doi:10.36849/JDD.9173

BRIEF COMMUNICATION

Journal of Drugs in Dermatology

ChatGPT Improves Readability of Patient Education Materials for Hidradenitis Suppurativa and Psoriasis

Sarah Romanelli BS, Marley Cutrona BS, Kateryna Karpoff BS, Alice B. Gottlieb MD PhD

Department of Dermatology, Icahn School of Medicine at Mount Sinai, New York, NY

INTRODUCTION

Patients increasingly rely on the Internet for health information, medical advice, telemedicine, and treatment. Given the vast and ever-expanding repository of online patient education materials (PEMs), ongoing oversight is essential, as misinformation or misinterpretation can have serious consequences.[1]

Health literacy plays a critical role in patient outcomes, with lower literacy levels consistently associated with poorer health outcomes. The National Institutes of Health (NIH) and Centers for Disease Control and Prevention (CDC) recommend that patient education materials (PEMs) be written at or below an eighth-grade reading level.[2,3] Despite these guidelines, many existing PEMs remain inaccessible to the average reader, often due to complex medical terminology, dense formatting, and other readability barriers.[4,5]

In this study, we assessed the ability of generative artificial intelligence (AI) to revise commonly available online materials for psoriasis (PsO) and hidradenitis suppurativa (HS), 2 prevalent dermatologic conditions, so they align with the recommended reading level.

MATERIALS AND METHODS

PEMs for HS and PsO were collected from the top 3 hospitals in each northeastern US state, based on the 2024 to 2025 US News and World Report Best Hospitals regional rankings. The northeast region was defined to include Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. In cases where multiple hospitals were tied for a top-three ranking, all tied institutions were included. Tied hospitals were considered a single entity, and the next highest-ranked hospitals were also included, even if they were not technically in the top three positions. For example, in New York, four hospitals were tied for first place, and all were included, along with additional hospitals ranked lower to maintain the top-three representation after accounting for ties.

PEMs were identified by searching each hospital's website for information related to HS and PsO. If no relevant PEMs were available, the hospital was excluded. The PEMs were converted to a text-only format, excluding any audiovisual multimedia. Readability was calculated using validated readability formulates: Flesch-Kincaid Reading Ease Score, Gunning Fog, Flesch-Kincaid Grade Level, Coleman-Liau Index, Simple Measure of Gobbledygook (SMOG) Index, Automated Readability Index, and Linsear Write Formula.[6]

ChatGPT-4 was used to rewrite the PEMs. The original texts were uploaded, and the following modifications were applied: (1) limit the total number of polysyllabic words to less than 30, (2) limit sentences to less than 10 words, (3) limit paragraphs to less than 5 sentences, (4) eliminate as much medical jargon without compromising accuracy, (5) when eliminating medical jargon is not possible, provide a brief explanation of the relevant concept, and (6) overall, rewrite this as if you were speaking to an eighth grader. These parameters were based on recommendations by the NIH and CDC.[2,3,7] Two independent authors reviewed all modified PEMs produced by ChatGPT to ensure content validity.

The modified PEMs were then reassessed using the aforementioned readability metrics. The average scores for each readability tool were correlated with grade level to determine the mean change in readability before and after ChatGPT modification, and paired t-tests evaluated the significance of this change.

RESULTS

At baseline, the majority of PEMs for both PsO and HS exceeded the recommended eighth-grade reading level across nearly all assessed readability metrics. For PsO, average readability scores ranged from 6.21 to 11.25, with most metrics indicating reading levels above the eighth-grade threshold. Notably, only the SMOG Index and Linsear Write Grade Level Formula yielded averages at or below the recommended level (Table 1). For HS, average baseline readability scores ranged from 7.82 to 12.58. Nearly all readability metrics exceeded the recommended level, except the Linsear Write Grade Level Formula (Table 2). Average and standard deviation statistics for each readability metric are summarized.

TABLE 1.

Psoriasis Patient Education Material Readability Scores Before ChatGPT Correction by Institution. Average and standard deviation are provided for each readability index.

