The Evaluation of Medical Terms Complexity Using Lexical Features and Large Language Models

Liliya Makhmutova
TU Dublin, Ireland
D21124385@mytudublin.ie

Giancarlo Dondoni Salton
Federal University of the Southern Border
Campus Chapecó, Brazil
gian@uffs.edu.br

Fernando Perez-Tellez
TU Dublin, Ireland
Fernando.PerezTellez@TUDublin.ie

Robert J. Ross
TU Dublin, Ireland
robert.ross@TUDublin.ie

Abstract

Understanding medical terminology is critical for effective patient-doctor communication, yet many patients struggle with complex jargon. This study compares Machine Learning (ML) models and Large Language Models (LLMs) in predicting medical term complexity as a means of improving doctor-patient communication. Using survey data from 252 participants rating 1,000 words along with various lexical features, we measured the accuracy of both model types. The results show that LLMs outperform traditional lexical-feature-based models, suggesting their potential to identify complex medical terms and lay the groundwork for personalised patient-doctor communication[1].

1 Introduction

Laypeople often struggle with medical terminology due to factors such as limited health literacy, language barrier, or personal problems. This limited understanding of medical terminology poses significant challenges in healthcare systems. Poor comprehension of health records, medication instructions, or public education materials can lead to treatment non-adherence, miscommunication during clinical consultations, and even avoidable hospitalisations. Studies show that many individuals have only basic health literacy, and some conceal poor reading skills, leading doctors to overestimate their understanding (Weiss, 2007). On the other hand, LeBlanc et al. (2014) found that soon-to-be physicians underestimated patients' medical text comprehension abilities. Even patient information leaflets may not be written in accessible language, further compounding these issues for vulnerable groups such as Medicaid recipients (O'Sullivan et al., 2020). Non-native speakers may also lack relevant medical vocabulary (Al Shamsi et al., 2020). Others may disengage from healthcare-related information due to low interest (McCarron et al., 2019), (dis)trust in the healthcare system (Tsai et al., 2018; Tunç et al., 2025), inaccessible health records (van Mens et al., 2020; Neves et al., 2020), or personal challenges (Murugesu et al., 2022). Addressing these gaps is critical for designing clearer Electronic Health Record (EHR) interfaces, patient information materials, and AI-driven tools that minimise jargon and improve accessibility of medical records.

Each person comes with a unique background and experience, which influences their understanding of medical terms. As a result, people can misinterpret medical terminology by relying on their existing knowledge. Notable examples of such misunderstandings were documented by Gotlieb et al. (2022), including cases where individuals misinterpreted "impressive x-ray findings" or "progressing tumour" as positive news due to the colloquial meanings of these terms. Another example is the word "fertilisation" which in medical contexts refers to "the joining of sperm and egg" while in agricultural contexts it suggests "(soil) enrichment" (both derived from the Latin word "fer" meaning "to bear, carry"). Language background can also lead to confusion. For instance, "angina" (short for "angina pectoris" – "chest pain") is commonly used to refer to *tonsillitis* in many languages.

Several studies focus on general medical/biological terms (Shardlow et al., 2020; Grabar et al., 2014), while others examine specific terminology (e.g., X-rays, diseases) (Sugihara et al., 2024; Pieterse et al., 2012; Zimmermann et al., 2021; Lalor et al., 2024). Most research exploring lexical understandability measures term recognition rather than comprehension of meaning. Few studies investigate medical term complexity

[1]The supplementary material will be available on https://huggingface.co/liliya-makhmutova and https://github.com/LiliyaMakhmutova

Proceedings of Recent Advances in Natural Language Processing,pages 682–693
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-079

with a personalization component (Tran et al., 2025; Asthana et al., 2024). Furthermore, some publications analyse this topic using limited sets of terms (Gotlieb et al., 2022; O'Connell et al., 2013; Lerner et al., 2000; Briem et al., 2004).

In this paper, we explore features and models for predicting the complexity of medical terms. This study may be helpful for clinicians in identifying potentially ill-understood jargon that could confuse patients. Specifically, our goal is to understand which fairly common medical words or phrases (jargon) are unknown or misunderstood by laypeople. Another key question addressed is how different lexical characteristics of words influence their comprehension, as well as the use of LLMs to evaluate medical term complexity. Our key contributions are: (1) A curated dataset of 1,000 medical terms, annotated with familiarity and complexity scores based on responses from 252 participants; (2) An evaluation of traditional machine learning models versus the use of state-of-the-art LLMs to predict term complexity, along with an analysis of informative features; and (3) A demonstration that LLMs slightly outperform traditional models, achieving the lowest prediction errors (RMSE: 0.21).

2 Related Work

Early research by Zeng et al. (2005) and Kauchak and Leroy (2016) highlights the limitations of conventional readability measures, such as syllable count, word length, or easy-word lists, in assessing the complexity of medical terms, arguing that these metrics often do not predict true comprehension barriers. Zeng et al. (2005) pioneered a personalised approach, modelling term familiarity based on individual factors (e.g., education, age, native language status) and word-level characteristics (e.g., corpus frequency, Dale-Chall easy-word percentage), finding education to be the strongest predictor. Similarly, Kauchak and Leroy (2016) showed that while word length correlates with perceived difficulty, term frequency better predicts actual understanding, although their forced-choice survey method may inflate guessing since false definitions are randomly selected. Likewise, Jiang and Xu (2024) demonstrate that general readability metrics (e.g., FKGL, ARI) perform poorly on medical texts unless augmented with jargon term counts, reinforcing the need for an alternative approach.

