MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI251046

Development of Methods to Assess Readability in Health and Medical Information: A Scoping Review

Anunita NATTAM[a], Himaja CHINTALAPALLI[a], Hexuan LIU[a], Danny T.Y. WU[a,b,1]
[a] University of Cincinnati, Cincinnati, Ohio, United States
[b] University of North Carolina, Chapel Hill, North Carolina, United States

Abstract. This study aimed to extract characteristics of studies involving development of readability measures for health and medical information (HMI) to explore the common themes (strategies) within such developments. Four databases were searched following PRISMA guidelines. The data analysis included statistical summary, thematic analysis, and idea webbing. A total of 1,129 articles were identified in the initial search, resulting in 14 articles included. More than half of the articles (n=8) were published between 2006 and 2015, with declining recent developments. Four main themes of development strategies were revealed: Language Features (n=10), Machine Learning (ML, n=6), Natural Language Processing (NLP, n=5), and Human Annotations (n=5). ML and NLP techniques were popularly used and oftentimes tested by professional and layperson judgement. However, despite these recent developments, none have been widely adopted. With further advancement of Artificial Intelligence and Large Language Models, opportunities exist to develop usable and reliable readability measures for HMI.

Keywords. Readability Assessment, Scoping Literature Review, Language Analysis Algorithms, Thematic Analysis

1. Introduction

Health and medical information (HMI) encompasses "the knowledge, facts, and news generated from various sources, necessary for good physical and mental condition of human beings" [1]. It includes provider-patient communication in the form of clinical trial information, electronic health records (EHRs), and online resources like medical websites, blogs, forums, and support groups [2]. However, in an era where patient autonomy and education are increasingly more valuable, several studies have shown that the quality of HMI, especially information on the internet, is inadequate in terms of its readability and understandability. It is a problem that persists despite ongoing research efforts [3,4].

Traditional readability measures (TRMs), such as Flesch-Kincaid Grade Level (FKGL), Gunning-Fog Index (GFI), and Simple Measure of Gobbledygook Index (SMOG), have been used to assess the readability of HMI based on surface-level characteristics such as sentence length and syllable count [5,6]. Although widely used (e.g., implemented in Microsoft Word), TRMs may be inadequate for HMI, as they are

[1] Corresponding Author: Danny T.Y. Wu, wutz@ucmail.uc.edu or wuty@unc.edu

often tricked by word complexity and failed to reflect the understandability and can produce misleading scores due to their focus on lexicology rather than semantics [7–9]. For example, the word "anesthesia" may appear less readable due to its syllable count, despite being familiar to the lay audience.

Developing proper readability measures (RMs) for HMI would ultimately help align the readability of the text with the audiences' health literacy, improving their ability to make informed decisions using HMI. Health literacy can be seen as a modifiable social determinant of health in our previous systematic review, [10]where interventions were recommended to focus on text simplification, cross-cultural communication, health literacy screening, and education programs to improve the comprehension and navigation of the healthcare system. Therefore, the present study aims to conduct a scoping review to summarize the literature with a focus on the recent developments of RMs for HMI. Specifically, this study aims to gain a summative understanding of the status and the design strategies of these RMs as an introduction to outlining potential applications and recommendations to improve patient readability.

2. Methods

This scoping review systematically examined the literature by adhering to the Arksey and O'Malley framework and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for scoping reviews (PRISMA-ScR) guidelines [12,13]. PubMed, ACM Digital Library, IEEE Xplore, and SCOPUS were searched using the keywords listed in the **Supplemental Table 1** uploaded on Zenodo [14]. Two reviewers screened the resulting articles based on inclusion and exclusion criteria (**Supplemental Table 2**) [14], with disagreements being resolved by a third reviewer. The inter-rater reliability for screening was calculated using the Prevalence Adjusted Bias Adjusted Kappa (PABAK) due to the high rate of excluded papers. References of included papers followed the same screening and eligibility determination process. A PRISMA flow diagram was generated to summarize the search process. An online Excel spreadsheet in Microsoft OneDrive was used to screen and store paper information.

Team members extracted information following the PRISMA guidelines. Key data from each paper, including publication year, patient population, PEM/validation measure, database, and algorithm changes, was extracted (**Supplemental Table 4**) [14]. To gain a summative understanding of the status and design strategies, an idea web was created using thematic analysis. The same reviewers individually identified themes within each article and created a list of primary and secondary themes. The idea web, centered on "Readability Measures," branched into themes with articles color-coded as primary or secondary. This provided an understanding of current RMs.

