Skin Health and Disease 2025; **5**:14–21
https://doi.org/10.1093/skinhd/vzae015
Advance access publication date: 20 January 2025

SHD
Skin Health and Disease
Original Article

Assessing the readability of dermatological patient information leaflets generated by ChatGPT-4 and its associated plugins

Dominik Todorov,[1] Jae Yong Park,[1] James Andrew Ng Hing Cheung,[1] Eleni Avramidou[2] and Dushyanth Gnanappiragasam[3]

[1]School of Medicine, Imperial College London, London, UK
[2]School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece
[3]Department of Dermatology, Imperial College Healthcare NHS Trust, London, UK

Correspondence: Dominik Todorov. Email: dominik.todorov19@imperial.ac.uk

Abstract

Background In the UK, 43% of adults struggle to understand health information presented in standard formats. As a result, Health Education England recommends that patient information leaflets (PILs) be written at a readability level appropriate for an 11-year-old.

Objectives To evaluate the ability of ChatGPT-4 and its three dermatology-specific plugins to generate PILs that meet readability recommendations and compare their readability with existing British Association of Dermatologists (BAD) PILs.

Methods ChatGPT-4 and its three plugins were used to generate PILs for 10 preselected dermatological conditions. The readability of these PILs was assessed using three readability formulas Simple Measure of Gobbledygook (SMOG), Flesch Reading Ease Test (FRET) and Flesch–Kincaid Grade Level Test (FKGLT) and compared against the readability of BAD PILs. A one-way ANOVA was conducted to identify any significant differences.

Results The readability scores of PILs generated by ChatGPT-4 and its plugins did not meet the recommended target range. However, some of these PILs demonstrated more favourable mean readability scores compared with those from the BAD, with certain plugins, such as Chat with a Dermatologist, showing significant differences in mean SMOG ($P=0.0005$) and mean FKGLT ($P=0.002$) scores. Nevertheless, the PILs generated by ChatGPT-4 were found to lack some of the content typically included in BAD PILs.

Conclusions ChatGPT-4 can produce dermatological PILs free from misleading information, occasionally surpassing BAD PILs in terms of readability. However, these PILs still fall short of being easily understood by the general public, and the content requires rigorous verification by healthcare professionals to ensure reliability and quality.

What is already known about this topic?

- Patient information leaflets (PILs) are essential tools in contemporary medical practice.
- Low health literacy is prevalent among a significant portion of the population, making it difficult to understand standard health information materials.
- ChatGPT-3.5 has demonstrated potential in creating dermatology-focused PILs.

What does this study add?

- ChatGPT-4 and its associated plugins can, in some cases, produce PILs with more favourable readability scores compared with the current British Association of Dermatologists (BAD) PILs.
- The content of PILs generated by ChatGPT-4 and its plugins may still be missing information typically found in existing BAD PILs.

In the UK, 43%[1] of adults are unable to comprehend commonly used health information materials. This percentage increases to 61%[1] when both health literacy and numeracy skills are assessed. This is especially concerning, given the increasing trend of patients seeking to be involved in treatment decisions.[2]

The importance of patient educational materials and written information is paramount in modern medical practice, as it plays a crucial role in achieving high treatment adherence. Providing patients with written materials allows them to revisit the information they may have forgotten after the consultation.[3] In dermatology, where patients can present

Accepted: 20 November 2024

with a wide range of conditions, having information leaflets readily available is particularly beneficial. However, with over 1500 skin conditions[4] and various presentations, creating individual leaflets for each condition can be extremely time-consuming.

The British Association of Dermatologists (BAD) has created a multitude of patient information leaflets (PILs).[5] These PILs undergo rigorous assessments to ensure they cover all relevant aspects, including a readability evaluation by the BAD's Patient Information Lay Review Panel.[6] Ultimately, readability is a crucial aspect when developing PILs, being included as one of the four main targets that need assessing when developing written patient information material. This is mainly due to the increasing availability of health information and low health literacy rates.[7]

Consequently, AI (artificial intelligence) may potentially help in the development of PILs that achieve lower readability levels, aiming for maximum inclusivity across the general population. John McCarthy defined AI as 'the science and engineering of making intelligent machines', a concept rooted in the ability to simulate human-like intelligence and critical thinking.[8] Since the 1950s, AI has expanded significantly, with a recent surge in popularity of large language models like Chat Generative Pretrained Transformer[9] (ChatGPT) (OpenAI, San Francisco, CA, USA). ChatGPT comes in multiple versions, including ChatGPT-4, which, at the time of writing this paper, remains behind a paywall and is not freely accessible. Additionally, the integration of plugins with ChatGPT-4 can enhance its capabilities by providing real-time, personalized and specific information beyond its training datasets, thereby improving the system's overall effectiveness.[10] Unfortunately, these plugins are only available to users with access to ChatGPT-4, making them inaccessible to free version users.