Several studies identify term frequency as a reliable complexity indicator. Cherednichenko et al. (2018) use frequency-based filtering for Ukrainian medical terms, while Pylieva et al. (2019) and Sugihara et al. (2024) integrate frequency features from medical corpora and Wikipedia into machine learning models. Shardlow et al. (2020) further validate that simple lexical features (word length, syllable count) outperform contextual embeddings in predicting single-word complexity. Their dataset also includes multi-word expressions, which tend to be more complex than individual words, particularly in biomedical texts.

The role of context emerges as critical in later studies. Kwon et al. (2022) focus on EHR jargon, pre-training models on Wikipedia hyperlinks to improve jargon span detection. Their work highlights challenges with abbreviations ("ENT", "q.6 h"), person-name-based medical concepts ("Azzopardi effect"), and device names ("BiPap"). They also found that the masked language model complements the model solely based on word frequencies. Jiang and Xu (2024) extend this by categorising jargon into Google-easy (searchable) and Google-hard (obscure) terms, proposing customised simplification strategies.

Discrepancies in laypeople's understanding of medical terminology are a recurring theme in the literature. Pylieva et al. (2019) and Sugihara et al. (2024) report that only 22–27% of medical terms are correctly understood by annotators. Sugihara et al. also note demographic variations (for example, younger males tend to struggle with terms related to pregnancy and childbirth). Additionally, Cheng Sheang et al. (2022) observe low inter-annotator agreement (κ = 0.175-0.316) to identify complex words in French clinical texts, further highlighting the subjectivity of complexity judgments.

To help laypeople understand medical jargon, numerous tools have been developed. Initially, these efforts focused on synonym dictionaries and direct word substitution (Alfano et al., 2020; Kandula et al., 2010; Keskisärkkä, 2012; Kvist, 2013). With advances in NLP, machine learning models were trained to address this task (Phatak et al., 2022; Flores et al., 2023; Lu et al., 2023; Botarleanu et al., 2020; van den Bercken et al., 2019; Van et al., 2020; Joseph et al., 2023; Basu et al., 2023). Subsequently, context-aware simplification systems emerged, leveraging NLP to identify and replace complex terms with lay-friendly alternatives in EHRs and patient portals (Jiang and Xu, 2024; Kwon et al., 2022). Similarly, mobile apps

like MediReader (Hendawi et al., 2022) and AI-driven chatbots (Bhatt and Vaghela, 2024; Khamaj, 2025) now provide real-time simplifications.

3 Methodology

In this study, the objective was to assess laypeople's understanding of common medical terminology they are likely to encounter during hospital admission, in EHRs, or in patient information leaflets. Although Pylieva et al. (2019) and Sugihara et al. (2024) have highlighted significant discrepancies in the comprehension of medical terms (with only around 25% of terms correctly understood by their audience), our work diverges by focusing on high-frequency terms to ensure practical relevance. Unlike studies with randomly selected distractors (such as Kauchak and Leroy (2016)), which can inadvertently bias responses toward the most plausible answer, we manually curated all incorrect options to rigorously test participants' true understanding. Our study also measures the level of term recognition prior to correct comprehension.

Our study was carried out in three main phases: (1) Dataset curation, where we systematically selected high-frequency medical terms from diverse sources to ensure relevance and coverage; (2) Annotation and labelling of the dataset, where we designed and administered a questionnaire to assess laypeople's comprehension; and (3) Dataset analysis, where we utilised different models and features to evaluate term complexity.

3.1 Terms Selection

The objective of this step was to compile a comprehensive and balanced set of medical terms likely to confuse patients, ensuring coverage across various sources while filtering for relevance, frequency, and practicality. To achieve this, three sources of terms were used.

The first source comprised medical terms identified through a conventional search engine (Google) query for pages related to confusing medical terminology. The query *"medical terms that confuse patients"* returned 198 web pages, though not all contained relevant unique content or were accessible. After the elimination of irrelevant or corrupted pages, 166 web pages remained, yielding an average of 130 unique terms per page (12,533 unique terms in total). Initially, Gemma2:27B (Gemma Team, 2024) and ChatGPT-4o-mini (OpenAI, 2024) were used to extract medical terms from 4,096-character text chunks of the web pages. The prompts requested enumerated lists of (1) *"terms used in hospitals, including anatomical ones"*, and (2) *"medical terms that can confuse patients"*. However, both models occasionally missed terms, necessitating a switch to manual extraction.

Secondly, multiple medical dictionaries (Baker, 2004; Stöppler, 2013; Merriam-Webster; EMA; Acland, a,b; Law and Martin, 2020) were used to obtain more academic medical terms (27,565 unique terms in total). To account for the lack of recent COVID-19 terminology, we added these terms separately, obtaining them from multiple web pages (50 terms found through the Google query *"medical terms related to COVID-19"*).

Finally, as the third source of medical terms, ChatGPT-4o-mini was used. The model was prompted with: *"Output an enumerated list of 1500 medical terms or phrases that usually confuse patients"*. Although instructed to produce 1,500 terms, the initial output contained significantly fewer. After three prompting attempts, the model generated approximately 1,950 non-unique terms. Following manual filtering, only 125 unique and relevant terms remained.