Institution	Automated Readability Index	Gunning Fog Readability	Flesch-Kincaid Grade Level	Coleman-Liau Index	Simple Measure of Gobbledygook (SMOG) Index	Linsear Write Grade Level Formula	FORCAST Readability Formula	Word Count
Mount Sinai	8.28	9.9	7.87	11.12	7.58	4.93	11.16	1,113
Cornell/NYP/Columbia	8.08	7.3	6.36	10.63	6.11	4.6	10.56	243
Northwell Health	12.76	12.3	12.16	12.48	10.94	11.96	11.46	446
NYU	13.14	13.4	12.36	13.22	11.34	12.58	11.36	3,299
Montefiore	8.30	7.6	6.62	10.88	6.32	4.74	10.67	246
Hospitals of the University of Pennsylvania-Penn Presbyterian	8.08	9.7	7.69	10.91	7.46	4.81	11.12	1,116
Thomas Jefferson University Hospitals-Jefferson Health	11.77	12	10.54	13.07	9.85	10.43	11.39	237
UPMC Presbyterian	8.59	10.1	8.78	9.71	8.84	8.04	10.15	341
Penn State	8.13	10.3	7.96	11.8	7.05	3.81	11.85	219
Brigham and Women's Hospital	6.36	8.2	6.11	9.05	6.39	3.98	10.59	1,331
Massachusetts General Hospital	6.36	8.2	6.11	9.05	6.39	3.98	10.59	1,331
Beth Israel	7.96	8.4	7.11	10.48	7.01	5.3	10.61	163
Tufts Medical Center	10.66	14.1	10.6	14.11	9.69	6.68	12.11	232
Dartmouth	8.3	7.6	6.62	10.88	6.32	4.74	10.67	246
Yale Medicine	10.82	12.1	10.45	12.45	9.85	8.35	11.35	1,419
Hartford Hospital	8.3	7.6	6.62	10.88	6.32	4.74	10.67	246
Maine Medical Center	7.47	11	8.22	10.93	7.02	3.38	11.47	232
University of Vermont Medical Center	8.3	7.6	6.62	10.88	6.32	4.74	10.67	246
Average	**8.98**	**9.86**	**8.27**	**11.25**	7.82	6.21	**11.03**	705.89
Standard Deviation	1.99	2.20	2.08	1.38	1.77	2.86	0.52	797.42

TABLE 2.

Hidradenitis Suppurativa Patient Education Material Readability Scores Before ChatGPT Correction by Institution. Average and standard deviation are provided for each readability index.

Institution	Automated Readability Index	Gunning Fog Readability	Flesch-Kincaid Grade Level	Coleman-Liau Index	Simple Measure of Gobbledygook (SMOG) Index	Linsear Write Grade Level Formula	FORCAST Readability Formula	Word Count
Mount Sinai	15.53	14.6	14.63	14.89	12.85	14.73	11.48	465
Cornell/NYP/Columbia	6.9	8	6.53	8.89	6.88	5.54	9.77	510
Northwell Health	11.58	12.1	12.01	15.3	8.77	5.62	12.87	164
Montefiore	12.53	14.8	12.61	16.45	10.49	7.53	13.04	166
Thomas Jefferson University Hospitals-Jefferson Health	11.48	14.5	11.26	13.86	10.67	8.57	11.87	273
UPMC Presbyterian	9.65	10	8.88	11.07	8.44	7.58	10.57	358
Penn State	10.65	11	9.58	12.15	8.89	7.98	11.45	328
Brigham and Women's Hospital	6.41	8.3	6.4	9.06	6.6	4.48	10.3	1,059
Beth Israel	13.28	11.7	11.24	13.16	10.13	11.88	10.97	442
Tufts Medical Center	8.8	11.7	9.2	12.54	7.45	3.69	11.73	245
Dartmouth	10.61	9	8.84	11.01	8.29	8.5	11.25	48
Yale Medicine	17.23	17.3	15.83	19.86	13.22	12.13	13.5	60
Hartford Hospital	6.97	8.1	6.59	8.97	6.94	5.59	9.92	512
University of Vermont Medical Center	6.97	8.1	6.59	8.97	6.94	5.59	9.92	512
Average	**10.61**	**11.37**	**10.01**	**12.58**	**9.04**	7.82	**11.33**	367.29
Standard Deviation	3.31	3.01	3.05	3.28	2.16	3.19	1.20	257.13

FIGURE 1. ChatGPT effectively reduces the overall readability (grade-level) of search result outputs for psoriasis and hidradenitis suppurativa. Overall readability was calculated based on the average scores of seven validated readability tools. Solid lines represent the original (pre-ChatGPT) average grade levels for PEMs from each institution, while dashed lines represent the same outputs after ChatGPT modification.