Previous responses generated by ChatGPT have shown a predominantly benign nature, suggesting potential utility in healthcare.[11] A prior study by Verran[12] demonstrated that ChatGPT-3.5 could generate relevant information in a comprehensible format for patients, particularly in situations where a condition-specific PIL was not readily accessible. However, the quality of information generated was lower compared with the standards upheld by the BAD.

Therefore, the aim of this study was to explore whether ChatGPT-4 and its associated dermatology-focused plugins can produce PILs that meet BAD-level standards. The study sought to assess whether these tools can improve readability and streamline the development of PILs, thereby potentially enhancing the availability of health information.

Materials and methods

Dermatological conditions selection

The 10 most prevalent dermatological conditions[13] in individuals aged 18 and older were selected for analysis: acne, atopic dermatitis, alopecia, psoriasis, rosacea, squamous cell carcinoma (SCC), basal cell carcinoma (BCC), hidradenitis suppurativa, vitiligo and melanoma. PILs for these conditions were generated using both ChatGPT-4 (version 4; OpenAI, San Francisco, CA, USA) and three ChatGPT-4 based plugins. As the BAD provides four distinct PILs for each stage of melanoma, the PIL for melanoma *in situ* was used instead to maintain a consistent one-to-one comparison between the PILs.

Generation of patient information leaflets using ChatGPT

Three ChatGPT plugins designed for dermatological advice, including 'Dr. Dermatology', 'Dermatology Adviser' and 'Chat with a Dermatologist', along with ChatGPT-4 were used in this study. These tools were employed to generate PILs for all 10 conditions using the prompt 'Create a patient information leaflet about [one of the ten conditions]' following prompt engineering techniques.[14] Maintaining a consistent prompt was crucial, as any modifications would potentially alter the resulting PIL.[15]

Patient information leaflet analysis

After an independent review by two authors (D.T. and J.Y.P.) to ensure the generated PILs were sensible and accurate, a readability analysis was conducted. Non-narrative text was removed, and the PILs were standardized in format to ensure accurate evaluation. The readability analysis was performed using Readable software (Readable by Added Bytes Ltd; available from: https://readable.com/), which includes readability measures that have been widely used in previous studies.[16] The readability results were then compared with the readability of PILs generated by ChatGPT-4 to assess the impact of dermatology-specific plugins on the produced PILs. Additionally, the readability of existing BAD PILs for the corresponding conditions was evaluated. The mean readability of BAD PILs was then compared with that of PILs generated by ChatGPT-4 and the three plugins. To evaluate the consistency of PILs produced by ChatGPT, acne PILs were regenerated using the three ChatGPT plugins and ChatGPT-4, with their readability assessed and compared with the readability of the initially generated PILs. After reviewing the content of the 10 selected BAD PILs, two authors (D.T. and J.Y.P.) independently identified the most commonly occurring subsections. Any disagreements were resolved through a consensus discussion with the senior author (D.G.). These identified subsections provided a framework for evaluating the content of PILs generated by ChatGPT-4 and its plugins. Each generated PIL was assessed on whether it adequately covered each subsection, with a point awarded for every relevant section addressed, reflecting typical content found in a BAD PIL. Lastly, the generated PILs were evaluated for the presence of any misleading content.

Readability assessment: Flesch Reading Ease Test, Simple Measure of Gobbledygook, Flesch–Kincaid Grade Level Test

The Flesch Reading Ease Test (FRET)[17] scores range from 0 to 100, with higher scores indicating easier readability, and a score of 100 representing text easily understood by an average 11-year-old. The score is calculated by evaluating the content's sentence and word length; short sentences with shorter words receive higher scores, making the text simpler and easier to read. Conversely, longer sentences with longer words result in lower scores, suggesting that the content may be more suited for advanced readers. In contrast, the Flesch–Kincaid Grade Level Test (FKGLT)[18] estimates the US school grade level required to comprehend the text. For example, a FKGLT score of 8 indicates that an eighth grader should be able to understand the material. Although FRET and the FKGLT use the same fundamental metrics, such as word length and sentence length, they apply distinct weighting factors to them, leading to a discrepancy resulting in an inverse relationship between the two tests. Therefore, a FRET score of 60–70 would be an equivalent to a FKGLT score of 8–9, making the text standard for understanding in the eighth grade. The Simple Measure of Gobbledygook (SMOG)[19] score estimates the number of years of education needed for the average person to understand the text. The score is based on the number of polysyllabic words in a text but does not account for sentence complexity or length, unlike FRET and FKGLT. While FRET and FKGLT are widely used for readability assessment, SMOG was also included, as it has been found to be more suitable for evaluating healthcare-related information.[16]