As the next step, for each term from the three sources, its absolute frequency was obtained from the NGRAMS website[2]. Terms were filtered to include only those with an absolute frequency between 400,000 and 1,500,000. This range was chosen as a compromise: it avoids selecting too many "obvious" terms (which can be difficult to define simply and concisely in the next step) while also excluding terms that may be unfamiliar to most people. Words within this frequency range were uniformly and randomly sampled from the three sources. During the creation of the questionnaires, many selected words were replaced with others of similar frequency for the following reasons:

1. **Overly obvious terms**: phrases whose meanings are easily guessed if the individual terms are known, definitions that include the term itself, or words that are difficult to simplify (e.g. *body surface, seriously ill, liquid form*);
2. **Low relevance to patients**: specialised scientific, chemical, or genomics terminology (e.g. *florid, hydrogen bond, messenger RNA, data bank, The Lancet, Gaussian distribution*);
3. **Duplication**: terms with minor differences (e.g., extra apostrophes, hyphens, or capitalisation) or

[2]https://ngrams.dev

different grammatical forms (e.g., adjective vs. adverb, plural vs. singular);

4. **Uncommon or ambiguous abbreviations**: e.g. *OCD, RBC, CSC*;
5. **Clarification modifications**: terms requiring additions for proper medical context (*UFO* → *the "UFO" procedure*; *Adonis (complex)*).

After these adjustments, a final set of 1,000 words was selected for the questionnaire.

3.2 Questionnaire Design and Execution

To gather data on the recognition and understanding of medical terminology along with socioeconomic information, we conducted a series of questionnaires. The 1,000 refined terms were distributed across 25 questionnaires (40 terms each). This distribution helped to strike a balance between participant numbers and average completion time, as lengthy surveys may reduce concentration and response quality (Sharma, 2022). Each questionnaire comprised four sections:

1. **Self-rated confidence**: Participants first rated their confidence in medical terminology on a 10-point Likert scale;
2. **Term familiarity**: Respondents then evaluated their familiarity with 40 terms using a 4-point forced-choice scale: (1) "I never heard of this term before", (2) "I possibly heard this term before but I don't know its meaning", (3) "I think I know something about this term and know something about its definition", (4) "I know this term and know what it means".
3. **Definition accuracy**: Participants selected the correct definition for each term from four options (matching the terms in section 2). Correct definitions were sourced from online medical articles, prioritizing simple wording. Non-essential details were omitted without compromising meaning. Terms requiring overly lengthy explanations were replaced with frequency-matched alternatives (see Section 3.1);
4. **Socioeconomic data collection** The final section covered the following broad topics: (1) Demographics, (2) Language background, (3) Geographical exposure, (4) Education, interests & profession, (5) History of serious illness (self or close family).

To generate three false but plausible definitions for each term, we utilised LLMs (ChatGPT-4o-mini and DeepSeek-V3 OpenAI, 2024; DeepSeek-AI, 2025b). However, fully automating this process proved unfeasible due to several limitations:

- **Repetitive output**: The models frequently produced similar definitions or variations on the same theme;
- **Formulaic structure**: Generated definitions often followed predictable patterns, such as: *"<Simple definition>, often caused by <...>, [typically occurring <...>, [especially <...>]]"*;
- **Overly personalized language**: The outputs sometimes used informal phrasing (e.g. "your body" instead of "the body"), which differs from standard medical definitions found online;
- **Topic divergence**: In some cases, the generated definitions were entirely unrelated to the medical term. For instance, for *"cardiogram"*, an LLM might propose a definition unrelated to the heart or diagnostic procedures. Such obviously incorrect options could make the correct answer easier to identify (even for someone completely unfamiliar with the term). Although guessing might help patients during consultations, and the exercise may have value in itself, it fails to test confusion between commonly misassociated medical terms (e.g., "bipolar disorder" vs "multiple personality disorder").

Consequently, we used manual data curation with LLM assistance to create the false definitions.

The questionnaire was administered via Prolific[3]. A total of 25 surveys were conducted, with ten participants recruited for each survey. The average completion time was 18 minutes (median = 15 minutes, SD = 10 minutes). The following pre-screening criteria were applied to most surveys: (a) English as the first and primary language; (b) Participants had to be born in, currently reside, and hold nationality of one of the following countries: Ireland, the United Kingdom, the United States, Canada, Australia, or New Zealand. All participants who had previously completed any of the surveys were excluded during pre-screening.

3.3 Dataset Analysis

Survey results were analysed from two perspectives: the participants' perspective (how well a person performs) and the terms' perspective (how complex a term is). For the latter, we compare how the term complexity score derived from participant performance differs from that estimated by LLMs. This analysis was performed using Gemma-3:27B (Gemma Team, 2025), Phi-4:14B (Abdin

[3]https://www.prolific.com/

et al., 2024), Mistral-Small-3.1:24B (Mistral AI, 2025), Llama-3.3:70B (Meta, 2024), DeepSeek-R1 (DeepSeek-AI, 2025a), and ChatGPT-4.1 (OpenAI, 2025)[4].

To instruct an LLM to output evaluation scores, Leng et al. (2023) recommends using an integer (not float) scale, such as 0-3 or 0-4, for better results. Although a larger scale (0-10) was not recommended due to the "difficulty in coming up with distinguishing criteria between all scores", we compared both the 0-4 and 0-10 scales to assess differences. For reliability, we performed five evaluations per term and averaged the results. To balance variability in LLM outputs with reliability and instruction adherence, we set the temperature to 0.2 (AI21, 2025).

