SciGisPy: a Novel Metric for Biomedical Text Simplification via Gist Inference Score

Chen Lyu
University of Warwick
chen.lyu@warwick.ac.uk

Gabriele Pergola
University of Warwick
gabriele.pergola.1@warwick.ac.uk

Abstract

Biomedical literature is often written in highly specialized language, posing significant comprehension challenges for non-experts. Automatic text simplification (ATS) offers a solution by making such texts more accessible while preserving critical information. However, evaluating ATS for biomedical texts is still challenging due to the limitations of existing evaluation metrics. General-domain metrics like SARI, BLEU, and ROUGE focus on surface-level text features, and readability metrics like FKGL and ARI fail to account for domain-specific terminology or assess how well the simplified text conveys core meanings (*gist*). To address this, we introduce *SciGisPy*, a novel evaluation metric inspired by Gist Inference Score (GIS) from Fuzzy-Trace Theory (FTT). SciGisPy measures how well a simplified text facilitates the formation of abstract inferences (gist) necessary for comprehension, especially in the biomedical domain. We revise GIS for this purpose by introducing domain-specific enhancements, including semantic chunking, Information Content (IC) theory, and specialized embeddings, while removing unsuitable indices. Our experimental evaluation on the Cochrane biomedical text simplification dataset demonstrates that *SciGisPy* outperforms the original GIS formulation, with a significant increase in correctly identified simplified texts (84% versus 44.8%). The results and a thorough ablation study confirm that *SciGisPy* better captures the essential meaning of biomedical content, outperforming existing approaches.

1 Introduction

Biomedical literature is often written in highly specialized language, making it challenging for non-experts to understand. The 2022 World Health Organization (WHO) report identifies low public health literacy as a significant global issue, affecting disease prevention and management (Osborne et al., 2022). Automatic text simplification (ATS) offers a potential solution by transforming complex biomedical language into simpler, more accessible text while preserving essential details. However, evaluating the effectiveness of ATS on biomedical texts remains a challenge.

Complex Medical Abstract – ABS

Technical abstract: The pooled evidence derived from trials of gabapentin suggests that it is not efficacious for the prophylaxis of episodic migraine in adults. Since adverse events were common among the gabapentin-treated patients, it is advocated that gabapentin should not be used in routine clinical practice. Gabapentin enacarbil is not efficacious for the prophylaxis of episodic migraine in adults. There is no published evidence from controlled trials of pregabalin for the prophylaxis of episodic migraine in adults.
GisPy score: -0.417
SciGisPy score: -5.596

Plain-Language Summary - PLS

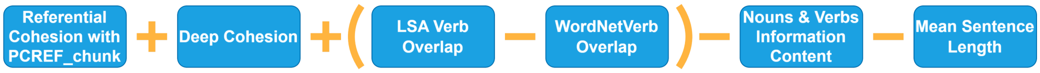
Plain-language summary: The studies showed that neither gabapentin nor gabapentin enacarbil was more effective than placebo at reducing the frequency of migraine headaches. Gabapentin commonly caused side effects, especially dizziness and somnolence (sleepiness). No studies of pregabalin were identified, and research on this drug is desirable.
GisPy score: 0.348
SciGisPy score: 3.599

Figure 1: An example of excerpts from a technical abstract (top) and its corresponding plain-language summary (bottom) from the Cochrane text simplification dataset (Devaraj et al., 2021). SciGisPy demonstrates better ability in distinguishing between ABS and PLS.

Existing metrics for biomedical text simplification are still limited. General-domain ATS metrics, such as SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004), focus on surface-level word edits and n-gram overlaps, relying heavily on the quality and variety of reference texts. Similarly, referenceless metrics, like BERTScore (Zhang et al., 2019a), Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), and Automated Readability Index (ARI) (Senter and Smith, 1967), focus on syntactic and lexical simplicity (e.g., shortening sentences, simplifying vocabulary), but fail to capture domain-specific terminology, and more importantly, they cannot ensure that the core meaning (or *gist*) is easily understood.

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 95–106
November 15, 2024

(a) Original Gist Inference Score (GIS) formula by (Wolfe et al., 2019)



(b) Enhanced GIS formula for Biomedical Text Simplification: SciGisPy

Figure 2: Original and enhanced GIS formula

These limitations are especially critical for biomedical texts because they do not provide an adequate measure of whether the text is effective in facilitating comprehension of complex medical concepts despite its linguistic simplicity.

To address this gap, we propose a novel, task-specific evaluation score, SciGisPy, based on the Gist Inference Score (GIS), inspired by the Fuzzy-Trace Theory (FTT). FTT posits that human cognition operates through two parallel representations: gist (the essential meaning) and verbatim (exact details) (Reyna, 2021). GIS measures how effectively a text conveys this *gist*, supporting decision-making. While previous GIS formulations have been explored in general domains (Hosseini et al., 2022), none have been optimized for the complexities of biomedical documents.

