

LANGUAGES

From Layperson to Medical Expert: A Corpus-Based Analysis of Linguistic Variation in Audience-Conditioned Biomedical Discourse

Quoc Khanh Nguyen^{*,1} and Nhat Quang Vo²

¹The University of Danang, Vietnam

²Ho Chi Minh University of Banking, Vietnam

*Corresponding author: **Quoc Khanh Nguyen** – The University of Danang, Vietnam.

Email: av32.1b2cva1d047@oncce.udn.vn

Received: 09/07/2026. **Revised:** 15/08/2026. **Accepted:** 25/08/2026. **DOI:** [10.53901/tjs.2026.v01.issue03.art11](https://doi.org/10.53901/tjs.2026.v01.issue03.art11)

Abstract

Biomedical text accessibility is often evaluated through readability formulas, although audience adaptation may involve changes across multiple linguistic dimensions. This study analyzed the Persona-Guided Controllable Biomedical Summarization (PERCS) corpus, an expert-curated, GPT-4-assisted resource in which persona-specific drafts were reviewed and revised by physicians. Following corpus-integrity screening, the analytic dataset comprised 1,984 English-language summaries derived from 496 biomedical source abstracts, with matched versions for laypersons, premedical students, non-medical researchers, and medical experts. Linear mixed-effects models accounted for the repeated-source structure and tested categorical audience differences, ordered trends, and nonlinear trajectories. Twelve theory-driven linguistic features were examined, with Flesch Reading Ease and Flesch-Kincaid Grade Level used as conventional readability baselines. All 12 linguistic features showed significant overall audience effects after false-discovery-rate correction. Information-packaging measures showed the clearest expertise-related progression, whereas several grammatical and cohesive features followed nonlinear or non-monotonic trajectories. Differences between the premedical and non-medical researcher conditions were consistently smaller than those involving the layperson or medical-expert endpoints. In a source-separated held-out evaluation, combining readability with the 12 linguistic features increased macro-F1 from 0.610 to 0.677 and reduced multiclass log loss from 0.804 to 0.664 relative to readability alone. The advantage remained under length-controlled sensitivity analyses. These findings indicate that audience-conditioned biomedical rewriting in PERCS is distributed across multiple linguistic domains and is uneven rather than reducible to a single readability gradient. Because PERCS is a controlled, persona-conditioned corpus, the observed linguistic profiles should not be interpreted as evidence of actual reader comprehension or as equivalent to naturally occurring human audience accommodation.

Keywords: audience adaptation; biomedical communication; corpus linguistics; linguistic complexity; plain language; readability

Introduction

Biomedical information increasingly reaches readers beyond the specialist communities for which scientific and medical texts are conventionally written. Research abstracts, patient-facing materials, online medical resources, and plain-language summaries may communicate related content, but their intended readers differ in scientific knowledge, medical familiarity, and information needs. Plain-language communication is one response because it aims to improve clarity without removing essential meaning [1]. Accessibility, however, cannot be reduced to a single text property. Passages with similar reading-grade estimates may differ in vocabulary, grammatical organization, cohesion, and the extent to which information is compressed into nominal structures.

Audience adaptation offers a useful way to concep-

tualize such differences. The broader audience-design principle holds that linguistic choices are shaped partly by assumptions about intended recipients [2]. In biomedical communication, audiences cannot always be represented by a simple expert versus non-expert distinction. A premedical student may have emerging domain knowledge, whereas a researcher outside medicine may have advanced scientific literacy but less biomedical specialization. A medical expert differs again in the amount of technical terminology, methodological detail, and conceptual compression that can be assumed. These four profiles encode different combinations of biomedical familiarity and broader scientific literacy. They are therefore treated here as distinct audience conditions rather than as psychologically equidistant points on a validated expertise scale.

Biomedical simplification research has traditionally

concentrated on professional-to-consumer adaptation. PLABA provides manually adapted biomedical abstracts [3], Med-EASi adds fine-grained annotations for simplification transformations [4], and MultiMSD extends professional-consumer simplification across nine languages [5]. More recent work has moved toward finer control over intended audiences and textual properties through persona-based summarization [6], attribute-based lay summarization [7], and readability-controlled generation [8]. These developments motivate a more specific question than whether a biomedical text is simply easier: how is the same content linguistically reorganized for different intended audiences?