As noted in multiple studies (Clarke and Dietz, 2024; Qin et al., 2024; Christiano et al., 2023), LLMs perform better as rankers than scorers. Thus, we sorted the terms in each questionnaire (40 terms per ranking) using Python's built-in sorting function with a custom LLM-based comparator. For the comparator, we set the temperature to 0.0 and used the following prompt: *You are a helpful AI assistant. Your task is to predict which term is likely to be correctly understood by the majority of people: term 1: <term_1> or term 2: <term_2 >. Output 1 if term 1 is likely to be correctly understood by the majority of people than term 2. Otherwise, output 2. Output just the number*. This approach was inspired by Wang et al. (2025), but we replaced their Bubble Sort algorithm with Python's Timsort (Peters, 2002) to reduce LLM calls. Timsort efficiently handles partially ordered data, and terms were pre-sorted by frequency – a known correlate of complexity (Hashimoto, 2021; Stewart et al., 2022).

Finally, these ranking results were transformed into complexity scores. For this conversion, we calculated each term's percentile within its questionnaire and mapped it to a number in the [0, 10] interval using the percent point function (PPF) of the Gaussian distribution $\mathcal{N}(5, 1.8^2)$ (University of Illinois, 2025; Virtanen et al., 2020). The output was then min-max scaled to the 0–1 range. This two-step procedure mitigates the numerical instability that arises from the small scale σ of the Gaussian distribution by getting 0-1 score in one step.

[4]The following prompt was used for each of the listed models: *You are a helpful AI assistant. A medical term is called simple if it is understood correctly by the majority of people. Your task is to evaluate the simplicity of the medical term "<...>" on a scale from 0 (very complex) to <MAX_GRADE> (very simple). Output just the simplicity score.*

3.4 Features for Terms Complexity Prediction

We selected features for ML models to predict term complexity based on three recent papers.

Inspired by Dalvean (2024), we used binary positional letter variables (e.g., "a2" = *True* – or 1 – if the *second* letter is "a"). From their study, we selected 66 significant binary letter position variables. Dalvean solved the complexity prediction task for individual words, whereas in our paper, a medical term may consist of multiple words. In these cases (e.g., *"kidney failure"*), values of positional letter variables are averaged across words (e.g., if "a2" is 0 for *"kidney"* and 1 for *"failure"*, the term-level value of "a2" variable is 0.5).

From Cheng Sheang et al. (2022), we adopted: (1) FastText embeddings (reduced from 300 to 10 dimensions via PCA to avoid the curse of dimensionality); (2) English Wikipedia absolute word frequency; (3) Word character, syllable and vowel count; and (4) Word rank (the frequency order from the FastText pre-trained model). All features are averaged for multi-word terms.

Finally, following Mosquera (2021), we included the following features:

- **Morphological**: Word character, syllable (Holtzscher, 2022), and morpheme counts (Cuko and Warren, 2023); medical word parts number (e.g. 2 for angiogram, as it consists of angio- and -gram Tsutsumi, 2017); is acronym/abbreviation (regex-based), word lemma/stem length (Bird et al., 2009);
- **Frequency**: Google n-gram[5] frequency (of a full medical term), Wikipedia absolute word frequency, Zipf frequency (Speer, 2023), Wikipedia documents count (where a term appeared), consonant frequency;
- **Lexical**: WordNet senses, synonyms, hypernyms, and hyponyms counts (Miller, 1994);
- **Psycholinguistic/other**: Average age of acquisition (Kuperman et al., 2012), first year of appearance (of a full term) using Google n-gram.

Similarly, the features are averaged for multi-word terms (except Google n-gram-related ones).

4 Results and Analysis

In this section, we present our analysis of the collected data (information obtained via question-

[5]https://books.google.com/ngrams/

naires and Prolific demographic data) on the medical terms complexity.

4.1 Respondents' Profiles and Terms Analysis

The mean and median age of all questionnaire respondents were 39.1 and 37.0, respectively. The age distribution was right-skewed, with one-third of participants under 30 years of age.

Most of the participants were from the US, the UK, South Africa and Canada, and more than 90% were native English speakers. The gender distribution was balanced, but the ethnicity was heavily skewed toward White (70%), followed by Black (20%) and Mixed/Asian/Other (10%).

For 231 respondents, it was known whether they were students or not. Approximately 70% (170 people) were not students at the time of the survey, with "Mathematics and Computer Science" and "Health Sciences" being the most popular fields of education. Only 3% of respondents reported having less than a high school diploma, while 23% had only a high school diploma. In addition to collecting socioeconomic data, we collected information on participants' confidence in medical terminology and whether they considered themselves medical professionals. Participants rated their confidence in medical terminology highly, with a mean of 6.3 on a 0-10 Likert scale, a median of 7.0, and a standard deviation of 2.4. Approximately 20% of the participants identified as trained or practising medical professionals (e.g., doctors, nurses, etc.).

To estimate participants' fine-grained self-reported knowledge of medical terms, a new variable, familiarity score (y), was introduced:
$y = 1 * x_1 + 2 * x_2 + 3 * x_3 + 4 * x_4$, where:
x_1 = number of times that the option "I had never heard of this term before" was selected;
x_2 = selections of "I possibly heard this term before but I don't know its meaning";
x_3 = selections of "I think I know something about this term and know something about its definition";
x_4 = selections of "I know this term and know what it means".