In this work, we introduce SciGisPy, the first GIS formulation for biomedical text. Figure 2 shows the original GIS formula and our enhanced version, which incorporates domain-specific adaptations. These include new indices and improvements, based on semantic-based chunking, Information Content (IC) theory, specialised embeddings, and an overall revision of the original GIS formulation.

Our contribution can be summarized as follows:

- We introduce SciGisPy, a novel GIS formulation specifically designed for biomedical text simplification, revising existing indices and eliminating those unsuitable for the biomedical domain.

- We introduce newly designed indices for SciGisPy, based on semantic-driven chunking, Information Content (IC) theory, and linguistic features of biomedical sentences.

- We conduct a comprehensive experimental evaluation of GIS as a metric for biomedical text simplification, analyzing the relevance of each index and its correlation with established simplification metrics.

2 Related Work

Text Simplification Metrics The development of automatic evaluation metrics tailored specifically for biomedical ATS remains under-explored. Due to the scarcity of specialized metrics, existing studies often rely on general-domain ATS metrics, which are insufficient for capturing the characteristics of biomedical text. Reference-based metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), compare simplified outputs to human-generated references, focusing on n-gram precision and recall. These metrics heavily depend on the quality of reference texts and may penalize valid simplifications that use different wording (Pergola et al., 2019, 2021a; Zhu et al., 2021). SARI (Xu et al., 2016), designed for text simplification, evaluates word-level edits but similarly focuses on surface-level features like n-gram overlaps, missing the deeper semantic aspects crucial for biomedical comprehension (Sulem et al., 2018; Pergola et al., 2021b; Alva-Manchego et al., 2021). BERTScore (Zhang et al., 2019a), leveraging contextual embeddings from BERT, improves semantic similarity evaluation but still underperforms in biomedical contexts (Sun et al., 2022; Zhu et al., 2022, 2023; Lu et al., 2023).

Readability metrics, such as Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Automated Readability Index (ARI) (Senter and Smith, 1967), are reference-less and rely on surface features like sentence length and word complexity. However, these metrics do not account for the accurate use of domain-specific terminology or semantic nuances critical to biomedical texts. Consequently, they often fail to effectively evaluate the readability and accuracy of simplified medical content, underscoring the need for more advanced evaluation methods in this domain.

Gist and GIS According to Fuzzy-Trace Theory (FTT) (Reyna, 2021), individuals encode multiple mental representations when processing text, ranging from verbatim, which captures surface-level details, to gist, which conveys the core meaning. In FTT, "gist" refers to the essential idea of a matter. Prior research (Reyna, 2021) suggests that gist representations significantly influence decision-making processes more than verbatim representations. Therefore, assessing gist representation can help measure a document's ability to generate clear and actionable mental models and effectively communicate its message.

The Gist Inference Score (GIS) was first introduced by Wolfe et al. (Wolfe et al., 2019) to evaluate how well a text enables readers to form gist inferences. Before the development of the GisPy library (Hosseini et al., 2022), which automated GIS evaluation, research on GIS was still developing (Reyna, 2021; Wolfe et al., 2019). GIS was initially proposed by leveraging Coh-Metrix, a multilevel linguistic framework analyzing over 100 variables related to text simplicity, such as referential cohesion, lexical diversity, and latent semantic analysis (LSA) (Wolfe et al., 2019; Graesser et al., 2011; Sun et al., 2024). However, Coh-Metrix lacks batch processing and efficiency. The GisPy library, building on these earlier methods and leveraging advanced NLP techniques, provides the first open-source solution for computing GIS across multiple documents. In GisPy (Hosseini et al., 2022), GIS is composed of seven indices: Referential Cohesion, Deep Cohesion, Verb Overlap, Word Concreteness, Word Imageability, and Hypernymy Nouns & Verbs, each associated with either a positive or negative coefficient. This work extends the GisPy library by modifying, removing, and adding to these indices for better alignment with biomedical simplification tasks.

2.1 GisPy and Other Text Simplification Metrics

GIS, as a reference-less metric, is more closely related and comparable to the readability metrics, such as FKGL and ARI. However, we argue that GIS captures different information from the text compared to FKGL and ARI.

While FKGL and ARI focus on surface-level "verbatim" features of text, GIS aims to measure the likelihood that readers will develop meaningful "gist inferences" from the text. Specifically, FKGL assesses readability based solely on surface-level features using sentence length and word syllable count, whereas GIS captures the underlying abstract meaning by considering more complex dimensions of text features such as cohesion and word concreteness. To validate this argument, we calculated the Pearson correlation coefficients between the reference-less metrics shown in A.3, where the results are all close to zero, proving that GIS is uncorrelated with these metrics.