The Persona-Guided Controllable Biomedical Summarization (PERCS) dataset provides a source-matched design for this question [9]. PERCS contains four persona-conditioned summaries for each biomedical source abstract, targeting laypersons, premedical students, non-medical researchers, and medical experts. Initial drafts were generated with GPT-4 and then reviewed and revised by physicians. PERCS is therefore treated here as an expert-curated, GPT-4-assisted resource rather than naturally authored discourse from the four audience groups. Its matched structure nevertheless permits audience-conditioned linguistic differences to be examined while holding the underlying biomedical source constant.

Conventional readability measures provide useful but limited information about such differences. Flesch-type formulas summarize surface characteristics such as sentence and word or syllable length, but they do not directly represent vocabulary commonness, grammatical organization, referential continuity, discourse signaling, or information density. Previous work has shown that multiple linguistic features can improve modeling of medical-text difficulty beyond a conventional readability measure [10], and that fine-grained lexical information, including jargon, can contribute to medical readability assessment [11]. Readability should not, however, be equated with comprehension. A recent scoping review identified extensive reliance on automated readability metrics and limited direct evaluation with lay readers [12]. The present outcome is therefore linguistic realization of intended audience adaptation, not measured understanding.

Corpus research has long shown that register and discourse variation may be distributed across multiple linguistic systems rather than captured adequately by a single continuum [13]. Accordingly, the present study examines a theory-driven set of lexical, grammatical, information-packaging, and cohesive features. These linguistic domains may show monotonic, threshold-like, plateau, or non-monotonic changes rather than moving in parallel. PlainMedScale similarly showed that measures distinguishing two medical registers do not necessarily form clean gradients across multiple levels [14]. Unlike that resource, PERCS supplies four audience-conditioned versions of the same source abstract, offering tighter source control over the un-

derlying biomedical content.

The original PERCS study reported technical validation showing persona-related differences in readability, vocabulary, and content depth [9]. However, these broad differences do not show how distinct linguistic systems vary when the biomedical source is held constant, whether those systems follow comparable or divergent trajectories across the four audience conditions, or whether fine-grained linguistic information contributes beyond conventional readability. The present study addresses this gap by modeling source-matched variation across lexical, grammatical, information-packaging, and cohesive features and by evaluating their incremental discriminative value beyond readability. The study therefore asks: RQ1, how do lexical, grammatical, information-packaging, and cohesive features vary across the four audience conditions? RQ2, which features show a consistent directional progression across the ordered PERCS audience sequence? RQ3, do different linguistic domains follow distinct trajectories across the four audience conditions? RQ4, to what extent do the selected linguistic features distinguish the four audience conditions beyond conventional readability measures?

Materials and Methods

Research design and corpus

This quantitative corpus-based repeated-measures study used PERCS [9]. Each biomedical source abstract has four audience-conditioned summaries for laypersons, premedical students, non-medical researchers, and medical experts. Initial persona-specific drafts were generated with GPT-4 and reviewed and revised by physicians for factual accuracy and persona alignment. The four summaries from each source were treated as matched observations, with source ID as the grouping unit. Audience was modeled primarily as a four-category factor. For the planned trend analysis, the four PERCS conditions were additionally coded 1 = layperson, 2 = premedical, 3 = researcher, and 4 = expert, creating an equally spaced analytic contrast across the dataset's ordered persona sequence. This coding was used only to test directional progression across the four designed conditions; it should not be interpreted as evidence that the audiences are psychologically or substantively equidistant in expertise.

The original release contained 500 source records and 2,000 summaries. A pre-analysis audit identified two duplicate source records and two sources with unusable persona outputs: one literal missing placeholder and one premedical summary copied from an unrelated source. These were excluded before feature extraction. The final matched corpus contained 496 source abstracts and 1,984 summaries. The official source-level split was retained: 348 training sources and 148 test sources, with PLOS and PLABA evenly represented. The retained partitions had no source-ID overlap and every source had exactly four audience versions. Raw files were preserved unchanged; deter-

Table 1: Composition of the analytic PERCS corpus

Partition	Sources	Summaries	PLOS	PLABA
Train	348	1,392	174	174
Test	148	592	74	74
Total	496	1,984	248	248

ministic Unicode repairs were applied only to analysis text and all exclusions were recorded before inferential analysis.

The expert prompt imposed a lower maximum length (250 words) than the other three conditions (350 words). Text length was therefore treated as a potential design confound and addressed in sensitivity analyses rather than used as a primary accessibility feature. The composition of the final analytic corpus is summarized in Table 1.

