The Health Literacy Penalty: Clinical vs. Lay Terminology and Medical Information Quality Across Ten Diseases

Divyanshi Dubey
Divyanshi Dubey, Alpharetta High School, Alpharetta, Georgia, USA

***Abstract*—This study examined whether Google search results differ when the same medical symptom is described using clinical terminology versus everyday language. Using a paired-query design across ten diseases, we found that lay-terminology searches returned significantly lower-quality medical information, with the largest gap observed for depression a finding with direct public health implications. Millions of people search for health information online using everyday language to describe their symptoms, yet the quality of results they receive compared to searches using clinical terminology has never been systematically examined across multiple disease categories. We hypothesized that Google searches using clinical terminology would return higher-quality medical information than searches using lay terminology, and that this quality gap would vary by disease category. We constructed paired clinical and lay queries for ten diseases spanning seven clinical domains and retrieved the top ten organic Google search results for each query, yielding 200 total results. Each result was scored on source quality tier, evidence-based language density, and alarming language density. Independent samples t-tests confirmed that clinical queries returned significantly higher-quality sources and significantly more evidence-based content than lay queries ($p < 0.05$ for both). The largest gap was observed for depression, where lay searchers received results with almost no evidence-based language. These findings suggest that the language patients use when searching online systematically determines the quality of health information they receive.**

***Index Terms*—consumer health informatics, health equity, health literacy, medical terminology, search engine quality**

I. INTRODUCTION

Consumer health information seeking on the internet has become one of the most prevalent health behaviors worldwide. A significant proportion of adults search for health information online, and Google serves as a primary gateway to this information [1]. How consumers search for and appraise that information has direct implications for patient education and health decision-making [2].

A fundamental but underexplored aspect of this information landscape is the role of search terminology. Patients and caregivers rarely use clinical language when describing symptoms. A person experiencing syncope searches "keep fainting what is wrong," not "syncope causes." A person with dyspnea searches for "can't catch my breath," not "dyspnea causes treatment." Prior research has shown that medical terminology and lay terminology for the same condition are treated differently in clinical settings physicians assign more diagnostic weight to medically labeled conditions, and patients perceive conditions described in clinical terms as more serious [3].

However, whether the terminology a patient uses when searching online determines the quality of information they receive has not been studied. Prior research has examined specific dimensions of online health information quality. Studies have evaluated search result quality for individual conditions such as depression [4] and rosacea [5], and broader analyses have characterized health misinformation in online communities [6, 7]. Systematic reviews of consumer health information have established that gaps in online health information are consistent across conditions and platforms rather than random occurrences [8]. However, no prior study has used a paired-query

design comparing the same symptom expressed in both clinical and everyday language across multiple disease categories simultaneously to determine whether terminology type predicts information quality.

We hypothesized that clinical-terminology Google searches would return significantly higher-quality medical information than lay-terminology searches describing identical symptoms, and that the magnitude of this gap would vary systematically by disease category with the largest gaps observed for psychiatric and autoimmune conditions where lay descriptions diverge most substantially from clinical vocabulary. We found that clinical queries returned significantly higher source quality and evidence-based content scores, with depression showing the largest disparity of any disease examined. These findings have direct implications for health literacy research, public health communication, and the design of consumer health search interfaces.

II. RESULTS

Across all 200 search results, clinical queries returned significantly higher evidence-based language scores than lay queries (mean 2.33 vs. 1.55; t = 2.422, p = 0.017, independent-samples t-test; Figure 1). Clinical queries also returned significantly higher-tier sources than lay queries (mean source score 2.97 vs. 3.41; t = -2.318, p = 0.022; Figure 1). Unexpectedly, clinical pages contained significantly more alarming language than lay pages (mean red flag score 0.34 vs. 0.19; t = 1.988, p = 0.048). Readability grade level did not differ significantly between query types (mean 12.80 vs. 11.57; p = 0.098). These results are summarized in Table I.

Table I: Overall Results Summary

Metric	Clinical Mean	Lay Mean	t-statistic	p-value
Source score (lower=better)	2.97	3.41	-2.318	0.022
Evidence score (higher=better)	2.33	1.55	2.422	0.017
Red flag score (lower=better)	0.34	0.19	1.988	0.048
Readability grade level	12.80	11.57	1.665	0.098

Evidence quality gap varies substantially by disease

The evidence score gap between clinical and lay queries varied substantially across the ten diseases examined (Figure 2; Table II). Depression showed the largest gap (clinical mean 3.70 vs. lay mean 1.22; gap = 2.48), indicating that everyday language searches for depressive symptoms returned results with almost no evidence-based content. Hypothyroidism (gap = 1.39), dyspnea (gap = 1.20), and syncope (gap = 1.11) also showed large disparities. Fibromyalgia (gap = 0.29), lupus (gap = 0.11), and angina (gap = -1.00) showed small or negative gaps. Angina was the only disease where lay queries outperformed clinical queries in evidence score, likely reflecting the extensive public health communication investment directed at chest pain symptoms.