With 40 terms per questionnaire, the possible *familiarity score* ranges from 40 (if all responses scored 1) to 160 (if all responses scored 4). The distribution of familiarity scores was tested for normality using the Kolmogorov-Smirnov test (KS statistic = 0.0708, p = 0.1523, n = 252), which did not reject the null hypothesis of normality. However, the Shapiro-Wilk and D'Agostino K^2 tests suggested non-normality. The scores had a mean of 112 and a median of 111 (SD = 23), with a skewness of 0.08 and kurtosis of -0.7.

Let us examine how accurately participants selected the correct definitions for all 40 terms in the survey (participants' *accuracy score*). The mean and median values of the correct definition selection score were approximately 0.6, indicating that on average each participant correctly identified about 24 out of the 40 terms.

Similarly, the term *complexity score* is determined by the success rate of participants in selecting its correct definition (calculated as the number of correct selections divided by the total number of attempts). A score close to zero indicates a very difficult term (understood correctly by none of the participants), while a score close to one represents a very simple term (understood correctly by all participants). The mean and median complexity scores of the medical terms were approximately 0.6. The distribution showed slight negative skewness (-0.24, indicating a tendency towards simpler terms) and platykurtic kurtosis (-0.71).

We calculated the correlation between n-gram frequencies (originally used for the selection of medical terms, from ngrams.dev) and term familiarity scores, finding both Pearson (r = 0.065) and Spearman (ρ = 0.087) correlations to be negligible. We then analysed absolute term frequencies using an English Wikipedia dataset from BEEspoke Data (2023) (counting total word occurrences). Here, the Spearman correlation remained weak, though slightly stronger (ρ = 0.25), while Pearson's correlation remained close to zero (r = 0.05). Finally, we examined the correlation between term familiarity scores and the number of Wikipedia documents containing each term, obtaining similar results (Pearson r = -0.03; Spearman ρ = 0.31).

4.2 Medical Terms Complexity Modelling

Understanding the complexity of medical terms is valuable for improving readability and accessibility in healthcare communication. It is interesting to examine the features that contribute to term complexity. Three sets of features (discussed in Section 3.4) were analysed. For the score prediction task (regression), multiple machine learning models were tested for each feature set, including linear regression, decision tree, random forest, multilayer perceptron (MLP), CatBoost, and support vector regression (SVR) (Pedregosa et al., 2011; Dorogush

et al., 2018), using 5-fold cross-validation.

First, as a baseline, a simple model based solely on term frequencies (including n-gram frequency from *ngrams.dev*, absolute term frequency from Wikipedia, and the count of Wikipedia documents in which a term appeared BEEspoke Data (2023)) was tested. The best-performing model was the decision tree regressor, with an RMSE of 0.23 and a Median Absolute Error (Median AE) of 0.17. A second baseline was tested, which simply predicted the mean term complexity score of 0.62 (for RMSE) and the median term complexity of 0.6 (for Median AE). For this baseline, the RMSE was 0.23 and the Median AE was 0.20. Third, the aforementioned models were trained using only the familiarity score. Here, the SVR and Decision Tree models performed best, achieving an RMSE of around 0.20 and a Median AE of around 0.13. However, the familiarity score was not used in subsequent experiments, as it is a data leak in a way.

The best-performing model using the features from Dalvean (2024) was CatBoost regression, with an RMSE of 0.23 and a Median AE of 0.18 (while the decision tree regression model achieved comparable results). The most important features, based on CatBoost's feature importance analysis, were *d1*, *n4*, *t1*, *u5*, *s5*, *s3*, *s2*, and *t3*.

Among the models using features from Cheng Sheang et al. (2022), random forest regression and CatBoost regression performed best, with RMSE values around 0.22 and median absolute errors around 0.16. For the CatBoost regressor, we analysed both feature importance and Shapley values. The most influential features were *Word Rank from FastText*, *Embedding_0*, and *Embedding_3*.

A model incorporating features from Mosquera (2021) was explored. Again, the CatBoost regressor performed best, while the decision tree regressor and random forest regressor achieved comparable results. The CatBoost model achieved an RMSE of around 0.22 and a median absolute error (AE) of 0.16. For this model, feature importance and Shapley values were analysed. The most important features included *Zipf Frequency*, *Morpheme Count*, *Google N-gram Frequency*, *Absolute Word Frequency*, *Wikipedia Document Count*, and *Average Age of Acquisition*. Finally, the BERT Base Uncased model (Devlin et al., 2018) was fine-tuned to predict the target scores based solely on the terms themselves. This model achieved an RMSE of 0.23 and a median AE of 0.16.

Overall, the best models performed comparably, with an RMSE of around 0.22 and a median AE of around 0.16. These results are close to the baseline model's mean and median scores (0.23 and 0.20, respectively). However, the performance remains poor given that the target variable's mean and median values are approximately 0.6. This suggests that the tested variables provide little predictive benefit.

4.3 LLMs Scoring and Ranking Capabilities of Medical Terms Complexity Evaluation

The ranking and scoring abilities of LLMs were explored to predict the complexity of medical terms. Table 1 presents the overall results for each LLM (using both 0–4 and 0–10 evaluation scales). As shown, Mistral-Small-3.1:24B and Phi-4:14B performed best in terms of both RMSE and Median AE. Contrary to the conclusion in Leng et al. (2023), the 0–10 scale evaluation yielded better results than the smaller 0–4 scale. The RMSE and Median AE scores suggest that LLMs have some potential for medical term complexity evaluation, outperforming the approach of training ML models using standard features (as discussed in the previous section). Additionally, refining prompts to include conditions aligned with the target audience may further improve performance.