Linguistic measures and feature screening

Feature selection was theory-driven across lexical characteristics, grammatical organization, information packaging, and cohesion. Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) were calculated separately as conventional readability baselines. The 12 primary linguistic measures are summarized in Table 2. Lexical diversity was represented by HD-D; MTLD and MATTR were calculated during screening but not simultaneously retained because of strong overlap. Lexical sophistication was represented by mean general-English log frequency, with lower values indicating less common vocabulary. Adjacent argument overlap captured recurrence of noun and pronoun stems across neighboring sentences.

Candidate measures were screened before the primary analyses. Extraction reliability and missingness were checked first. Redundancy was assessed using the training partition only; $|\text{Spearman } \rho| \geq 0.85$ was pre-specified as a near-redundancy flag rather than an automatic deletion rule. Flagged variables were reviewed for conceptual overlap and interpretability. The final multivariate linguistic set had variance inflation factors below 5. A provisional source-corpus rarity feature was removed before RQ1-RQ4 because construct-validity checking showed that it could assign high rarity to ordinary words absent from the training source corpus. No feature was retained or removed because of its later audience effect. All measures were generated through a deterministic Python pipeline. The retained feature definitions, analysis specification, and computational record are provided in Supplementary File S1.

Statistical analysis

Audience differences

For RQ1, a separate linear mixed-effects model was fitted for each linguistic feature with audience as a fixed categorical effect and source ID as a random intercept. Outcomes were standardized for comparable coefficients, while raw-unit estimated means and

contrasts were retained for interpretation. The global audience effect was tested against a random-intercept null model. Benjamini-Hochberg false-discovery-rate correction was applied across the 12 primary features. Six model-based pairwise contrasts per feature were adjusted with Holm correction, with effect estimates and 95% confidence intervals retained.

Ordered trends and trajectory classification

For RQ2, the planned 1–4 coding described above was entered as a numeric fixed effect with a source-ID random intercept. The resulting slope tests directional change across equally spaced analytic steps in the PERCS persona sequence and is not interpreted as a psychometrically calibrated measure of expertise. A feature was classified as monotonic only when the Benjamini-Hochberg-adjusted ordered trend had $q < 0.05$ and all three adjacent estimated transitions had the same direction; strong stepwise monotonicity additionally required all three adjacent contrasts to remain significant after Holm correction. For RQ3, the ordered and categorical models were compared by likelihood-ratio test, with nonlinearity requiring FDR-adjusted $q < 0.05$. A monotonic nonlinear feature received an early-, middle-, or late-threshold label only when its largest absolute adjacent standardized change occurred at that boundary and was at least 1.5 times the second-largest adjacent change; otherwise it remained monotonic nonlinear. Significant ordered trends with directional reversals were labeled directional non-monotonic. These labels describe trajectory shape rather than an accessibility scale.

Discrimination beyond readability

RQ4 compared regularized multinomial logistic regression models using: A, FRE and FKGL; B, the 12 linguistic features; and C, readability plus the 12 linguistic features. Model development used the 348 training sources; the 148 test sources remained untouched until final evaluation. Five-fold grouped cross-validation kept all four summaries from a source in the same fold. Predictor standardization occurred inside the pipeline, and regularization strength was selected on training data by minimum mean multi-class log loss. Held-out evaluation reported accuracy, macro-F1, multiclass log loss, and a confusion matrix. Uncertainty was estimated with 2,000 source-cluster bootstrap resamples, resampling all four summaries from each source together. The primary incremental comparison was Model C versus Model A.

Robustness and reproducibility

Three planned sensitivity analyses addressed major design concerns: mixed-effects models were repeated with log word count as a covariate; audience-by-source-collection interactions tested PLOS versus PLABA pattern differences; and RQ4 was repeated after adding word count to the readability baseline and with an alternative readability implementation including numerical expressions. Preprocessing, feature

Table 2: Linguistic measures used in the primary analysis

Domain	Feature	Operational role
Lexical	HD-D	Lexical diversity
Lexical	General log frequency	General-English vocabulary commonness
Grammatical	Subordinators / 1,000 words	Subordinating relations
Grammatical	Coordinators / 1,000 words	Coordinating relations
Grammatical	Passives / 100 sentences	Passive construction frequency
Grammatical	PPs / 100 words	Prepositional-phrase elaboration
Information packaging	Noun density	Nominal informational load
Information packaging	Content-word density	Lexical information density
Information packaging	Complex NPs / 100 words	Internally complex noun phrases
Information packaging	Premodifiers / NP	Noun-phrase premodification
Cohesion	Discourse connectives / 1,000 words	Explicit discourse signaling
Cohesion	Adjacent argument overlap	Referential continuity