3 Method

In this section, we first review the indices in the original GisPy formulation (Hosseini et al., 2022) and assess their suitability for biomedical text simplification. For each index, we propose adaptations by either (i) introducing novel approaches, (ii) improving the existing indices with more specialized methods, or (iii) removing them if unsuitable for the biomedical domain.

3.1 Enhancing GIS indices for Biomedical Document Simplification

As shown in Figure 2, the original GIS formula includes seven indices, with some positively and others negatively weighted. These indices cover five dimensions of text features. Through analysis, we posit that only four dimensions are beneficial for evaluating biomedical text simplicity – *Referential Cohesion, Deep Cohesion, Verb Overlap, Hypernymy Nouns & Verbs*, while *Word Concreteness and Imageability* is not.

Hypernymy Nouns & Verbs: This index measures word specificity, based on the idea that more specific words are harder to understand for a general audience without specialized knowledge. Simplifying biomedical texts often requires translating technical terms into concepts that are accessible to a broader audience, thus making this metric valuable for evaluating simplified biomedical documents.

To achieve this, the index (`WRDHYPnv`) uses WordNet's hierarchy of concepts and penalizes words with greater depth in the hierarchy, as these represent more specialized terms. In particular, the specificity is quantified by the average hypernym path length of synonym sets.

In the original GIS formula, this index evaluates word specificity by listing all nouns and verbs in a document, identifying their synonym sets in WordNet, and calculating the average hypernym path length. Instead, we propose three more fine-grained alternatives to address limitations when applied to specialized texts: (i) the first ensures proper comparison of noun and verb paths via normalisation (`WRDHYP_norm`), (ii) the second introduces and adapts the concept of Information Content `WRDIC`, and (iii) the third resolves a development issue found in the original GisPy library.

i. Hypernym Root Normalisation: In WordNet, unlike noun synsets that all trace back to the hypernym root '*entity*', verb synsets can trace back to different hypernym roots, with some synsets having multiple roots. For example, in the biomedical domain, the verb *administer* could trace back to *apply* (in the context of giving treatment) or *manage* (in the context of overseeing care). Consequently, the GisPy approach of averaging hypernym paths for all synsets can lead to incomparable path lengths if the roots are different, as these roots may have hypernym hierarchies of varying scales.

To address this issue, we propose an alternative approach, where instead of averaging all hypernym paths, we group the paths that lead to the same root and apply L1 normalization within each group to balance the scales of the hypernym hierarchies. For synsets with multiple hypernym roots, we select the longest hypernym path and its corresponding root as the representative. Finally, we compute the average of the normalized path lengths across all groups to obtain the final result. We indicate this new index with `WRDHYP_norm`, where the suffix stands for "*root normalization*", formalised as follows:

$$\texttt{WRDHYP\_norm} = \frac{1}{n}\sum_{i=1}^{n}\frac{L_i}{||L_i||_1}$$

Where L_i is the path length for the i-th hypernym path group, $||L_i||_1$ is the L1 normalization of the path length for each root group, and n is the total number of hypernym path groups.

ii. Information Content: To improve the GIS metric for biomedical text simplification, we propose to replace the Wordnet hypernym-based solution `WRDHYP` with a new approach based on Information Content (IC) (Cover and Thomas, 2006), namely `WRDIC`. Simple hypernym path counting can be insufficient in some cases, as it fails to account for the frequency and relevance of terms within specific domains. For instance, two biomedical terms may have the same path length but differ significantly in importance or specificity within a corpus. Information Content addresses this issue by considering the probability of encountering a term in a given corpus. In information theory, IC is a measure derived from the probability of a specific event occurring from a random variable (Cover and Thomas, 2006). In this context, the IC of a word can be defined as $-\log(P(c))$, where $P(c)$ is the probability of encountering a hypernym of word c in a corpus.

To compute the new index, similar to the strategy in `WRDHYP`, we first identify all nouns and verbs in the text. Then, we calculate the average of their IC values to generate the final result. IC provides a more accurate measure of word specificity, with higher IC values indicating more specialized words. This approach enhances the ability to measure text simplification by considering both the structure of the language and its actual use in the domain, making it particularly suitable for specialized domains.

iii. Mean Hypernym Paths Length: The original GisPy score (Hosseini et al., 2022) uses WordNet (Miller, 1995) and its Synset objects from the NLTK library[1] to compute hierarchical paths (`WRDHYP`). However, for some verb synsets, there are multiple hypernym roots, resulting in several hypernym paths leading to different roots. This issue is critical because having different roots makes the hypernym paths non-comparable, as the hierarchical structures vary in depth and scope.