Table II: Evidence Score Gap by Disease

Disease	Clinical Mean	Lay Mean	Gap (Clinical-Lay)
Depression	3.70	1.22	+2.48
Hypothyroidism	2.50	1.11	+1.39
Dyspnea	2.20	1.00	+1.20
Syncope	2.78	1.67	+1.11
Seborrheic Dermatitis	2.71	1.78	+0.94
Cellulitis	1.67	1.11	+0.56
Dysphagia	1.78	1.25	+0.53
Fibromyalgia	3.00	1.71	+0.29
Lupus	1.89	1.78	+0.11
Angina	1.00	2.00	-1.00

Source quality gap mirrors evidence gap

Depression also showed the largest source quality gap across diseases (clinical mean 2.90 vs. lay mean 4.44; Figure 3), indicating that lay searches for depressive symptoms frequently returned social media and forum-based sources rather than institutional health pages. Hypothyroidism (clinical 2.75 vs. lay 3.78) and cellulitis (clinical 2.89 vs. lay 3.67) also showed notable source quality disparities. Syncope was the only disease where lay searches returned higher-tier sources than clinical searches (lay mean 2.44 vs. clinical mean 2.89), consistent with its anomalous evidence score pattern.

The red flag paradox

A counterintuitive finding emerged in the red-flag language analysis: clinical query results contained significantly more alarming language than lay query

results across most diseases (Figure 4). This pattern was most pronounced for seborrheic dermatitis (clinical 0.71 vs. lay 0.22) and fibromyalgia (clinical 0.43 vs. lay 0.29). This finding suggests that clinical and institutional health pages directed at medical or health-literate audiences employ explicit risk language as standard clinical communication, while consumer health platforms moderate their tone to avoid alarming general readers.

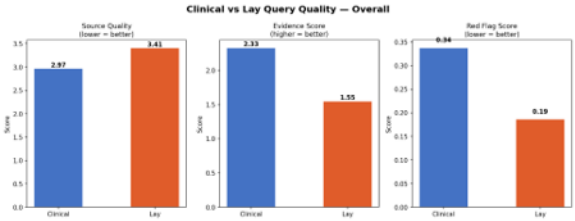

Figure 1: Overall results comparing clinical and lay query results

III. DISCUSSION

This study provides the first systematic, multi-disease, paired-query analysis of Google search result quality by medical terminology type. Our results confirmed the primary hypothesis: clinical-terminology searches returned significantly higher-quality information than lay-terminology searches. Critically, we demonstrated that this quality gap is not uniform across disease categories it is most severe for mental health conditions, and specifically for depression, where the disparity in evidence-based content was nearly three times larger than the average gap across other diseases studied.

The depression finding carries substantial public health implications. Depression affects approximately 280 million people globally and is a leading cause of disability worldwide [9]. Our data suggest that when individuals experiencing depressive symptoms search using natural language as most patients would they are systematically returned results with almost no evidence-based content, compared to the rich clinical content returned by formal terminology searches. This disparity may reinforce misconceptions about the nature of depression, delay recognition of symptoms as a treatable medical condition, and discourage professional help-seeking at the moment when accurate information is most needed. Future research should examine whether this information gap contributes to the well-documented delay between symptom onset and treatment initiation in depression [10].

The red flag paradox in which clinical queries returned more alarming language than lay queries is a novel finding that challenges the assumption that higher source quality is synonymous with more reassuring content. Rather than indicating that clinical results are more alarmist, this pattern most likely reflects a systematic difference in writing convention between clinical and consumer health communication. Institutional and clinical pages routinely include explicit risk and urgency language intended for health-literate audiences, while consumer health platforms deliberately moderate their tone to prevent unnecessary alarm in general readers. This finding suggests that source tier and evidence language density alone are insufficient proxies for overall information quality and that tone calibration represents an additional dimension of consumer health content evaluation.