Note. HD-D = hypergeometric distribution D; NP = noun phrase; PP = prepositional phrase.

extraction, model fitting, sensitivity analyses, and table generation were scripted. The source PERCS corpus is available from the original authors' OSF repository (doi: [10.17605/OSF.IO/UMBJ7](https://doi.org/10.17605/OSF.IO/UMBJ7)). Supplementary File S1 provides the retained feature definitions, analysis specification, and computational record, including software information, for the present study. Raw PERCS text is not redistributed; the source corpus remains available from the original authors' OSF repository.

Results

Audience-level differences

All 12 primary linguistic features showed significant overall differences across audience conditions after FDR correction (all $q < 0.001$). Mean FRE declined from 56.37 for layperson summaries to 34.22, 30.87, and 16.69 across the premedical, researcher, and expert conditions; FKGL increased from 10.79 to 13.99, 14.58, and 16.78, respectively.

Linguistic differences and trajectories across domains

Information-packaging measures showed the most consistent directional progression. Noun density increased from 0.310 to 0.349, 0.356, and 0.397; the standardized adjacent changes were 0.79, 0.15, and 0.84, all Holm-significant. Content-word density increased from 0.626 to 0.651, 0.654, and 0.690, with the largest adjacent contrast at researcher-to-expert (0.91, 95% CI [0.84, 0.99]). Complex noun phrases rose from 5.26 to 6.41, 6.65, and 7.60 per 100 words, and premodifiers per noun phrase from 0.392 to 0.542, 0.562, and 0.692. All four measures met the strong stepwise-monotonicity criterion, although their categorical models fitted better than a simple linear progression.

General-English log frequency decreased from 1.601 to 1.338, 1.298, and 1.123, with the largest standardized

change at layperson-to-premedical (-1.11, 95% CI [-1.16, -1.06]); HD-D was non-monotonic. Subordinators declined from 39.27 to 25.75, 24.18, and 15.09 per 1,000 words, whereas coordinators increased from 34.10 to 37.02, 38.12, and 45.69. Passives and prepositional phrases were directional non-monotonic. Discourse-connective incidence was similar across the first three conditions before declining from 10.21 to 6.89 per 1,000 words at the expert transition. Adjacent argument overlap declined from 0.132 to 0.117, 0.115, and 0.110. Across the 12 features, eight met the estimated-monotonicity criterion and seven the stricter stepwise criterion; five were monotonic nonlinear, two early-threshold monotonic, one late-expert-threshold monotonic, and four directional non-monotonic. Table 3 summarizes the audience-level means and trajectory classifications for all 12 linguistic features.

Premedical and researcher conditions

The premedical-to-researcher transition was generally smaller than the outer transitions. Standardized contrasts were 0.15 for noun density, 0.09 for content-word density, 0.13 for complex noun phrases, 0.12 for premodification, -0.11 for subordinators, and 0.08 for coordinators. In the held-out combined model, 91.2% of layperson and 83.1% of expert summaries were correctly classified, compared with 60.1% of premedical and 38.5% of researcher summaries. Researcher summaries were classified as premedical in 43.9% of cases, while 26.4% of premedical summaries were classified as researcher.

Discrimination beyond conventional readability

On the untouched test partition, the readability-only model achieved accuracy = 0.625, macro-F1 = 0.610, and log loss = 0.804. The linguistic-only model produced accuracy = 0.628, macro-F1 = 0.623, and log loss = 0.778. The combined model performed best, with accuracy = 0.682, macro-F1 = 0.677, and log loss = 0.664. Relative to readability alone, the combined model increased accuracy by 0.057, 95% CI [0.017,