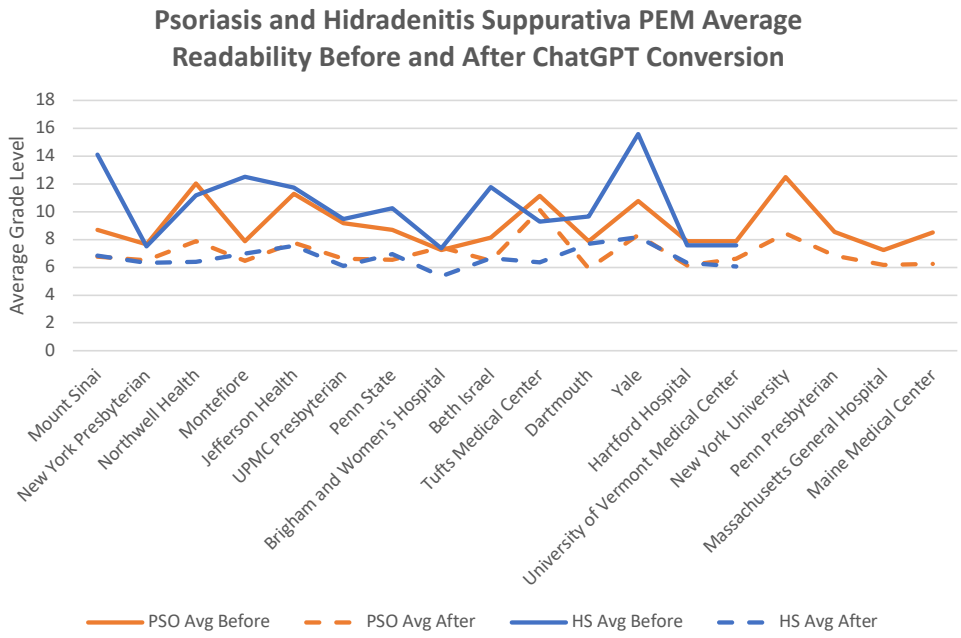

TABLE 3.

Difference in Readability Following ChatGPT Correction of Psoriasis and Hidradenitis Suppurativa Patient Education Material. *P*-values compare literacy scores before and after ChatGPT correction.

Institution	Automated Readability Index	Gunning Fog Readability	Flesch-Kincaid Grade Level	Coleman-Liau Index	Simple Measure of Gobbledygook (SMOG) Index	Linsear Write Grade Level Formula	FORCAST Readability Formula	Word Count
Psoriasis (Mean Reduction +/- STDEV)	2.30+/-1.37	2.29+/-1.27	2.44+/-1.34	1.98+/-0.99	1.93+/-1.19	2.54+/-2.22	0.46+/-0.54	404.61+/-663.21
Psoriasis (*P*-value)	1.61E-06	6.71E-07	5.88E-07	1.69E-07	2.64E-06	0.000152	0.00220	0.0191
Hidradenitis Suppurativa (Mean Reduction +/- STDEV)	4.37+/-2.80	4.56+/-2.51	4.50+/-2.40	3.97+/-2.65	3.43+/-1.79	3.93+/-2.87	1.17+/-0.93	145.43+/-180.07
Hidradenitis Suppurativa (*P*-value)	5.76E-05	1.24E-05	8.89E-06	8.62E-05	7.13E-06	0.000193	0.000402	0.00982

To generate an overall readability score, we averaged the values from all 7 readability metrics assessed per institution, both before and after ChatGPT modification (Figure 1). Significant reductions in readability scores were observed across all institution for both conditions following ChatGPT modification. The average reduction in grade-level readability scores following AI modification ranged from 0.46 to 2.54 and 1.17 to 4.56 for PsO and HS, respectively. Reductions were statistically significant (P<0.01) for most comparisons (Table 3).