The exception case of angina where lay searches outperformed clinical searches in evidence score is also informative and potentially optimistic. Chest pain is one of the most heavily resourced consumer health topics, reflecting decades of public health investment in cardiovascular disease awareness. This suggests that targeted public health communication investment can close or even reverse the terminology-driven information gap for high-priority conditions. Depression, which showed the largest gap in this study, may benefit from analogous investment in lay-language search quality.

Several limitations of this study should be acknowledged. First, the sample of ten diseases limits statistical power for per-disease comparisons, and future work should expand the disease panel to 20-25 conditions to enable more robust subgroup analyses. Second, results were collected from a single geographic location in Atlanta, Georgia, in a single session; Google search results vary by location, user history, and time, and our findings may not generalize to other geographic or temporal contexts. Third, our source quality scoring relied on automated URL pattern matching and keyword-based content scoring rather than expert evaluation; future studies should incorporate validated instruments such as the DISCERN scale for a subset of results. Fourth, this study examined only Google; comparable analyses of Bing, AI-powered search interfaces, and non-English

search contexts would substantially extend these findings.

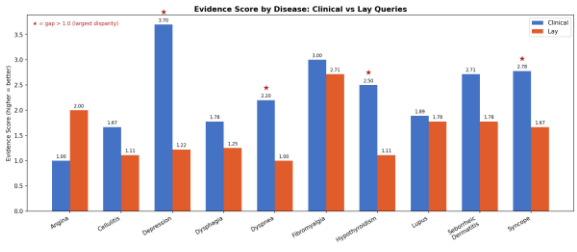

Figure 2: Evidence score gap by disease

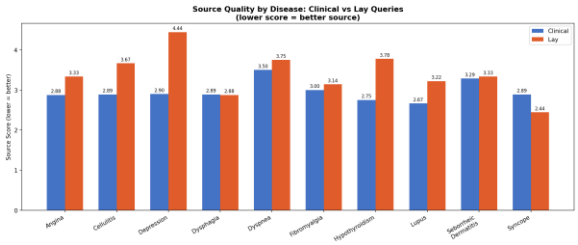

Figure 3: Source quality gap across diseases

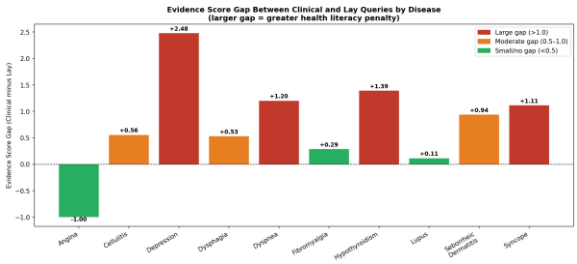

Figure 4: Red flag language density across diseases

IV. MATERIALS AND METHODS

Disease and query selection

Ten diseases were selected to represent a range of clinical domains: cardiology (angina pectoris), neurology (syncope), gastroenterology (dysphagia), psychiatry (major depressive disorder), rheumatology/autoimmune (systemic lupus erythematosus), pulmonology (dyspnea), endocrinology (hypothyroidism), rheumatology (fibromyalgia), infectious disease/dermatology (cellulitis), and dermatology (seborrheic dermatitis). For each disease, a paired query set was constructed consisting of one clinical-terminology query using standard medical vocabulary and one lay-terminology query using language representative of a patient with no medical background describing the same symptoms (Table III).

Table III: Paired Query Sets

Disease	Clinical Query	Lay Query
Angina	Angina pectoris symptoms	Chest tightness, what is it
Syncope	Syncope causes	Keep fainting what is wrong
Dysphagia	Dysphagia symptoms cause	Trouble swallowing food
Depression	Major depressive disorder symptoms	Feeling sad and empty all the time
Lupus	Systemic lupus erythematosus symptoms	Red rash on my face and joints keep hurting
Dyspnea	Dyspnea causes treatment	Can't catch my breath all the time
Hypothyroidism	Hypothyroidism symptoms	Always tired of gaining weight for no reason
Fibromyalgia	Fibromyalgia symptoms diagnosis	Pain all over my body all the time
Cellulitis	Cellulitis symptoms cause	Skin red, swollen, and hot
Seborrheic Dermatitis	Seborrheic dermatitis symptoms	Flaky, itchy scalp what is it

Data collection

Search results were retrieved programmatically using the SerpAPI Google Search Application Programming Interface (API) with a Python script (Python 3.14.5). For each of the 20 queries, the top 10 organic search results were retrieved, yielding 200 total result records. Each record included the result URL, page title, and meta-description snippet. All data were collected in a single session in May 2026 to minimize temporal and personalization variation in search results.