3. Results

3.1. Characteristics of Included Articles

The characteristics of the 14 included studies are summarized in the **Supplemental Table 3** [14]. The majority (57.2%) were published between 2011 and 2020, with only two (13.33%) from 2021-2023, despite advancements in NLP and ML. The rest (28.6%) were from before 2010. Most studies targeted general laypersons (n=9), with others

focusing on specific groups like internet users (n=4) and clinical trial readers (n=1). Common health information types included EHR (n=5), scientific articles (n=4), and Wikipedia (n=4). Less common sources were clinical trial information (n=3), blogs (n=1), patient education materials (n=3), and news articles (n=3). Predominantly, studies were from PubMed (n=7), with technical databases such as IEEE Xplore (n=5) and ACM Digital (n=2) also contributing, highlighting the interdisciplinary nature of readability assessments.

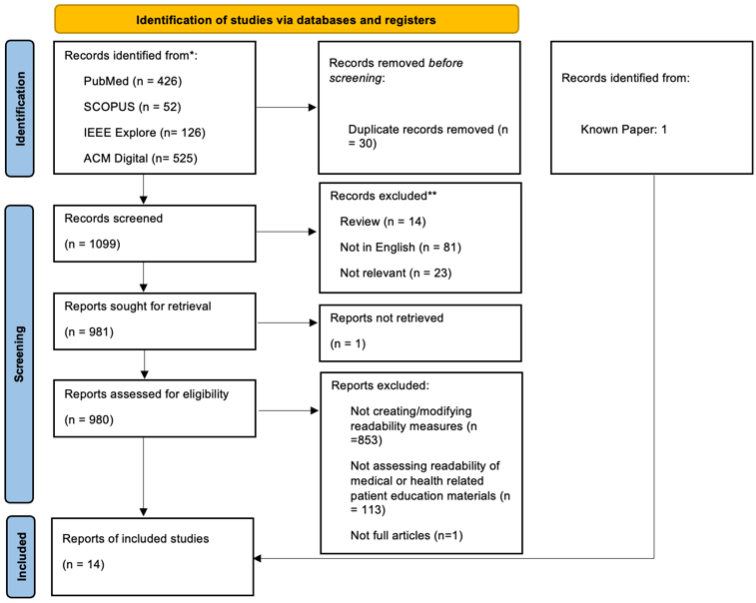

Figure 1: PRISMA Flowchart identifying the total number of papers included and excluded throughout the screening process.

3.2. Thematic Analysis / Idea Webbing

Theme 1: Language Analysis

Ten papers (#1, 2, 4-6, 8-11, and 13 in **Supplemental Table 4**) [14] assessed readability through syntactic, semantic, and linguistic features, including term familiarity and grammatical characteristics. External dictionaries, such as the Open Access and Collaborative Consumer Health Vocabulary (CHV) and Unified Medical Language System (UMLS), were often used to enhance these analyses.

Theme 2: Natural Language Processing (NLP)

Five papers (#1, 8, and 12-14) employed NLP techniques like text tokenization, sentence splitting, and part-of-speech tagging. Some studies developed metrics such as noun phrase complexity and semantic familiarity (#13), while others used information retrieval metrics or advanced NLP pipelines (#1 and 14).

Theme 3: Machine Learning (ML)

Four papers (#2, 3, 7, and 9) applied ML models to readability assessment, utilizing algorithms such as SVM, KNN, Naïve Bayes, and ANN. Supervised learning was

common, with models trained on labeled data and evaluated using metrics like F-measure, accuracy, precision, and recall. Feature engineering involved extracting linguistic, syntactic, and lexical features from texts. Comparisons of various ML algorithms aimed to find the most effective approach, with some studies including traditional readability metrics (#2 and 7).

Theme 4: Human Annotations

Five studies used human annotations to evaluate algorithm performance (#3, 7, 8, 11 and 14). Annotators included experts and laypersons. Two papers (#8 and 14) used only expert raters, while two (#3 and 7) used layperson annotators. One study (#11) employed both. Expert annotators rated document comprehensibility and found correlations with layperson ratings, though discrepancies with traditional readability metrics were noted. The study using both types of annotators revealed that laypersons often perceived difficulty more intensely for others than for themselves.

4. Discussion and Conclusions

This scoping review identified 14 articles and summarized their characteristics. Emphasis was placed on linguistic analysis, integrating techniques such as ML and NLP, and human annotation. Some studies employed multiple approaches in their development. These findings could guide the development of novel RMs suitable for HMI. Meanwhile, recent studies are increasingly focusing on the understandability of HMI. Effective HMI should enhance health literacy, enabling individuals to interpret and use health information more effectively. With the growing availability of HMI, it's essential to ensure their understandability to laypersons to improve self-management, treatment compliance, and the clinician-patient relationship [15], which may lead to better outcomes. Recent advances in Artificial Intelligence and Large Language Models (LLMs) present opportunities for more accurate readability assessments and automated text simplification.