In the UK, Health Education England[20] recommends that materials be targeted to the reading level of an average 11- to 14-year-old for broader accessibility, which corresponds to a SMOG score between 5 and 8, with 5 representing the readability level of an 11-year-old. Ideally, FRET scores should fall between 60 and 70, and FKGLT scores should range between 8 and 9 to meet the recommended readability standards.

Statistical analysis

To assess the normality of the data, the Shapiro–Wilk test was applied. Normally distributed data were presented as mean (SD), while non-normally distributed data were presented as median (interquartile range; IQR). The Levene test was used to assess the equality of two variances. If equal, a one-way ANOVA was conducted to identify any significant differences between the groups, followed by a Tukey post hoc analysis. A probability value of 0.05 or less was considered as a significant result. Statistical analysis was performed using IBM SPSS (Released 2023; IBM SPSS Statistics for Windows, Version 29, IBM, Armonk, NY, USA). All figures were created using Microsoft Excel (Released 2021; Microsoft).

Results

Overall results

A total of 44 PILs were generated, with 1 PIL being generated for each condition using ChatGPT-4 and the 3 ChatGPT plugins, resulting in 40 PILs. Additionally, the acne PILs were re-generated, adding four more to the total. Examples of the generated PILs can be found in Appendix S1 (see Supporting Information).

Readability analysis

Readability analysis was conducted for both ChatGPT-4 and the ChatGPT plugins, with results presented as raw scores along with their mean values and SDs, as shown in Table 1. The mean SMOG scores ranged from 10.7 to 11.5, with Chat with a Dermatologist having the lowest mean score and Dr. Dermatology the highest. None of the outputs from ChatGPT-4 or the plugins met the recommended SMOG readability score. The highest mean FRET score was 54.2, which falls below the recommended FRET range. Additionally, mean FKGLT scores ranged from 8.3 to 9.5, indicating that the PILs generated by ChatGPT may not be easily understandable by the general public. For specific conditions like vitiligo and alopecia, the SMOG and FKGLT scores were notably higher when compared with the other dermatological conditions. The statistical comparison of mean SMOG scores showed no significant differences across the four groups. However, the mean FRET score analysis revealed a significant difference ($P = 0.005$), with the Tukey post hoc test identifying significant differences between Dr. Dermatology and Chat with a Dermatologist with a 9.13 difference in mean readability scores (45.07 vs. 54.20; $P = 0.006$). Additionally, Dermatology Adviser and Chat with a Dermatologist showed a 7.95 difference in mean readability scores (46.25 vs. 54.20; $P = 0.023$). Finally, the FKGLT score analysis indicated a significant difference among the four groups ($P = 0.023$), with a significant difference between Dr. Dermatology and Chat with a Dermatologist, with a 1.23 difference in mean readability scores (9.53 vs. 8.30; $P = 0.018$).

Table 1 Readability scores for patient information leaflets generated by ChatGPT-4 and ChatGPT plugins across 10 selected dermatological conditions, including mean (SD) for each metric

	Dr. Dermatology			Dermatology Adviser			Chat with a Dermatologist			ChatGPT-4		
	SMOG	FRET	FKGLT	SMOG	FRET	FKGLT	SMOG	FRET	FKGLT	SMOG	FRET	FKGLT
Acne	10.9	43.4	9.5	11.1	46.1	9.2	9.2	64.8	6.5	9.8	52.6	8.1
Atopic dermatitis	11.4	38.4	10.1	10.1	45.1	8.7	9.9	57.6	7.3	11.7	41.9	9.9
Alopecia	12.6	36.3	10.9	12.2	41.5	10.1	11.9	42.6	9.9	12.2	41.2	10.3
Psoriasis	11.9	40.1	10.3	10.6	43.7	9.0	10.8	54.2	8.2	11.8	48.3	9.5
Rosacea	10.4	50.6	8.4	10.3	43.1	9.0	11.2	49.7	8.8	11.0	48.2	8.9
SCC	11.0	52.3	8.5	11.5	48.6	9.3	10.3	60.5	7.6	11.3	53.9	8.7
BCC	11.1	53.0	8.6	11.2	54.9	8.5	10.1	62.1	7.6	10.8	54.9	8.2
Hidradenitis suppurativa	10.8	48.4	9.0	10.8	47.9	9.0	10.4	55.0	8.0	11.1	44.4	9.4
Vitiligo	12.6	40.1	10.7	12.4	43.4	10.2	12.2	45.3	10.2	12.7	40.8	10.7
Melanoma	11.8	48.1	9.3	11.8	48.2	9.4	11.4	50.2	8.9	11.4	48.3	9.0
Mean (SD)	11.45 (0.76)	45.07 (6.14)	9.53 (0.93)	11.20 (0.78)	46.25 (3.88)	9.24 (0.55)	10.74 (0.94)	54.20 (7.28)	8.30 (1.16)	11.38 (0.80)	47.45 (5.26)	9.27 (0.86)