The original GisPy paper assumes a single hypernym path per root and does not account for this issue. We addressed this by modifying the index to use the mean length of all available hypernym paths for a given synset when multiple paths are available, which we call `WRDHYP_mean` [2].

Verb Overlap: According to the FFT, abstract verb overlaps promote the formation of gist representations, aiding readers in understanding the text's core meaning. To capture this, the GisPy score uses two indices: `SMCAUSe` (positively weighted) and `SMCAUSwn` (negatively weighted) (Hosseini et al., 2022). For biomedical text simplification, `SMCAUSe` is important because it promotes simplicity by emphasizing abstract overlaps between verbs, while `SMCAUSwn` penalizes the redundant repetition of identical or similar verbs.

In its original implementation, this index is based

[1] https://www.nltk.org

[2] We flagged the issue regarding multiple hypernym paths for verb synsets on the GisPy GitHub repository. The authors implemented a solution using the maximum path length, but our preliminary experiments indicated that averaging the path lengths offers a more balanced measure of specificity.

on the *en_core_web_trf*[3] from SpaCy, a RoBERTa-based pre-trained language model (Liu et al., 2021), to generate token vector embeddings for each verb, and then computes cosine similarity of the embeddings. To better suit the characteristics of biomedical texts, often featuring technical and compound terminology, we propose a simple yet effective modification, adopting embedding models specialised for technical documents. We identify two embedding models for this index, fastText (Bojanowski et al., 2017) and BioWordVec (Zhang et al., 2019b).

FastText (Bojanowski et al., 2017) is a widely used word embedding library, particularly suitable for technical documents. It learns word embeddings on a sub-word basis, which allows it to represent out-of-vocabulary words. This is particularly useful for dealing with biomedical language characterized by many compound words (Pergola et al., 2018). We adopt pre-trained embeddings provided by fastText and name this index `SMCAUSf`.

Our second alternative is BioWordVec (Zhang et al., 2019b), based on a benchmark biomedical word embedding library. BioWordVec combines subword information from unlabeled biomedical text with the widely-used Biomedical Subject Headings (MeSH) vocabulary (Lipscomb, 2000). Pre-trained using FastText embeddings, BioWordVec is the most commonly used biomedical word embedding model in the recent literature. We adopt it to improve the `SMCAUS` index, and name it `SMCAUSb`.

Referential Cohesion: Referential Cohesion measures word and idea overlaps across sentences, making it a suitable metric to characterise simplicity in the biomedical text. In Hosseini et al. (2022) this dimensions is captured with two indices: `PCREF` and `CoREF` (both positively weighted). `PCREF` calculates cosine similarity between sentence embeddings, while the `CoREF` focuses on coreference resolution across sentences. A high overlap in both indices typically indicates that the text maintains consistent ideas and vocabulary, which helps readers follow complex biomedical content more easily, thus promoting text simplicity. To improve the detection of referential cohesion in biomedical texts, we introduce (i) a novel index based on *semantic chunking*, which posits that a lower number of semantic chunks indicates stronger coherence, and (ii) more suitable sentence embedding models designed for technical and biomedical documents.

i. Semantic Chunking: We introduce an alternative solution for measuring Referential Cohesion to substitute `PCREF`. This new approach is based on the concept of *semantic chunking*[4]. Unlike traditional methods that chunk text using a fixed size, semantic chunking adaptively determines breakpoints between sentences based on embedding similarity of customizable window size. This ensures that each chunk contains sentences that are semantically related. Similar to `PCREF`, the semantic chunking method uses cosine similarity between sentences to represent overlap across sentences.

We argue that a higher number of chunks indicates more diverse semantics and topics within the text. Therefore, minimizing the number of chunks ensures textual coherence and enhances simplicity. Inspired by this, we designed a new index, `PCREF_chunk`, built using a semantic chunker to replace the original `PCREF`. We selected BioSimCSE (Kanakarajan et al., 2022) as the sentence embedding model, as its biomedical-domain embeddings capture semantics more accurately. We apply a negative coefficient to this index, indicating that fewer semantic chunks correspond to higher coherence and simplicity.

ii. Specialized Sentence Embeddings The original index *PCREF* calculates cosine similarity between sentences using the pretrained MPNet model [5] from Hugging Face as the sentence embedding model.

We experimented with five state-of-the-art sentence embedding models. First, we adopted two leading general-purpose models from the Massive Text Embedding Benchmark (MTEB) Leaderboard mxbai-embed-large-v1 (Li and Li, 2023) and the e5-mistral-7b-instruct (Jiang et al., 2023; Wang et al., 2023, 2022) embedding models. Additionally, we utilized three state-of-the-art biomedical domain embedding models based on BERT: BioSimCSE (Kanakarajan et al., 2022) and BioBERT (Lee et al., 2020), which generate contextual embeddings, and a context-free embedding model, BioSentVec (Chen et al., 2019). These models are known for their robustness in biomedical text processing.