To further evaluate the consistency of ChatGPT's rewriting performance, we analyzed institutions that presented identical PEMs on their websites. NewYork-Presbyterian, Dartmouth, Hartford Hospital, and University of Vermont Medical Center were found to have shared PEMs for PsO, while NewYork-Presbyterian, Hartford Hospital, and University of Vermont Medical Center shared PEMs for HS. ANOVA testing revealed that, following ChatGPT modification, these institutions had statistically indistinguishable readability scores for both conditions (PsO *P*=0.99999905; HS *P*=0.99950239).

DISCUSSION

This multi-institutional analysis of PEMs for PsO and HS found that nearly all original PEMs exceeded the recommended eighth-grade reading level, potentially limiting patient accessibility. These findings underscore the importance of allocating time during clinical visits for patient education and clarification, given that most readily accessible online materials exceed the average patient's comprehension level.

Following ChatGPT-4 modification of PEMs, significant reductions in readability scores across all readability metrics assessed were observed. Importantly, ChatGPT-4 achieved these improvements without sacrificing content accuracy, as confirmed by dual author review. These results reinforce the potential for generative AI to serve as an effective tool in creating

health-literate educational content, particularly in complex or jargon-heavy specialties such as dermatology.

The magnitude of improvement was more pronounced for HS compared to PsO, likely reflecting the initially higher baseline complexity of HS-related content. This suggests that ChatGPT may be especially useful for conditions with less well-known or more medically complex terminology. Furthermore, our analysis of institutions hosting identical PEMs demonstrated high consistency in post-ChatGPT readability outcomes. ANOVA testing confirmed statistically indistinguishable grade-level scores across these sites, underscoring the reproducibility of AI-generated outputs when applied to standardized content.

Our analysis was limited to text-based PEMs from top-ranked northeastern hospitals and may not capture national variability in content or accessibility. Additionally, while readability scores offer a validated measure of textual complexity, they do not fully assess patient comprehension, cultural appropriateness, or engagement. Overall, our findings suggest that AI-driven tools such as ChatGPT have the potential to enhance health literacy by making complex medical information more readable to patients.

DISCLOSURES

SR, MC, and KK have no conflicts of interest to disclose. ABG has received honoraria as an advisory board member and consultant for Amgen, Eli Lilly, HighlightsTherapeutics, Janssen, Novartis, Sanofi, SunPharma, Takeda, Teva, UCB, and Xbiotech (stock options for RA); and research/educational grants from Avalo Therapeutics, Bristol-Myers Squibb, Janssen, Moonlake, and UCB Pharma (all paid to Mount Sinai School of Medicine).

Funding: International Dermatology Outcome Measures (IDEOM) nonprofit organization.

REFERENCES

1. Borges do Nascimento IJ, Pizarro AB, et al. Infodemics and health misinformation: a systematic review of reviews. *Bull World Health Organ.* 2022;100(9):544-561.
2. Clear & simple. (2021). Accessed: October 8, 2024: https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple?utm_medium=email&utm_source=transaction
3. Centers for Disease Control and Prevention. Simply put; a guide for creating easy-to-understand materials. Accessed April 5, 2020. https://www.cdc.gov/healthliteracy/pdf/simply_put.pdf
4. Zirwas MJ, Holder JL. Patient education strategies in dermatology—Part 2: Methods. *J Clin Aesthetic Dermatol.* 2009;2(12):28–34.
5. Tulbert BH, Snyder CW, Brodell RT. Readability of Patient-oriented Online Dermatology Resources. *J Clin Aesthet Dermatol.* 2011;4(3):27-33.
6. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Educ Behav.* 2006;33(3):352-373. doi:10.1177/1090198105277329
7. Miskiewicz M, Capotosto S, Wang ED. Evaluation of readability of patient education materials on lateral epicondylitis (tennis elbow) from the top 25 orthopedic institutions. *JSES Int.* 2023;7(5):877-880. doi:10.1016/j.jseint.2023.05.006

AUTHOR CORRESPONDENCE

Sarah Romanelli BS
E-mail:..... sarah.romanelli@mssm.edu