BCC, basal cell carcinoma; FKGLT, Flesch–Kincaid Grade Level Test; FRET, Flesch Reading Ease Test; SCC, squamous cell carcinoma; SMOG, Simple Measure of Gobbledygook.

BAD patient information leaflet readability assessment

The readability of the BAD PILs was also analysed, with results presented as raw scores along with their mean values and SDs, as shown in Table 2. The analysis indicated that the mean scores for SMOG, FRET and FKGLT did not align with the recommended readability ranges. A one-way ANOVA was performed to compare the mean readability of BAD PILs with those generated by ChatGPT-4 and the three plugins, revealing a significant difference in mean SMOG scores ($P=0.001$), particularly between BAD and Dermatology Adviser, with a 1.06 difference in mean readability score of (12.26 vs. 11.20; $P=0.025$), and BAD and Chat with a Dermatologist, with a 1.52 difference in mean readability score (12.26 vs. 10.74; $P=0.0005$), as determined by Tukey post hoc analysis. While there were no significant differences in mean FRET scores between the ChatGPT-generated PILs and BAD PILs, there was a significant difference in mean FKGLT scores ($P=0.004$), specifically between BAD and Chat with a Dermatologist, with a 1.51 difference in mean readability score (9.81 vs. 8.30; $P=0.002$). The mean readability scores for SMOG and FKGLT for the PILs generated by ChatGPT-4, its three plugins and the BAD PILs are shown in Figure 1. Additionally, the mean readability scores for FRET are shown in Figure 2.

Table 2 Readability scores for existing British Association of Dermatologists (BAD) patient information leaflets, covering selected dermatological conditions, with melanoma replaced by melanoma *in situ*

	SMOG	FRET	FKGLT
Acne	12.2	52.5	10.0
Atopic dermatitis	12.9	49.9	9.9
Alopecia	12.4	51.8	9.8
Psoriasis	12.4	44.3	10.2
Rosacea	11.8	49.9	9.2
SCC	11.9	53.9	9.4
BCC	11.6	58.0	9.0
Hidradenitis suppurativa	12.8	43.8	10.8
Vitiligo	12.7	45.6	10.6
Melanoma *in situ*	11.9	57.4	9.2
Mean (SD)	12.26 (0.45)	50.71 (5.04)	9.81 (0.61)

BCC, basal cell carcinoma; FKGLT, Flesch–Kincaid Grade Level Test; FRET, Flesch Reading Ease Test; SCC, squamous cell carcinoma; SMOG, Simple Measure of Gobbledygook.

Assessing consistency through comparing the readability of re-generated patient information leaflets

Table 3 presents the results of the readability analysis for acne PILs, re-generated using ChatGPT-4 and the three ChatGPT plugins. It also includes readability scores from the initially created PILs as a benchmark for comparison. The findings reveal slight variations in readability scores, with the most significant change observed in the PIL re-generated by the ChatGPT plugin Dr. Dermatology.

Analysing the similarity between ChatGPT-4 and ChatGPT plugin-generated patient information leaflet content against BAD patient information leaflets

Table 4 presents the assessment, comparing the content similarity of PILs generated by ChatGPT-4 and the three ChatGPT plugins against BAD PILs.

All generated responses include a section advising consultation with a healthcare professional. Notably, no instances of misleading information were found in any of the PILs produced by ChatGPT-4 or the three ChatGPT plugins.