Detailed implementation information is provided in Section 4, where a thorough ablation study highlights the impact of each of them. Each model

[3] `https://spacy.io/models/en#en_core_web_trf`

[4] LlamaIndex Semantic Chunking Documentation

[5] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

is indicated by a suffix to the index name (e.g., `PCREF_mxbai`) specifying the embedding used.

Deep Cohesion: Indicated by the *PCDC* index (positively weighted), it measures the extent to which a text uses causal and intentional connectives, detected using regular expression patterns. The index is still highly relevant to biomedical text comprehension, as it supports logical relationships between sentences, crucial for ensuring that readers can follow complex biomedical information. Therefore, we retain the original design of this index as in Hosseini et al. (2022).

Mean Sentence Length: FTT suggests that people extract both verbatim (exact details) and gist (core meaning) from texts. Longer or more complex sentences may increase cognitive load, making it harder to focus on gist and potentially promoting reliance on verbatim processing. In contrast, shorter sentences with clear structures could help readers extract gist more easily because the underlying meaning is more accessible. Research in readability and health communication has shown that shorter sentences enhance readability by making information more accessible (Rudd et al., 2023), reducing cognitive load (Graesser et al., 2011), improving comprehension (National Institutes of Health, 2012), and maintaining consistency and focus (Weiss, 2007).

To address this gap within the original GIST score, we propose a new composite index called Mean Sentence Length (`MSL`) . This index, rewards the reduction of the average sentence length. Concretely, we calculate the mean sentence length by counting the number of words in each sentence and averaging these counts across the entire text. Despite its simplicity, preliminary exploration on this index showed promising results, with a more detailed assessment presented in Section 4.

3.2 Removing Word Concreteness and Imageability

Unlike the previous indices, the dimension we find potentially detrimental to representing biomedical text simplicity is *Word Concreteness and Imageability*. In the original GisPy score, *Word Concreteness* (`PCCNCz`) measures how concrete and image-evoking words are, while *Imageability* (`WRDIMGc`) indicates how easily a word can evoke a mental image. For instance, high imageability words like "hammer" are more specific and easily visualized compared to low imageability words like "reason". These indices are negatively weighted in the GIS formula, suggesting its promotion of abstractness. However, we argue that words with high concreteness and imageability (such as “heart”) help understand scientific and biomedical texts, are easier to visualize and thus improve comprehension and make it easier to follow for diverse audiences. Therefore, we hypothesize that removing this concreteness-penalizing index from GIS formula could enhance its performance in biomedical text simplification task.

In conclusion, while the GIS formula promotes abstractness in text to generate Gist, this may not align with promoting simplicity, especially in biomedical texts. Relevant indices may need to be modified or removed to better evaluate simplicity in biomedical documents.

4 Experiments

In this section, we present an experimental evaluation to assess the effectiveness of the GIS metric and our proposed index enhancements in the biomedical domain, using a Cochrane Library dataset (Devaraj et al., 2021) containing technical documents paired with simplified versions, firstly detailed in Section 4.1. In Section 4.2, we report the results of our evaluation on this dataset, exploring the impact of different index combinations on simplified texts. Specifically, we analyze which index combinations produce the most significant improvements in gist abstraction by measuring the GIS differences between the technical and simplified documents. Finally, we assess the generalization of our findings by testing on several benchmark datasets from previous GIS literature to evaluate how well our GIS enhancements can be applied beyond the biomedical domain.

4.1 Datasets

We conducted GIS analysis and tested our indices enhancement on the Cochrane paragraph-level biomedical text simplification dataset (Devaraj et al., 2021), which is sourced from the Cochrane library[6] of systematic reviews. The Cochrane text simplification dataset comprises 4,459 parallel pairs of technical abstracts (ABS) and their plain-language summaries (PLS) crafted by domain experts, where the PLS texts are simplified versions of original technical abstracts. Figure 1 presents a sample excerpt of a technical abstract