This study has a few limitations. First, it covered only four databases (PubMed, Scopus, ACM Digital, IEEE Explore) and did not include additional relevant sources found through Google Scholar. The absence of a guiding framework for thematic analysis was mitigated by team collaboration. Second, the study focused solely on articles creating new readability measures, excluding those validating existing metrics or using non-health-related materials, which might have provided additional insights. Future research will concentrate on developing and disseminating reliable readability assessment algorithms, incorporating LLMs. Current work includes evaluating readability of summaries from ClinicalTrial.gov using ChatGPT [16] and a fine-tuned BART-Large-CNN model. Additionally, creating repositories of HMI with human annotations will aid in developing effective measures.

This scoping review outlined the characteristics and implementation strategies of existing readability assessment methods. The identified themes can inform further comparative studies on readability methods to provide recommendations for the creation of next-generation RMs and associated applications, aiming at improving health literacy and supporting better patient outcomes.

References

[1]. O.E. Nwafor-Orizu, and O.T.U. Onwudinjo, Availability and Use of Health Information Resources by Doctors in Teaching Hospitals in South East Nigeria, **5** (2015).

[2]. G. Eysenbach, Infodemiology: the epidemiology of (mis)information, *Am. J. Med.* **113** (2002) 763–765. doi:10.1016/S0002-9343(02)01473-0.

[3]. J.Y. Tan, Y.C. Tan, and D. Yap, Readability and quality of online patient health information on parotidectomy, *J. Laryngol. Otol.* **137** (2023) 1378–1383. doi:10.1017/S0022215123000336.

[4]. L. Boutemen, and A.N. Miller, Readability of publicly available mental health information: A systematic review, *Patient Educ. Couns.* **111** (2023) 107682. doi:10.1016/j.pec.2023.107682.

[5]. U.C. Bureau, Educational Attainment in the United States: 2018, *Census.Gov*. (n.d.). https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html (accessed February 19, 2024).

[6]. R. Flesch, A new readability yardstick., *J. Appl. Psychol.* **32** (1948) 221–233. doi:10.1037/h0057532.

[7]. Barry D. Weiss, Health Literacy: A Manual For Clinicians, (2006). http://lib.ncfh.org/pdfs/6617.pdf.

[8]. A. Nattam, T. Vithala, T.-C. Wu, S. Bindhu, G. Bond, H. Liu, A. Thompson, and D.T.Y. Wu, Assessing the Readability of Online Patient Education Materials in Obstetrics and Gynecology Using Traditional Measures: Comparative Analysis and Limitations, *J. Med. Internet Res.* **25** (2023) e46346. doi:10.2196/46346.

[9]. H. Kim, S. Goryachev, G. Rosemblat, A. Browne, A. Keselman, and Q. Zeng-Treitler, Beyond surface characteristics: a new health text-specific readability measurement, *AMIA Annu. Symp. Proc. AMIA Symp.* **2007** (2007) 418–422.

[10]. S. Bindhu, A. Nattam, C. Xu, T. Vithala, T. Grant, J.K. Dariotis, H. Liu, and D.T.Y. Wu, Roles of Health Literacy in Relation to Social Determinants of Health and Recommendations for Informatics-Based Interventions: Systematic Review, *Online J. Public Health Inform.* **16** (2024) e50898. doi:10.2196/50898.

[11]. G.H. Mc Laughlin, SMOG Grading - a New Readability Formula, *J. Read.* **12** (1969) 639–646.

[12]. H. Arksey, and L. O'Malley, Scoping studies: towards a methodological framework, *Int. J. Soc. Res. Methodol.* **8** (2005) 19–32. doi:10.1080/1364557032000119616.

[13]. A.C. Tricco, E. Lillie, W. Zarin, K.K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M.D.J. Peters, T. Horsley, L. Weeks, S. Hempel, E.A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M.G. Wilson, C. Garritty, S. Lewin, C.M. Godfrey, M.T. Macdonald, E.V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, and S.E. Straus, PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation, *Ann. Intern. Med.* **169** (2018) 467–473. doi:10.7326/M18-0850.

[14]. A. Nattam, H. Chintalapalli, H. Liu, and D.T.Y. Wu, Supplemental Information: Developing Readability Measures for Health and Medical Information – A Scoping Review, (2024). doi:10.5281/ZENODO.14019625.

[15]. D. Kauchak, and G. Leroy, Moving Beyond Readability Metrics for Health-Related Text Simplification, *IT Prof.* **18** (2016) 45–51. doi:10.1109/MITP.2016.50.

[16]. ChatGPT, (n.d.). https://chatgpt.com/ (accessed October 18, 2024)