responses from ChatGPT's to medication-related questions between Japanese and English.

In this study, we arbitrarily selected target OTC drugs. Hence, the current findings may not accurately reflect the issues arising from consumers' inquiries to ChatGPT regarding OTC drugs. However, to better simulate real-world scenarios for this issue, we selected mainly class II drugs as the target drugs (18/22) because most of the active ingredients in OTC drugs sold in Japan are categorized as class II (Ministry of Health, Labour and Welfare of Japan 2022; Ministry of Health, Labour and Welfare of Japam 2023). Furthermore, consultations with medical professionals are not mandatory when selling class II OTC drugs in Japan, although these drugs have a relatively high risk, and thus safety information should be provided to prevent rare but possible severe adverse reactions that may require hospitalization (Nomura et al. 2016). Indeed, most of the target drugs in the present study considerably overlap with OTC drugs that are frequently purchased or used improperly by consumers in Japan. While Masuyama conducted a review of OTC drug sales in Japan, reporting 164 cases of improper purchases due to complications, duplications, and more, it is worth noting that 66 out of these 164 cases (40%) involved OTC drugs targeted in the current study (Masuyama 2021). He also surveyed 31 cases of OTC drug overdose and reported that the causative agents were diphenhydramine, caffeine, antipyretic analgesics, antitussives, and cold remedies. Notably, 61% of these cases included the drugs targeted in the present study. Taken together, although we arbitrarily selected the target drugs, the present study is considered to well simulate the real-world scenarios. Because a risk of potentially harmful responses from ChatGPT was identified in this study, consumers are advised to not rely on ChatGPT for medication counseling related to OTC drugs.

Recently, the use of machine-learning technologies to assist pharmacists has been reported, including the detection of overdose and underdose (Nagata et al. 2021), along with models for detecting prescriptions that require pharmacist intervention (Balestra et al. 2021). However, the use of generative AI in the pharmacy is less popular. Meanwhile, an example of using AI to improve patients' medication management has been reported (Eggerth et al. 2020). This technology checks whether patients have taken their medications and reminds them to do so if they have not. However, these applications have been developed from the perspective of healthcare professionals and are not yet widely applied. This study is the first to analyze consumers' use of generative AI to collect medical information and related issues.

One of the limitations of this study is the rapidly evolving nature of generative AI, raising uncertainty regarding the duration of the study's validity. Furthermore, as previously noted, caution should be exercised when generalizing the present results, as only a limited number of OTC drugs were arbitrarily selected as the study's focus, although a significant portion of OTC-related issues are potentially associated with the targeted drugs.

In conclusion, this study revealed that ChatGPT is not a reliable tool for obtaining information about OTC drugs. At present, consumers are advised not to use ChatGPT for OTC medication counseling, especially for Japanese brand names. As the technologies underlying generative AI continue to advance, it will be important to continuously monitor the trends in consumer use of generative AI as well as its capability for providing medication counseling.

4. Experimental

4.1. Statistical analysis

Statistical differences in occurrence probabilities between the two groups were assessed using Fisher's exact test, and a P value of less than 5% was considered statistically significant.

Conflicts of interest: None declared.

References

Ashkanani H, Asery R, Bokubar F, AlAli N, Mubarak S, Buabbas A, Almajran (2019) Web-Based Health Information Seeking Among Students at Kuwait University: Cross-Sectional Survey Study. JMIR Form Res 3: e14327.

Balestra M, Chen J, Iturrate E, Aphinyanaphongs Y, Nov O (2021) Predicting inpatient pharmacy order interventions using provider action data. JAMIA Open 4: ooab083.

Bergmo TS, Sandsdalen V, Manskow US, Småbrekke L, Waaseth M (2023) Internet use for obtaining medicine information: Cross-sectional survey. JMIR Form Res 7: e40466.

Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW (2002) Patients' use of the internet for medical information. J Gen Intern Med 17: 180–185.

Eggerth A, Hayn D, Schreier G (2020) Medication management needs information and communications technology-based approaches, including telehealth and artificial intelligence. Br J Clin Pharmacol 86: 2000–2007.

Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D (2023) How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 9: e45312.

Hämeen-Anttila K, Pietilä K, Pylkkänen L, Pohjanoksa-Mäntylä M (2018) Internet as a source of medicines information (MI) among frequent internet users. Res Social Adm Pharm. 14: 758–764.

Kishimoto K, Yoshida T, Fukushima N (2009) Factors related to purchasing over-the-counter medications online. Yakugaku Zasshi 129: 1127–1136.

Kunze KN, Jang SJ, Fullerton MA, Vigdorchik JM, Haddad FS (2023) What's all the chatter about?. Bone Joint J 105-b: 587–589.

Masuyama, K (2021) Exploring Effective Methods for Pharmacists to Manage and Provide Over-the-Counter Drug Information (in Japanese), Report of the Health, Labour and Welfare Scientific Research, Ministry of Health, Labour and Welfare, Japan
https://mhlw-grants.niph.go.jp/project/155838 (accessed 26 October 2023)

Ministry of Health, Labour and Welfare of Japan (2022) https://www.mhlw.go.jp/content/001025495.pdf. (accessed 3 July 2023).

Ministry of Health, Labour and Welfare of Japan (2023) https://www.mhlw.go.jp/content/001083984.pdf. (accessed 3 July 2023).

Mononen N, Airaksinen MSA, Hämeen-Anttila K, Helakorpi S, Pohjanoksa-Mäntylä M (2019) Trends in the receipt of medicines information among Finnish adults in 1999-2014: a nationwide repeated cross-sectional survey. BMJ Open 9: e026377.

Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, Hoppe-Tichy T, Green K (2023) Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. Eur J Hosp Pharm doi: 10.1136/ejhpharm-2023-003750.

Nagata K, Tsuji T, Suetsugu K, Muraoka K, Watanabe H, Kanaya A, Egashira N, Ieiri I (2021) Detection of overdose and underdose prescriptions-An unsupervised machine learning approach. PLoS One 16: e0260315.

Nakai K, Tanaka T (2015) Introducing internet retailing of OTC drugs in Japan: Revision of the pharmaceutical affairs law. Ther Innov Regul Sci 49: 284–288.

Nomura K, Kitagawa Y, Yuda Y, Takano-Ohmuro H (2016) Medicine reclassification processes and regulations for proper use of over-the-counter self-care medicines in Japan. Risk Manag Healthc Policy 9: 173–183.

Sajjad M, Saleem R (2023) Evolution of healthcare with ChatGPT: A word of caution. Ann Biomed Eng 51: 1663–1664.

Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 11: 887.

Seghier ML (2023) ChatGPT: not all languages are equal. Nature 615: 216.

van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL (2023) ChatGPT: five priorities for research. Nature 614: 224–226.