[6]https://www.cochranelibrary.com/

Index		Mean GIS Diff	% + GIS Diff	% inc GIS Diff	% - to +
Original GisPy (Hosseini et al., 2022)		-0.461	43.38%	N/A	N/A
Without *PCCNC* & *WRDIMG*		-0.022	49.91%	60.94%	**12.87%**
Hypernymy Nouns & Verbs	*WRDHYP_mean*	0.158	52.85%	74.54%	11.58%
	WRDHYP_norm	-1.355	31.25%	17.56%	1.27%
	WRDIC	0.211	51.38%	**78.77%**	9.74%
Verb Overlap	*SMCAUSe*	-0.728	39.25%	41.36%	3.95%
	SMCAUSb	-0.678	40.53%	43.38%	4.78%
Referential Cohesion	*PCREF_mxbai*	-0.328	45.59%	61.12%	4.41%
	PCREF_BioSimCSE	-0.629	40.53%	37.87%	2.48%
	PCREF_BioBERT	-0.655	40.99%	43.75%	3.68%
	PCREF_mistral	-0.503	44.39%	53.22%	3.86%
	PCREF_BioSentVec	-0.686	40.99%	37.04%	2.57%
	PCREF_chunk	**0.418**	**54.32%**	61.12%	5.97%
Mean Sentence Length (MSL)		0.321	52.94%	76.93%	11.40%
SciGisPy (Our)		**2.312**	**85.39%**	**85.57%**	**44.21%**

Table 1: Results of GIS Enhancements on Cochrane simplification development set. The second column displays the average GIS difference across the entire dataset. The third column indicates the percentage of documents with a positive GIS difference. The fourth column shows the percentage of documents where the GIS difference increased following the enhancement. The fifth column reports the percentage of documents that initially had a negative GIS difference but shifted to a positive value.

and its corresponding PLS. Since GIS score computation does not require any training, we sample 4,334 document pairs as our *development set* to determine the best index configurations; while the remaining additional subset, used as *test set*, will be introduced in the following sections. Since SciGisPy does not involve any training, no training set is required.

In previous literature, three benchmark general-domain datasets have been used for evaluating GIS metrics: News Reports vs. Editorials, Journal Article Methods vs. Discussion, and Disneyland Measles Outbreak Data (Wolfe et al., 2019; Hosseini et al., 2022). This subset serves as our *test set* to assess the generalisation of our biomedical-specialized GIS.

4.2 Results and Discussion

To investigate the effectiveness of applying GisPy GIS in evaluating the simplicity of biomedical text, we first computed GIS values for all Abstracts (ABS)s and Plain Language Summaries (PLS)s in the development set using the best configuration reported for the GisPy library, following the evaluation process outlined in Hosseini et al. (2022). After we obtained the GIS score for all documents, we calculated the *GIS difference* between each pair of ABS and PLS:

$$\text{GIS Difference} = \text{GIS}_{\text{PLS}} - \text{GIS}_{\text{ABS}}$$

In the rest part of this paper, "GIS difference" refers to the difference calculated using the above equation. A *positive GIS difference* for a pair of documents suggests that audiences will more easily abstract the gist from the simplified text (PLS) compared to the original text (ABS), subsequently showing that GIS can be a good indicator of simplification. Consistent with previous literature, we compare the average GIS difference among all documents under the different GIS formulations to determine the more effective alternatives, namely mean GIS difference.

To evaluate the impact of each individual enhancement, we ran GisPy with each enhancement applied separately, while keeping all other indices identical to those in the original formula. Additionally, when testing combinations of enhancements that modify the same index, for those that modify the same index, we ensured that only one change was applied at a time to prevent overlapping calculations.

4.2.1 GIS for Biomedical Text Simplification

First, we report the GIS scores resulting from the original GisPy on the development set of the Cochrane Simplification Dataset in Table 1. The average GIS for ABS texts is 0.225, while for PLS texts, the mean GIS is -0.225, resulting in a mean GIS difference of -0.450; only 43% of document pairs have a positive GIS difference. These results show that the original GIS formulation struggles to distinguish between simplified and unsimplified

Index	Mean GIS Diff	% + GIS Diff	% inc GIS Diff	% - to +
Original GisPy (Hosseini et al., 2022)	-0.311	44.8%	N/A	N/A
SciGisPy (Our)	2.295	84%	79.2%	45.6%

Table 2: Results of GIS Enhancements on Cochrane simplification test set. See Table 1 for column descriptions.

texts for most biomedical documents.

We proceed to discuss the impact of the enhancements proposed in this work; for each index enhancement listed, Table 1 reports the results obtained by running GisPy with only the corresponding enhancements while keeping the rest of the formula unchanged.

Removing Word Concreteness and Imageability: To test our hypothesis that removing word concreteness and imageability promotes biomedical text easier to comprehend (Sec. 3.2), we ran GIS without *PCCNC* and *WRDIMG* and observed positive results, as reported in Table 1: the average GIS difference increased from -0.450 to -0.022, with **49.91%** of documents now exhibiting a positive GIS difference, compared to 43.38% with the original GIS formula. This finding supports our initial analysis and confirms the need to tailor the roles of the indices when dealing with specialised domains.