Source quality scoring

Each URL was classified into one of five source quality tiers based on domain type using automated pattern matching: Tier 1 (government and international health bodies: NIH, CDC, WHO,

MedlinePlus), Tier 2 (academic medical centers: Mayo Clinic, Johns Hopkins Medicine, Cleveland Clinic), Tier 3 (established consumer health media: WebMD, Healthline, Medical News Today), Tier 4 (general reference: Wikipedia, NHS), and Tier 5 (social media and forums: Reddit, Quora, Yahoo Answers). Lower-tier numbers represent higher source quality. Classification was performed using a predefined domain dictionary applied with the Python panda's library.

Content quality scoring

The full text of each URL was retrieved using the Python requests library and parsed using BeautifulSoup4. Navigation elements, scripts, and footers were removed before text extraction. Each page was scored on two content metrics. Evidence scores were calculated as the count of evidence-based language terms present in the page text, including "clinical trial," "study found," "research shows," "peer reviewed," "meta-analysis," "systematic review," "guidelines," "diagnosis," and "physician." Red flag score was calculated as the count of alarming or sensationalist language terms, including "miracle," "cure," "detox," "toxins," "guaranteed," and "ancient secret." Readability was assessed using the Flesch-Kincaid Grade Level formula via the Python textstat library.

Statistical analysis

Independent-samples t-tests were used to compare clinical and lay query groups on source score, evidence score, red flag score, and readability grade level. All statistical analyses were performed in Python using the SciPy.stats library. A significance threshold of $p < 0.05$ was applied. All data collection and analysis code is included in the Appendix.

V. CONCLUSION

Consumer health search quality is significantly lower when users employ everyday symptom language rather than clinical terminology, and this gap is most severe for mental health conditions, particularly depression. These findings suggest that public health communicators, search engine developers, and health literacy researchers should prioritize improving the quality and accessibility of results returned to lay-terminology queries the language that most patients actually use when they are most in need of accurate information.

REFERENCES

[1] Bundorf, M. K., et al., "Who searches the internet for health information?" Health Services Research, vol. 41, no. 3, pp. 819–836, 2006. doi: 10.1111/j.1475-6773.2006.00510.x.

[2] Bizzotto, N., G. J. de Bruijn, and P. J. Schulz, "Buffering against exposure to mental health misinformation in online communities on Facebook: The interplay of depression literacy and expert moderation," BMC Public Health, vol. 23, Art. no. 1577, 2023. doi: 10.1186/s12889-023-16404-1.

[3] Eysenbach, G., and C. Köhler, "How do consumers search for and appraise health information on the World Wide Web?" BMJ, vol. 324, no. 7337, pp. 573–577, 2002. doi: 10.1136/bmj.324.7337.573.

[4] Nelson, H. C., et al., "The reliability gap: How traditional search engines outperform AI chatbots in rosacea public health information quality," Cureus, vol. 17, no. 6, Art. no. e86543, 2025. doi: 10.7759/cureus.86543.

[5] Pryor, T. A. M., K. A. Reynolds, P. L. Kirby, and M. T. Bernstein, "Quality of late-life depression information on the internet: Website evaluation study," JMIR Formative Research, vol. 6, no. 9, Art. no. e36177, 2022. doi: 10.2196/36177.

[6] Sharma, K., et al., "Combating health misinformation on social media," ACM SIGKDD Explorations Newsletter, vol. 23, no. 1, pp. 19–27, 2021. doi: 10.1145/3468507.3468510.

[7] Sun, Y., Y. Zhang, J. Gwizdka, and C. B. Trace, "Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators," Journal of Medical Internet Research, vol. 21, no. 5, Art. no. e12522, 2019. doi: 10.2196/12522.

[8] Wang, P. S., P. Berglund, M. Olfson, et al., "Failure and delay in initial treatment contact after first onset of mental disorders in the National Comorbidity Survey Replication," Archives of General Psychiatry, vol. 62, no. 6, pp. 603–613, 2005. doi: 10.1001/archpsyc.62.6.603.

[9] World Health Organization, "Depression," WHO Fact Sheets. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. Accessed: May 29, 2026.

[10] Young, M. E., G. R. Norman, and K. R. Humphreys, "The role of medical language in changing public perceptions of illness," PLoS ONE, vol. 3, no. 12, Art. no. e3875, 2008. doi: 10.1371/journal.pone.0003875.