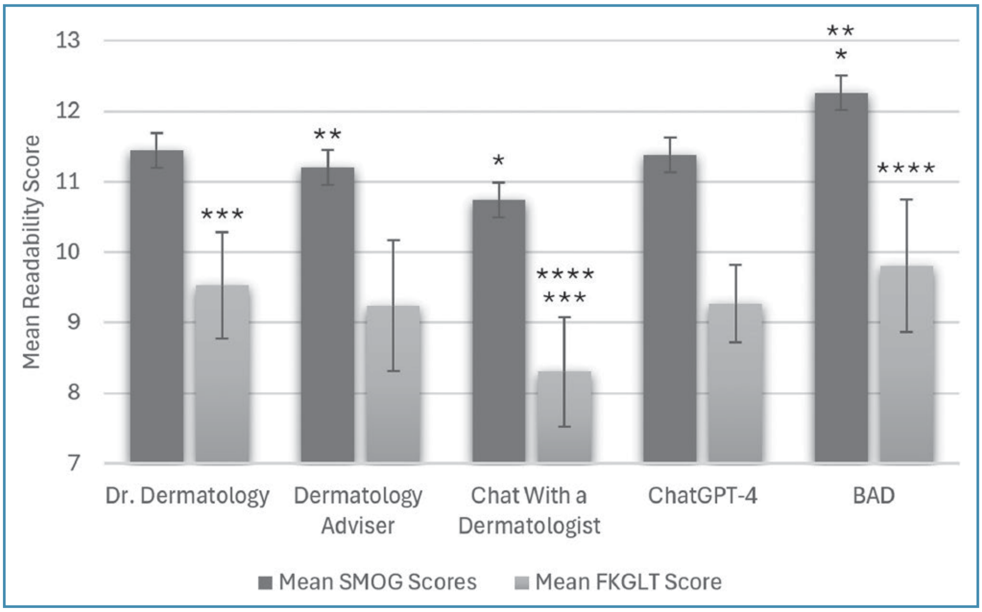

Figure 1 Visual representation of the differences in mean readability scores (SMOG and FKGLT) across ChatGPT-4, ChatGPT plugins and the BAD patient information leaflets, displayed as mean readability score (SD). Statistically significant differences between the mean readability scores are indicated on the figure, with specific comparisons highlighted: * denotes a significant difference between BAD and Chat with a Dermatologist' ($P=0.0005$), ** between BAD and Dermatology Adviser ($P=0.025$), *** between Dr. Dermatology and Chat with a Dermatologist ($P=0.018$) and **** between BAD and Chat with a Dermatologist ($P=0.002$). BAD, British Association of Dermatology; FKGLT, Flesch–Kincaid Grade Level Test; SMOG, Simple Measure of Gobbledygook.

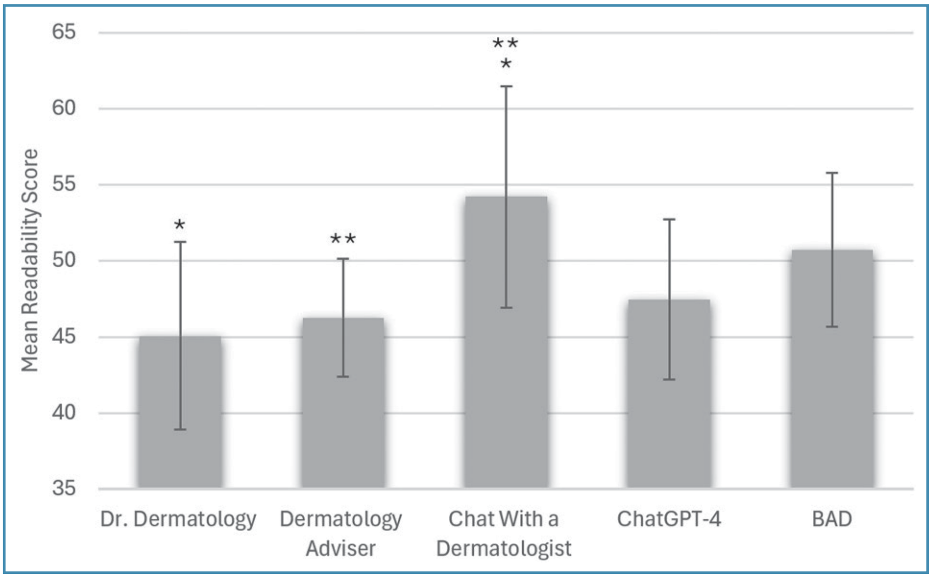

Figure 2 Visual representation of the differences in mean Flesch Reading Ease Test readability scores across ChatGPT-4, ChatGPT plugins and the British Association of Dermatologists (BAD) patient information leaflets, represented as mean readability score (SD). Statistically significant differences between the mean scores are marked on the figure, with specific comparisons highlighted: * indicates a significant difference between Dr. Dermatology and Chat with a Dermatologist (*P*=0.006), while ** marks the difference between Dermatology Adviser and Chat with a Dermatologist (*P*=0.023).