LLM	RMSE	MAE
DeepSeek-R1	0.24 (0.26)	0.17 (0.20)
Gemma-3:27B	0.28 (0.30)	0.20 (0.22)
ChatGPT-4.1	0.28 (0.30)	0.20 (0.20)
Llama-3.3:70B	0.27 (0.28)	0.20 (0.20)
Mistral-Small-3.1:24B	**0.21** (0.22)	0.14 (0.15)
Phi-4:14B	**0.21** (**0.21**)	**0.12** (0.15)

Table 1: Summary of LLMs complexity prediction results; MAE means Medium Absolute Error. The results in parentheses are for 0-4 scale.

To compare the correlation between human and LLM errors, we analysed the errors of DeepSeek-R1 with a 0–10 scale (terms sorted by the absolute difference between predicted and actual complexity scores) and human error (terms sorted by the absolute difference between normalised term familiarity scores and actual complexity scores). The correlation values were approximately 0.42 for Pearson and Spearman rank correlations and 0.3 for Kendall's Tau, indicating weak to moderate relationships. We also evaluated the Pearson and Spearman rank correlations between LLM-evaluated complexity scores and ground-truth human-based

scores to assess how well LLMs align with human judgments. DeepSeek-R1 (0–10 scale) achieved the highest overall correlation (0.51 Pearson, 0.52 Spearman), while other models showed comparable results (Pearson: 0.42-0.50; Spearman: 0.46-0.51).

The ranking abilities of LLMs were tested for the task of ranking medical terms by complexity. First, term complexities were derived from rankings using the procedure described in Section 3.3. The lowest RMSE score (0.25) was achieved by the DeepSeek-R1 and ChatGPT-4.1 models, which was very close to the other models' scores (with a maximum RMSE of 0.27). The median absolute error was 0.20 for all models. Pearson and Spearman rank correlations between LLM-evaluated and ground-truth complexity scores were also measured. ChatGPT-4.1 had the highest overall correlation (Pearson: 0.48; Spearman: 0.50), while Gemma-3:27B and Phi-4-14B had the lowest (Pearson: 0.36; Spearman: 0.39). Similarly, to compare human and LLM errors, we analysed the correlation between the absolute error of LLM rankings and human error. This analysis yielded weaker correlations: 0.25 for Pearson and Spearman, and 0.18 for Kendall's Tau.

5 Discussion

The study analysed the complexity and awareness of medical terms among respondents, mainly from English-speaking countries. Most of the respondents were confident in their knowledge of medical terminology. Machine learning models showed limited success in predicting the complexity of medical terms. CatBoost regression performed the best but was only slightly better than the mean-guessing baseline. LLMs demonstrated potential in assessing term complexity, outperforming traditional ML models. However, in terms of RMSE scores, both traditional ML models and LLM achieved comparable results (0.22 and 0.21, respectively), which were only marginally better than baseline (0.23). This suggests the inherent difficulty of the task and the limitations of the features used for complexity prediction. An important consideration is whether the computational cost of LLMs (especially, "reasoning" ones) justifies such a small improvement.

6 Limitations

While Prolific provided a diverse participant pool, its users do not fully represent real patient populations due to platform access requirements (e.g., often English proficiency, digital literacy). In addition, participants viewed isolated medical terms without any context, which may affect understanding in real world settings. Furthermore, although we tested basic comprehension of terms from multiple medical domains, patients are typically more familiar with terms related to their personal health conditions. Medical terms are inherently complex, and brief definitions (e.g., defining ADHD as "a neurodevelopmental disorder with inattention/hyperactivity symptom") cannot capture all nuances (e.g. subtypes, life impact, or treatment options). Thus, even selecting an accurate definition does not ensure full comprehension of a term.

7 Future work

For direct extensions of this work, testing patients on medical terms from their own health records would better assess their understanding of relevant terminology. Another important consideration is public misconceptions about medications. For example, in our study, only 1 in 10 participants correctly identified *metronidazole* as an antibiotic; most incorrectly associated it with viral infections or blood pressure management. Although based on a small sample, such misunderstandings could contribute to antibiotic resistance (Naghavi) and warrant further investigation.

8 Conclusion

The results suggest that participants generally rate their knowledge of medical terms highly, although their actual understanding may vary. Despite the modest performance gains of machine learning models, LLMs offer a promising alternative for assessing the complexity of medical terms. Further development and integration of contextual and demographic data could improve predictive accuracy in future studies.

Acknowledgments

This work was funded by Taighde Éireann – Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183) and the ADAPT SFI Research Centre for AI-Driven Digital Content Technology under Grant No. 13/RC/2106_P2.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 Technical Report.

Robert D. Acland. a. Acland's Video Atlas of Human Anatomy: A-Z INDEX. Accessed in October, 2024.

Robert D. Acland. b. Acland's Video Atlas of Human Anatomy: Glossary. Accessed in October, 2024.

AI21. 2025. What is LLM temperature? Accessed in May, 2025.

Hilal Al Shamsi, Abdullah G. Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for Healthcare: A systematic review. *Oman Medical Journal*, 35(2).

Marco Alfano, Biagio Lenzitti, Giosuè Lo Bosco, Cinzia Muriana, Tommaso Piazza, and Giovanni Vizzini. 2020. Design, development and validation of a system for automatic help to medical text understanding. *International Journal of Medical Informatics*, 138:104–109.

Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. Evaluating LLMs for targeted concept simplification for domain-specific texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226, Miami, Florida, USA. Association for Computational Linguistics.