Table 3: Audience-level means and trajectory classifications

Feature	Lay	Premedical	Researcher	Expert	Trajectory
HD-D	0.822	0.820	0.819	0.826	Directional non-monotonic
General log frequency	1.601	1.338	1.298	1.123	Early-threshold monotonic
Subordinators / 1,000	39.27	25.75	24.18	15.09	Monotonic nonlinear
Coordinators / 1,000	34.10	37.02	38.12	45.69	Late-expert-threshold monotonic
Passives / 100 sentences	19.60	25.82	25.54	24.12	Directional non-monotonic
PPs / 100 words	15.96	16.25	16.17	15.25	Directional non-monotonic
Noun density	0.310	0.349	0.356	0.397	Monotonic nonlinear
Content-word density	0.626	0.651	0.654	0.690	Monotonic nonlinear
Complex NPs / 100 words	5.26	6.41	6.65	7.60	Monotonic nonlinear
Premodifiers / NP	0.392	0.542	0.562	0.692	Monotonic nonlinear
Discourse connectives / 1,000	10.70	10.10	10.21	6.89	Directional non-monotonic
Adjacent argument overlap	0.132	0.117	0.115	0.110	Early-threshold monotonic

Note. All 12 features had an FDR-corrected global audience effect of $q < 0.001$. Trajectory classifications follow the pre-specified criteria described under “Ordered trends and trajectory classification.” HD-D = hypergeometric distribution D; NP = noun phrase; PP = prepositional phrase.

Table 4: Held-out audience discrimination

Model / contrast	Accuracy	Macro-F1	Log loss
A: Readability	0.625	0.610	0.804
B: Linguistic	0.628	0.623	0.778
C: Combined	0.682	0.677	0.664
C - A	+0.057 [0.017, 0.100]	+0.067 [0.024, 0.111]	-0.139 [-0.186, -0.096]

Note. C - A intervals are 95% source-cluster bootstrap confidence intervals based on 2,000 resamples. Lower log loss is better. Displayed model metrics are rounded to three decimal places; the C - A contrasts were calculated from the unrounded estimates.

0.100], and macro-F1 by 0.067, 95% CI [0.024, 0.111], while log loss decreased by 0.139, 95% CI [-0.186, -0.096]. These held-out results are summarized in Table 4.

Robustness analyses

After log word count was added, all 12 primary features retained a significant audience effect after FDR correction. Three features showed FDR-significant audience-by-source interactions: general log frequency, noun density, and discourse-connective incidence. When word count was added to the readability baseline, adding linguistic features increased macro-F1 from 0.645 to 0.702 (gain 0.057, 95% CI [0.018, 0.096]), improved accuracy by 0.055, 95% CI [0.019, 0.091], and decreased log loss by 0.146, 95% CI [-0.193, -0.102]. The alternative readability calculation including numerical expressions produced the same substantive conclusion, with improvement in all three metrics and bootstrap intervals excluding zero.

Discussion

Audience adaptation extends beyond readability

The principal finding is that audience-conditioned biomedical rewriting in PERCS cannot be represented adequately by readability alone. All four linguistic domains varied across audience conditions, extending evidence that medical-text difficulty includes information not captured by surface word- and sentence-

length measures [10, 11]. The matched design adds source control: the biomedical abstract remains constant while the intended audience condition changes. The observed differences can therefore be examined as alternative linguistic realizations of the same source material, while remaining products of the PERCS prompt-and-review procedure. Importantly, the contribution is not simply to establish that the four prompted conditions differ. The analysis identifies which linguistic systems differentiate those conditions, where the largest transitions occur, which features fail to follow a simple directional progression, and whether this information adds to what conventional readability measures already capture.

The clearest convergence occurred in information packaging. Noun density, content-word density, complex noun phrases, and premodification increased across conditions. This pattern is consistent with corpus evidence that registers differ in how information is distributed across linguistic structures [13]. In PERCS, expert-oriented summaries increasingly concentrated information within nominal structures. Replacing uncommon vocabulary or shortening sentences could therefore improve a readability score while leaving specialist information packaging intact. Accessibility is better treated here as a profile of linguistic choices than as one number, but that profile is not a new comprehension scale.

Audience adaptation is structured but not uniformly linear

The trajectories challenge the idea that the four PERCS audience conditions can be represented adequately by one smooth linguistic progression. Several measures progressed directionally, but changes were uneven across transitions; others were non-monotonic. Subordination decreased with intended expertise, coordination increased, passives and prepositional phrases reversed, and discourse-connective use was stable through the first three conditions before falling at the expert endpoint. Adaptation toward less specialized audiences can therefore involve greater explicitness in some systems rather than a universal reduction of linguistic complexity. This supports the warning from PlainMedScale that measures distinguishing two registers need not behave as smooth scales across multiple levels [14]. PERCS demonstrates a related pattern under source matching.