Discussion

This study assessed the use of ChatGPT-4 and ChatGPT plugins in the creation of dermatological PILs. The results showed that while the ChatGPT-generated PILs did not cover all the content typically included in a standard BAD PIL, their readability, in particular cases, was better than the corresponding PILs from the BAD. However, the readability of these PILs still did not meet the standards recommended by Health Education England. Previous research[12] supports this finding, showing that while ChatGPT-3.5 can generate PILs with appropriate content, their readability scores may still render them inaccessible to most people. Similarly, a study looking at ChatGPT-4's responses on hypothyroidism during pregnancy[21] indicated high readability scores, suggesting that the content produced would require a college level of understanding.

ChatGPT has numerous potential applications in medicine, from aiding in medical research and diagnosis to managing patients' health conditions.[22] This study specifically focused on the paid version of ChatGPT, ChatGPT-4, rather than the freely available ChatGPT-3.5, due to its superior performance compared with ChatGPT-3.5.[23] Notably, ChatGPT-4 has been shown to outperform, in certain instances, even human professionals when answering medical questions.[24] Most importantly, previous research has demonstrated that

Table 3 Uniformity in readability of re-generated acne patient information leaflets by ChatGPT-4 and three ChatGPT plugins, evaluated using three readability metrics

	PIL type	SMOG	FRET	FKGLT
ChatGPT-4	Original	9.8	52.6	8.1
	Re-generated	9.4	52.3	7.8
Dr. Dermatology	Original	10.9	43.4	9.5
	Re-generated	11.7	40.1	10.2
Dermatology Adviser	Original	11.1	46.1	9.2
	Re-generated	11.2	41.8	9.9
Chat with a Dermatologist	Original	9.2	64.8	6.5
	Re-generated	9.2	64.8	6.5

FKGLT, Flesch–Kincaid Grade Level Test; FRET, Flesch Reading Ease Test; PIL, patient information leaflet; SMOG, Simple Measure of Gobbledygook.

Table 4 Comparison of the most common subsections in existing British Association of Dermatology patient information leaflets (PILs) with content inclusion in patient information leaflets generated by ChatGPT-4 and ChatGPT plugins

	Dr. Dermatology	Dermatology adviser	Chat with a Dermatologist	ChatGPT-4
Aims of the PIL	0	0	0	0
More about the condition	10	10	10	10
What causes the condition	10	10	10	10
Is the condition hereditary?	3	5	1	6
What does the condition look like?	10	10	10	10
What does the condition feel like?/Symptoms	10	9	10	10
How is the condition diagnosed?	8	8	7	9
Can the condition be cured?	4	3	4	3
What is the treatment?	10	10	10	10
Self-care advice	10	10	10	10
Additional information for the patient	0	10	0	0

ChatGPT-4 improves multiple aspects of patient information, including how accurate and up-to-date the information is, compared with patient information generated by ChatGPT-3.5.[25] Given the abundance of dermatological misinformation available online,[26] it is crucial to combat this with readily available, high-quality patient information.

This study found that ChatGPT-4 and its plugins can generate PILs with appropriate condition-specific information; however, no significant difference was observed between ChatGPT-4 and its plugins in terms of readability. Despite the additional features dermatology-orientated plugins aim to provide, they still fall short of the specialized expertise that healthcare professionals offer. Additionally, our analysis also revealed that the BAD produces PILs with texts that are generally more challenging to comprehend, as shown by our readability assessments. This finding is consistent with a previous study[27] examining over 200 BAD-produced PILs, which found that many are too complex for some patients to fully understand. Although none of the ChatGPT plugins demonstrated an overall improvement in readability compared with ChatGPT-4, the Chat with a Dermatologist plugin produced PILs with significantly better mean readability than those generated by both the BAD and the other two plugins. This suggests that certain plugins can play a valuable role in developing and improving future PILs.

Since none of the PILs generated in this study fulfilled the recommended readability, future studies should explore specifying a desired readability level in the prompt. A study[28] on online orthopaedic surgery patient education materials demonstrated that ChatGPT was able to lower the reading level to that suitable for a sixth-grade student by specifying a desired reading level in the prompt. This suggests that targeting a specific readability level could lead to the creation of more accessible PILs.