Lindsey Baker. 2004. Dictionary of Medical Terms (4th edition). *Reference Reviews*, 18(7):1–37.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-EASI: Finely annotated dataset and models for controllable simplification of medical texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14093–14101.

BEEspoke Data. 2023. Bee-spoke-data/wikipedia-20230901.en-deduped. Accessed in March, 2025.

Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating Neural Text Simplification in the Medical Domain. In *The World Wide Web Conference*, WWW '19, page 3286–3292, New York, NY, USA. Association for Computing Machinery.

Ahan Bhatt and Nandan Vaghela. 2024. Med-Bot: An AI-Powered Assistant to Provide Accurate and Reliable Medical Information.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Robert-Mihai Botarleanu, Mihai Dascalu, Scott Andrew Crossley, and Danielle S. McNamara. 2020. Sequence-to-Sequence Models for Automated Text Simplification. In *Artificial Intelligence in Education*, pages 31–36, Cham. Springer International Publishing.

Birgir Briem, Thornorlákur Karlsson, Geir Tryggvason, and Olafur Baldursson. 2004. Public comprehension of medical terminology. *Laeknabladid*, 90:111–119.

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 116–125, Avignon, France. ATALA.

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 116–125, Avignon, France. ATALA.

Olga Cherednichenko, Nadiia Babkova, and Olga Kanishcheva. 2018. Complex term identification for Ukrainian medical texts. In *International Workshop on Informatics and Data-Driven Medicine*.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Charles L. A. Clarke and Laura Dietz. 2024. LLM-based relevance assessment still can't replace human relevance assessment. *CoRR*, abs/2412.17156.

Enkeleda Cuko and Paul Warren. 2023. morphemes. Version 1.2.0.

Michael Coleman Dalvean. 2024. Using letter positional probabilities to assess word complexity. *SSRN Electronic Journal*.

DeepSeek-AI. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

DeepSeek-AI. 2025b. DeepSeek-V3 Technical Report.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support.

EMA. EMA's medical terms simplifier. *EMA/329258/2022 Rev*, 1.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873, Singapore. Association for Computational Linguistics.

Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size.

Gemma Team. 2025. Gemma 3 Technical Report.

Rachael Gotlieb, Corinne Praska, Marissa A. Hendrickson, Jordan Marmet, Victoria Charpentier, Emily Hause, Katherine A. Allen, Scott Lunos, and Michael B. Pitt. 2022. Accuracy in patient understanding of common medical phrases. *JAMA Network Open*, 5(11).

Natalia Grabar, Thierry Hamon, and Dany Amiot. 2014. Automatic diagnosis of understanding of medical words. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 11–20, Gothenburg, Sweden. Association for Computational Linguistics.

Brett James Hashimoto. 2021. Is Frequency Enough?: The Frequency Model in Vocabulary Size Testing. *Language Assessment Quarterly*, 18(2):171–187.

Rasha Hendawi, Shadi Alian, and Juan Li. 2022. A smart mobile app to simplify medical documents and improve health literacy: System design and feasibility validation. *JMIR Formative Research*, 6(4).

Michael Holtzscher. 2022. syllapy. Version 0.7.2.

Chao Jiang and Wei Xu. 2024. MedReadMe: A Systematic Study for Fine-grained Sentence Readability in Medical Domain.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A Semantic and Syntactic Text Simplification Tool for Health Content. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:366–70.

David Kauchak and Gondy Leroy. 2016. Moving beyond readability metrics for health-related text simplification. *IT Professional*, 18(3):45–51.

Robin Keskisärkkä. 2012. Automatic Text Simplification via Synonym Replacement.

Abdulrahman Khamaj. 2025. AI-enhanced chatbot for improving healthcare usability and accessibility for older adults. *Alexandria Engineering Journal*, 116:202–213.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Maria Kvist. 2013. Professional Language in Swedish Radiology Reports - Characterization for Patient-Adapted Text Simplification. *Scandinavian Conference on Health Informatics*, 91(12):55–59.

Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 2022. MedJEx: A medical jargon extraction model with Wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John P Lalor, David A Levy, Harmon S Jordan, Wen Hu, Jenni Kim Smirnova, and Hong Yu. 2024. Evaluating expert-layperson agreement in identifying jargon terms in electronic health record notes: Observational Study. *Journal of Medical Internet Research*, 26.

Jonathan Law and Elizabeth Martin, editors. 2020. *Concise Medical Dictionary*. Oxford University Press.

Thomas W. LeBlanc, Ashley Hesson, Andrew Williams, Chris Feudtner, Margaret Holmes-Rovner, Lillie D. Williamson, and Peter A. Ubel. 2014. Patient understanding of medical jargon: A survey study of U.S. medical students. *Patient Education and Counseling*, 95(2):238–242.

Quinn Leng, Kasey Uhlenhuth, and Alkis Polyzotis. 2023. Best practices for LLM evaluation of rag applications. Accessed in March, 2025.

E.Brooke Lerner, Dietrich V.K. Jehle, David M. Janicke, and Ronald M. Moscati. 2000. Medical communication: Do our patients understand? *The American Journal of Emergency Medicine*, 18(7):764–766.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023. NapSS: Paragraph-level Medical Text Simplification via Narrative Prompting and Sentence-matching Summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1079–1091, Dubrovnik, Croatia. Association for Computational Linguistics.