The weak separation between premedical and non-medical researcher summaries further cautions against interpreting the four conditions as a simple calibrated expertise scale. Their overlap appeared across linguistic features and in held-out classification errors. Premedical students and researchers outside medicine plausibly represent different combinations of biomedical familiarity and broader scientific literacy rather than adjacent points separated by a known amount of expertise. Their positions in the planned 1–4 coding are therefore analytic rather than psychometric. The observed similarity may reflect how the prompts operationalized the two audiences, similarities in the communicative profiles encoded by those prompts, or both. Validation with actual members of these groups is required before attributing the pattern to reader knowledge or processing.

Linguistic information beyond readability

The held-out model comparison provides the clearest evidence for the beyond-readability claim. Readability alone was already informative, and the linguistic-only model did not dramatically outperform it. The important result was incremental: combining readability with the selected linguistic features improved audience-condition discrimination, and the advantage remained when word count entered the baseline. This extends evidence that fine-grained linguistic information contributes to medical readability and difficulty modeling [10, 11]. The classifier, however, predicts which PERCS audience condition is associated with a linguistic profile; it does not predict actual comprehension. This distinction matters because biomedical simplification has often been evaluated without direct lay-reader validation [12].

For biomedical communication and controlled text generation, a target reading grade is therefore not a complete specification of an audience. Vocabulary commonness, nominal compression, grammatical signaling, and cohesion may require separate evaluation because they change by different amounts

at different audience transitions. This complements persona-sensitive and controllable approaches [6–8]. The 12 measures used here are diagnostic rather than prescriptive and should not be treated as a universal accessibility checklist.

Limitations and future research

First, PERCS is a controlled synthetic, persona-conditioned corpus rather than naturally occurring audience-adapted discourse. GPT-4 generated initial persona-specific drafts, which physicians reviewed and revised [9]. Audience adaptation here therefore refers specifically to linguistic differentiation produced by that prompt-and-review procedure, not spontaneous accommodation by human writers or naturally authored discourse from the four audience groups.

Second, the persona prompts are part of the data-generating mechanism. They differed in requested terminology, explanation, sentence structure, and quantitative detail, and the expert condition had a lower maximum length [9]. Length-controlled analyses reduce the concern that findings are solely length-driven but cannot remove the broader influence of prompt design.

Third, the study measures textual properties rather than reader comprehension. No laypersons, premedical students, non-medical researchers, or clinicians evaluated the 1,984 summaries. The trajectories and predictive results therefore do not demonstrate comprehension, recall, perceived difficulty, or communicative effectiveness. Future studies should connect the linguistic profiles to comprehension accuracy, recall, reading time, perceived difficulty, and confidence, consistent with calls for evaluation with intended readers [12].

Fourth, the feature inventory was deliberately selective rather than exhaustive. Highly redundant measures were removed, and measures without defensible operationalization were not forced into the analysis. Additional work could incorporate independently validated biomedical terminology norms, conceptual-familiarity measures, or semantic-cohesion measures where reproducible.

Finally, external validity is limited by the English-language PLOS and PLABA source domain. The source-collection sensitivity analyses demonstrate robustness within these collections, not generalization beyond them. Future work should test independently authored patient materials, consent documents, medication instructions, clinical communication, other scientific genres, and other languages; PLABA and MultiMSD provide complementary settings for such validation [3, 5].

Conclusion

Using the four-way source-matched structure of PERCS, this study found that audience-conditioned biomedical rewriting varies across multiple linguistic domains rather than forming a single readability

gradient. Lexical characteristics, grammatical organization, information packaging, and cohesion varied across audience conditions, but their trajectories were uneven. Information packaging showed the clearest directional progression across audience conditions, several grammatical and cohesive features were non-linear or non-monotonic, and the premedical and researcher conditions overlapped substantially.

The selected linguistic features also contributed information beyond conventional readability in source-separated held-out prediction, including under length-controlled sensitivity analyses. The contribution is not a new comprehension scale but a differentiated account of which linguistic systems change and where those changes occur across intended audiences. Future work should validate these profiles with actual readers and naturally produced biomedical communication across additional genres and languages.

Acknowledgements

Not applicable.

Ethics Approval and Consent to Participate

Not applicable. This study involved secondary computational analysis of a publicly released textual corpus and recruited no new human participants or collected new human-participant data.

Funding

This research received no external funding.

Competing Interests

The authors declare no competing interests.