However, as shown in Table 4, sections on cure and heritability were notably absent from most of the ChatGPT-generated PILs, which tend to be of interest to patients. This highlights the advantage of human-created PILs addressing the specific needs and questions more effectively. Given that the content depth and level of detail in ChatGPT-generated PILs do not match the comprehensive content found in BAD-produced PILs, this would indicate that relying solely on ChatGPT-generated material would not be advisable.[29] The literature on ChatGPT-generated content is mixed, with some studies reporting accurate and appropriate content,[30] while others highlight a lower quality of content.[31] Given ChatGPT-4's potential, its integration should be accompanied by professional oversight to enhance the production of patient education materials, rather than relying on it as the sole creator.

Additionally, for specific dermatological conditions, such as vitiligo and alopecia, the SMOG and FKGLT readability scores were higher compared with other conditions. Several factors contribute to this discrepancy. Firstly, these conditions may be less well understood in terms of their pathophysiology and management, which can make it challenging to develop PILs that adhere to recommended readability levels. Another contributing factor could be a clinical lack of awareness or biases against these conditions. For instance, in a study examining patients with vitiligo, primary concerns included a perceived lack of awareness about available treatments among general practitioners, along with the perception that vitiligo is frequently dismissed by healthcare professionals as a primarily 'cosmetic' issue.[32] Such factors may lead to the production of PILs that are more challenging to understand, potentially exacerbating patients' uncertainty and diminishing the perceived credibility of online resources. This, in turn, could limit public understanding and reduce awareness of these conditions. Therefore, it is essential to improve the readability of PILs for conditions like vitiligo and alopecia, while also addressing patient concerns regarding the accessibility of reliable and trustworthy information.

Like any study, this one has certain limitations. For example, the choice of font, such as sans serif, can significantly improve text readability for individuals with dyslexia,[33] an aspect that may not be captured by readability assessment software. Additionally, it is important to recognize that automated readability tools are not infallible or perfectly consistent,[34] making it inappropriate to rely on a single tool for evaluating readability and patient suitability. It is recommended to seek validation from a patient panel or conduct multiple readability assessments when developing and reviewing PILs. Furthermore, the cross-sectional design of this study reflects outcomes at a specific point in time, which are subject to change over time.

Given the relatively small sample size of this study, future research should involve a more extensive examination across various conditions. Moreover, as demonstrated in Table 3, the regenerated PILs displayed a level of variability in readability scores, despite maintaining an identical prompt. Consequently, this is a factor that should be carefully considered and investigated in future work.

Another noteworthy aspect was the decision to focus exclusively on BAD's melanoma *in situ* PIL for direct comparison. This choice could influence the results, as it limits the assessment to just one of BAD's melanoma PILs. By not evaluating the full spectrum of BAD's melanoma PILs, the comparison may have overlooked the comprehensive information BAD provides, potentially biasing the result in favour of the more simplified PILs generated by ChatGPT.

Lastly, it is crucial to recognize that the current patient demographic is highly diverse, not only in terms of literacy levels, but also in language preferences for patient information materials. When designing PILs, careful consideration must be given to how translations will be conducted to ensure that readability remains satisfactory across different languages, while preserving the accuracy of the content. Despite these limitations, the findings highlight ChatGPT's potential to play a significant role in the future development of PILs.

Ultimately, the findings from this study underscore the potential of ChatGPT in the development of PILs, with some instances where it achieved improved readability compared with those currently provided by the BAD. However, despite these improvements, the readability of the generated PILs frequently did not meet the recommended standards for easy comprehension and omitted specific content sections typically included in standard BAD PILs. These results suggest that ChatGPT could serve as a valuable supplementary tool in the creation of PILs but should not fully replace conventional methods.

Funding sources

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

The full data set, including all the ChatGPT-generated PILs, will be made available upon reasonable request to the corresponding author.

Ethics statement

Not applicable.

Patient consent

Written patient consent for publication was obtained.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website.

References

1 Rowlands G, Protheroe J, Winkley J *et al.* A mismatch between population health literacy and the complexity of health information: an observational study. *Br J Gen Pract* 2015; **65**:e379–e386.

2 Gaston CM, Mitchell G. Information giving and decision-making in patients with advanced cancer: a systematic review. *Soc Sci Med* 2005; **61**:2252–64.

3 Kitching JB. Patient information leaflets—the state of the art. *J R Soc Med* 1990; **83**:298–300.

4 DermNet. Principles of dermatological practice: an overview of dermatology. 2008. Available at: https://dermnetnz.org/cme/principles/an-overview-of-dermatology#:~:text=Although%20relatively%20straightforward%20to%20examine,skin%20diseases%20and%20many%20variants (last accessed 11 March 2024).