Semantic Chunking: As mentioned earlier, the mean GIS difference is our primary metric for evaluating the performance of new GIS formula. A larger difference indicates better distinction between simplified and original documents. Based on the experiment results shown in Table 1, most of our enhancements produced positive outcomes. The enhancement *PCREF_chunk* achieved the most significant improvement, leading to the mean GIS difference increase from -0.461 to 0.418. This enhancement also led to 54.32% of documents obtaining a positive GIS difference, an increase of 10.94% compared to the original GisPy, which achieved 43.38%.

In addition, we tracked the impact of each enhancement on individual documents. The fourth column in Table 1 presents the percentage of documents where the GIS difference increased after the enhancement. This indicates that the enhanced GIS formula can better distinguish between the original ABS text and the simplified PLS text compared to the original GIS. The table's last column also shows the percentage of documents that originally had a negative GIS difference but switched to positive; this represents cases where the original GIS failed to evaluate simplicity, but the new GIS succeeded.

Looking at ABS-PLS pairs in Table 1, more than half of our indices enhancements yielded positive results. Some indices demonstrated significant improvements, with *WRDIC* by achieving the highest increase with 78.77% of documents in the development set transiting to a positive GIS difference.

Best Formulation: Based on the experimental results on the development set, we identified the best combination of our enhanced GIS formula, as shown in Figure 2. We adopted the enhancements of Referential Cohesion with Semantic Chunking (`PCREF_chunk`), Hypernyms with Information Content (IC) (`WRDIC`), and Mean Sentence Length (`MSL`), together with the removal of indices `PCCNC` and `WRDIMG`. The significant results of this biomedical text simplification-targeted GIS formula are presented in the last row of Table 1.

Generalisation: To test the generalisation of this finding, we also applied the enhanced formula to the Cochrane test set. The results, presented in the last row of Table 2, demonstrate a significant improvement, with the new GIS successfully identifying 84% of simplified texts, doubling the original number. This confirms the effectiveness of our new GIS for evaluating biomedical text simplification.

4.2.2 Gist Inference Benchmarks

To assess whether our enhancements improve the evaluation of Gist abstraction in the general domain, the original objective of GIS, we tested all index enhancements on the benchmark datasets used in the original GisPy paper. The News Reports vs. Editorials dataset comprises 50 pairs of documents per category, totaling 100 documents. The Journal Article Methods vs. Discussion dataset includes 25 pairs, amounting to 50 documents. The Disneyland dataset consists of 191 articles in total. To ensure comparability with these datasets, we randomly sampled 125 document pairs from the Cochrane dataset.

The experimental results were less significant compared to the previous results on the Cochrane simplification dataset since our enhancements were targeted at biomedical text simplification. However, we still identified a combination of index en-

Benchmark	Approach	Distance	t-statistic	p-value
Reports vs. Editorials	GisPy with *PCREF_mistral* & *MSL*	**3.260**	4.068	$* 2 \times 10^{-4}$
	GisPy (Hosseini et al., 2022)	2.551	3.643	$* 7 \times 10^{-4}$
	Coh-Metrix (Graesser et al., 2011)	2.535	3.826	$* 3 \times 10^{-4}$
	(Wolfe et al., 2019)	0.368	-	-
Methods vs. Discussion	GisPy with *SCAUSf*	**5.200**	5.916	$* 3 \times 10^{-7}$
	GisPy	5.012	7.188	$* 3 \times 10^{-9}$
	Coh-Metrix	5.010	6.331	$* 7 \times 10^{-8}$
	(Wolfe et al., 2019)	0.747	-	-
Disney	GisPy with *MSL*	**2.442**	3.492	$* 6 \times 10^{-4}$
	GisPy	2.418	3.440	$* 7 \times 10^{-4}$
	Coh-Metrix	0.998	1.878	6×10^{-2}

Table 3: Comparison of GIS scores generated by GisPy with our enhancement indices vs. original GisPy vs. other methods for all benchmarks

hancements that outperformed the original GisPy formula on the benchmark dataset. The results are presented in Table 3, where we also performed a student's t-test with the null hypothesis following GisPy (Hosseini et al., 2022) paper, which shows how good a GIS score can significantly distinguish these ABS texts and PLS texts. This positive result demonstrates that our proposed solutions are not only beneficial for simplification evaluation, but also enhance the measure of how easily GIS can be inferred.

5 Conclusion

In this study, we addressed the challenge of evaluating biomedical automatic text simplification by introducing a novel referenceless evaluation metric, SciGisPy, inspired by the Gist Inference Score (GIS) from Fuzzy-Trace Theory. This metric was specifically adapted and enhanced for biomedical text simplification through rigorous feasibility analysis and domain-specific enhancements. Our comprehensive experimental assessment on the Cochrane text simplification dataset demonstrates that SciGisPy significantly outperforms the original GIS metric in assessing the simplicity of biomedical texts.