Tamara L. McCarron, Thomas Noseworthy, Karen Moffat, Gloria Wilkinson, Sandra Zelinsky, Deborah White, Derek Hassay, Diane L. Lorenzetti, and Nancy J. Marlett. 2019. Understanding the motivations of patients: A co-designed project to understand the factors behind patient engagement. *Health Expectations*, 22(4):709–720.

Hugo J. van Mens, Mirte M. van Eysden, Remko Nienhuis, Johannes J. van Delden, Nicolette F. de Keizer, and Ronald Cornet. 2020. Evaluation of lexical clarification by patients reading their clinical notes: A quasi-experimental interview study. *BMC Medical Informatics and Decision Making*, 20(S10).

Merriam-Webster. Merriam-Webster medical dictionary. Accessed in October, 2024.

Meta. 2024. Llama 3.3.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mistral AI. 2025. Mistral Small 3.1.

Alejandro Mosquera. 2021. Alejandro mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559, Online. Association for Computational Linguistics.

Laxsini Murugesu, Monique Heijmans, Jany Rademakers, and Mirjam P. Fransen. 2022. Challenges and solutions in communication with patients with low health literacy: Perspectives of Healthcare Providers. *PLOS ONE*, 17(5).

Mohsen Naghavi. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*, 404(ue 10459):1199 – 1226.

Ana Luisa Neves, Lisa Freise, Liliana Laranjo, Alexander W Carter, Ara Darzi, and Erik Mayer. 2020. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. *BMJ Quality & Safety*, 29(12):1019–1032.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence.

OpenAI. 2025. Introducing GPT-4.1 in the API.

R.L. O'Connell, S.K. Hartridge-Lambert, N. Din, E.R. St John, C. Hitchins, and T. Johnson. 2013. Patients' understanding of medical terminology used in the Breast Clinic. *The Breast*, 22(5):836–838.

Lydia O'Sullivan, Prasanth Sukumar, Rachel Crowley, Eilish McAuliffe, and Peter Doran. 2020. Readability and understandability of clinical research patient information leaflets and consent forms in Ireland and the UK: A retrospective quantitative analysis. *BMJ Open*, 10(9).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tim Peters. 2002. [Python-Dev] Sorting. Accessed in May, 2025.

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, and Vijay Mago. 2022. Medical text simplification using reinforcement learning (TESLEA): Deep learning–based text simplification approach. *JMIR Medical Informatics*, 10(11).

Arwen H. Pieterse, Nienke A. Jager, Ellen M.A. Smets, and Inge Henselmans. 2012. Lay understanding of common medical terminology in oncology. *Psycho-Oncology*, 22(5):1186–1191.

Hanna Pylieva, Artem Chernodub, Natalia Grabar, and Thierry Hamon. 2019. RNN embeddings for identifying difficult to understand medical words. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 97–104, Florence, Italy. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Hunny Sharma. 2022. How short or long should be a questionnaire for any research? researchers dilemma in deciding the appropriate questionnaire length. *Saudi Journal of Anaesthesia*, 16(1):65–68.

Robyn Speer. 2023. wordfreq. Version 3.1.1.

Jeffrey Stewart, Joseph P. Vitta, Christopher Nicklin, Stuart McLean, Geoffrey G. Pinchbeck, and Brandon Kramer. 2022. The Relationship between Word Difficulty and Frequency: A Response to Hashimoto (2021). *Language Assessment Quarterly*, 19(1):90–101.

Shiel William C. Stöppler, Melissa Conrad. 2013. *Webster's New World Medical Dictionary*, 3rd edition edition. Wiley.

Soichiro Sugihara, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2024. Semi-automatic construction of a word complexity lexicon for Japanese medical terminology. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 329–333, Mexico City, Mexico. Association for Computational Linguistics.

Hieu Tran, Zonghai Yao, Won Seok Jang, Sharmin Sultana, Allen Chang, Yuan Zhang, and Hong Yu. 2025. MedReadCtrl: Personalizing medical text generation with readability-controlled instruction learning. *medRxiv*.

Tzu-I Tsai, Wen-Ry Yu, and Shoou-Yih D Lee. 2018. Is health literacy associated with Greater Medical Care Trust? *International Journal for Quality in Health Care*, 30(7):514–519.

Yutaka Tsutsumi. 2017. Medical terminology. Accessed in March, 2025.

Taner Tunç, Hasan Fehmi Demirci, and Aydan Ermiş. 2025. The mediating role of Health Literacy in the relationship between trust in Public Health Authorities and distrust in Health Systems. *BMC Public Health*, 25(1).

University of Illinois. 2025. Python functions for random distributions. Accessed in April, 2025.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: The Autocomplete for Medical Text Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1424–1434, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and et al. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

Tianshu Wang, Xiaoyang Chen, Hongyu Lin, Xuanang Chen, Xianpei Han, Le Sun, Hao Wang, and Zhenyu Zeng. 2025. Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 96–109, Abu Dhabi, UAE. Association for Computational Linguistics.

Barry D. Weiss. 2007. *Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians. 2nd ed.* American Medical Association Foundation.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *Proceedings of the 6th International Conference on Biological and Medical Data Analysis*, ISBMDA'05, page 184–192, Berlin, Heidelberg. Springer-Verlag.

Agnieszka Zimmermann, Anna Pilarska, Aleksandra Gaworska-Krzemińska, Jerzy Jankau, and Marsha N. Cohen. 2021. Written Informed Consent—Translating into Plain Language. A Pilot Study. *Healthcare*, 9(2).