5 British Association of Dermatologists. Patient information leaflets. Available at: https://www.bad.org.uk/patient-information-leaflets (last accessed 20 March 2024).

6 British Association of Dermatologists. Patient and public involvement. Available at: https://www.bad.org.uk/clinical-services/service-guidance/patient-and-public-involvement/ (last accessed 11 March 2024).

7 Lampert A, Wien K, Haefeli WE *et al.* Guidance on how to achieve comprehensible patient information leaflets in four steps. *Int J Qual Health Care* 2016; **28**:634–8.

8 Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017; **69S**:S36–S40.

9 OpenAI. Introducing ChatGPT. 2022. Available at: https://openai.com/blog/chatgpt (last accessed 11 March 2024).

10 OpenAI. ChatGPT plugins. 2023. Available at: https://openai.com/blog/chatgpt-plugins (last accessed 11 March 2024).

11 Dash D, Thapa R, Banda JM *et al.* Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. *arXiv* 2023; https://doi.org/10.48550/arXiv.2304.1374 (preprint).

12 Verran C. Artificial intelligence-generated patient information leaflets: a comparison of contents according to British Association of Dermatologists standards. *Clin Exp Dermatol* 2024; **49**:711–4.

13 Richard MA, Paul C, Nijsten T *et al.* Prevalence of most common skin diseases in Europe: a population-based study. *J Eur Acad Dermatol Venereol* 2022; **36**:1088–96.

14 Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023; **25**:e50638.

15 Eid K, Eid A, Wang D *et al.* Optimizing ophthalmology patient education via ChatBot-generated materials: readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery patient brochures. *Ophthalmic Plast Reconstr Surg* 2024; **40**:212–6.

16 Wang L, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Res Social Adm Pharm* 2013; **9**:503–16.

17 Flesch R. A new readability yardstick. *J Appl Psychol* 1948; **32**:221–33.

18 Kincaid JP, Fishburne RP, Rogers RL *et al.* Derivation of new readability formulas (automated readability index, fog count and Flesch Reading Ease Formula) for navy enlisted personnel. Institute for Simulation and Training, 1975. Available at: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary (last accessed 22 February 2024).

19 Mc Laughlin GH. SMOG grading – a new readability formula. *J Read* 1969; **12**:639–46.

20 NHS Health Education England. Improving health literacy. Available at: https://www.hee.nhs.uk/our-work/knowledge-library-services/improving-health-literacy (last accessed 16 February 2024).

21 Onder CE, Koc G, Gokbulut P *et al.* Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep* 2024; **14**:243.

22 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023; **6**:1169595.

23 Brin D, Sorin V, Vaid A *et al.* Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep* 2023; **13**:16492.

24 Moshirfar M, Altaf AW, Stoakes IM *et al.* Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus* 2023; **15**:e40822.

25 Currie G, Robbie S, Tually P. ChatGPT and patient information in nuclear medicine: GPT-3.5 versus GPT-4. *J Nucl Med Technol* 2023; **51**:307–13.

26 O'Connor C, Rafferty S, Murphy M. A qualitative review of misinformation and conspiracy theories in skin cancer. *Clin Exp Dermatol* 2022; **47**:1848–52.

27 Hunt WTN, Sofela J, Mohd Mustapa MF; British Association of Dermatologists' Clinical Standards Unit. Readability assessment of the British Association of Dermatologists' patient information leaflets. *Clin Exp Dermatol* 2022; **47**:684–91.

28 Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? *Clin Orthop Relat Res* 2023; **481**:2260–7.

29 Ferreira AL, Chu B, Grant-Kels JM *et al.* Evaluation of ChatGPT dermatology responses to common patient queries. *JMIR Dermatol* 2023; **6**:e49280.

30 Momenaei B, Wakabayashi T, Shahlaee A *et al.* Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina* 2023; **7**:862–8.

31 Haidar O, Jaques A, McCaughran PW *et al.* AI-generated information for vascular patients: assessing the standard of procedure-specific information provided by the ChatGPT AI-language model. *Cureus* 2023; **15**:e49764.

32 Teasdale E, Muller I, Abdullah Sani A *et al*. Views and experiences of seeking information and help for vitiligo: a qualitative study of written accounts. *BMJ Open* 2018; **8**:e018652.

33 Rello L, Baeza-Yates R. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '13)*. New York, NY: Association for Computing Machinery, 2013; Article 14.

34 Mac O, Ayre J, Bell K *et al*. Comparison of readability scores for written health information across formulas using automated vs. manual measures. *JAMA Netw Open* 2022; **5**:e2246051.