References

- [1] S. Stableford and W. Mettger, "Plain language: A strategic response to the health literacy challenge," *Journal of Public Health Policy*, vol. 28, no. 1, pp. 71–93, 2007, doi: [10.1057/palgrave.jphp.3200102](https://doi.org/10.1057/palgrave.jphp.3200102).
- [2] A. Bell, "Language style as audience design," *Language in Society*, vol. 13, no. 2, pp. 145–204, 1984, doi: [10.1017/S004740450001037X](https://doi.org/10.1017/S004740450001037X).
- [3] K. Attal, B. Ondov, and D. Demner-Fushman, "A dataset for plain language adaptation of biomedical abstracts," *Scientific Data*, vol. 10, Art. no. 8, 2023, doi: [10.1038/s41597-022-01920-3](https://doi.org/10.1038/s41597-022-01920-3).
- [4] C. Basu, R. Vasu, M. Yasunaga, and Q. Yang, "Med-EASi: Finely annotated dataset and models for controllable simplification of medical texts," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14093–14101, 2023, doi: [10.1609/aaai.v37i12.26649](https://doi.org/10.1609/aaai.v37i12.26649).
- [5] K. Horiguchi, T. Kajiwar, T. Ninomiya, S. Wakamiya, and E. Aramaki, "MultiMSD: A corpus for multilingual medical text simplification from online medical references," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 9248–9258, doi: [10.18653/v1/2025.findings-acl.481](https://doi.org/10.18653/v1/2025.findings-acl.481).
- [6] A. Mullick, S. Bose, R. Saha, A. Bhowmick, P. Goyal, N. Ganguly, P. Dey, and R. Kokku, "On the persona-based summarization of domain-specific documents," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 14291–14307, doi: [10.18653/v1/2024.findings-acl.849](https://doi.org/10.18653/v1/2024.findings-acl.849).
- [7] Z. Zhang, T. Goldsack, C. Scarton, and C. Lin, "ATLAS: Improving lay summarisation with attribute-based control," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, Short Papers, 2024, pp. 337–345, doi: [10.18653/v1/2024.acl-short.32](https://doi.org/10.18653/v1/2024.acl-short.32).
- [8] H. Tran, Z. Yao, L. Li, and H. Yu, "ReadCtrl: Personalizing text generation with readability-controlled instruction learning," in *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, 2025, pp. 19–36, doi: [10.18653/v1/2025.in2writing-1.3](https://doi.org/10.18653/v1/2025.in2writing-1.3).
- [9] R. C. Salvi, C. Chawla, D. Jain, S. Panigrahi, M. S. Akhtar, and S. Yadav, "PERCS: Persona-guided controllable biomedical summarization dataset," *arXiv preprint arXiv:2512.03340*, 2025, doi: [10.48550/arXiv.2512.03340](https://doi.org/10.48550/arXiv.2512.03340).
- [10] S. A. Crossley, R. Balyan, J. Liu, A. J. Karter, D. McNamara, and D. Schillinger, "Predicting the readability of physicians' secure messages to improve health communication using novel linguistic features: Findings from the ECLIPPSE study," *Journal of Communication in Healthcare*, vol. 13, no. 4, pp. 344–356, 2020, doi: [10.1080/17538068.2020.1822726](https://doi.org/10.1080/17538068.2020.1822726).
- [11] C. Jiang and W. Xu, "MedReadMe: A systematic study for fine-grained sentence readability in medical domain," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17293–17319, doi: [10.18653/v1/2024.emnlp-main.958](https://doi.org/10.18653/v1/2024.emnlp-main.958).
- [12] C. Krenn, C. Loder, N. Berger, K. Jeitler, T. Semlitsch, A. Siebenhofer, and D. Wilfling, "Automated approaches of text simplification of patient education materials: Scoping review," *Journal of Medical Internet Research*, vol. 28, Art. no. e88365, 2026, doi: [10.2196/88365](https://doi.org/10.2196/88365).
- [13] D. Biber, *Variation Across Speech and Writing*. Cambridge, U.K.: Cambridge University Press, 1988, doi: [10.1017/CBO9780511621024](https://doi.org/10.1017/CBO9780511621024).
- [14] B. Brocai, I. Papagno, and M. Ohta, "PlainMedScale: A corpus of multi-level simplified medical texts in German and English," *arXiv preprint arXiv:2608.01158*, 2026, doi: [10.48550/arXiv.2608.01158](https://doi.org/10.48550/arXiv.2608.01158).