6 Limitations

A limitation of this study is the reliance on a single benchmark, the Cochrane simplification dataset, due to the limited availability and suitability of biomedical text simplification datasets at the document level. Validating our methodology across multiple datasets would strengthen its robustness.

Additionally, while we introduced several improvements to the individual GIS indices, the coefficient magnitudes currently remain fixed at 1. Developing an automated method to dynamically adjust these coefficients based on text distributions could further improve the accuracy and versatility of SciGisPy in text simplification.

Lay Summary

Medical research papers are often written in very complex and technical language, which makes it difficult for non-experts to understand. To solve this problem, automatic text simplification (ATS) systems try to rewrite these texts in a simpler way while keeping the important information intact. However, it's hard to evaluate how well these systems simplify medical texts because current tools focus too much on the surface details, like word counts and sentence length, without considering whether the text still conveys the core meaning (the *gist*).

In this study, the researchers developed a new evaluation tool called SciGisPy, designed specifically to measure how well simplified medical texts communicate the essential meaning. It builds on an existing concept called the *Gist Inference Score* (GIS), which measures how easily a reader can understand the gist of a text. SciGisPy adds new features like focusing on medical terms, simplifying complex sentences, and improving coherence between ideas. The study shows that SciGisPy significantly improves the evaluation of simplified medical texts compared to existing methods, helping to make complex medical information more accessible to a broader audience.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Pedram Hosseini, Christopher Wolfe, Mona Diab, and David Broniatowski. 2022. GisPy: A tool for measuring gist inference score in text. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 38–46, Seattle, United States. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kamal Raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. Biosimcse: Biomedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

C.E. Lipscomb. 2000. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

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.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

National Institutes of Health. 2012. Clear communication: An NIH health literacy initiative. *http://www. nih. gov/clearcommunication/healtHealth Literacyiteracy. htm*.

Richard Osborne, Shandell Elmer, Melanie Hawkins, Christina Cheng, Roy Batterham, Sónia Dias, Suvajee Good, Maristela Monteiro, Bente Mikkelsen, Ranjit Nadarajah, and Guy Fones. 2022. Health literacy development is central to the prevention and control of non-communicable diseases. *BMJ Global Health*, 7.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Gabriele Pergola, Lin Gui, and Yulan He. 2019. TDAM: A topic-dependent attention model for sentiment analysis. *Information Processing & Management*, 56(6):102084.

Gabriele Pergola, Lin Gui, and Yulan He. 2021a. A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2870–2883, Online. Association for Computational Linguistics.

Gabriele Pergola, Yulan He, and David Lowe. 2018. Topical phrase extraction from clinical reports by incorporating both local and global context. In *The 2nd AAAI Workshop on Health Intelligence*, pages 499–506. Association for the Advancement of Artificial

Intelligence. 2018 Workshop on Health Intelligence (W3PHIAI 2018).

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021b. Boosting low-resource biomedical QA via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985, Online. Association for Computational Linguistics.

Valerie F Reyna. 2021. A scientific theory of gist communication and misinformation resistance, with implications for health, education, and policy. *Proceedings of the National Academy of Sciences*, 118(15):e1912441117.

Rima E Rudd, Jennie Epstein Anderson, Sarah Oppenheimer, and Charlotte Nath. 2023. Health literacy: an update of medical and public health literature. In *Review of Adult Learning and Literacy, Volume 7*, pages 175–204. Routledge.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Technical report, DTIC document.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhaoyue Sun, Gabriele Pergola, Byron Wallace, and Yulan He. 2024. Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–357, St. Julian's, Malta. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Barry D Weiss. 2007. *Health literacy and patient safety: Help patients understand. Manual for clinicians*. American Medical Association Foundation.

Christopher R Wolfe, Mitchell Dandignac, and Valerie F Reyna. 2019. A theoretically motivated method for automatically evaluating texts for gist inferences. *Behavior research methods*, 51:2419–2437.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019b. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.

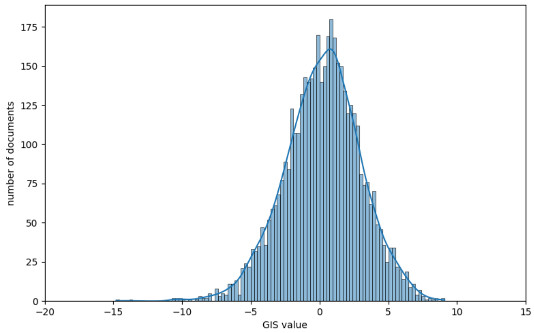
Lixing Zhu, Zheng Fang, Gabriele Pergola, Robert Procter, and Yulan He. 2022. Disentangled learning of stance and aspect topics for vaccine attitude detection in social media. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1580, Seattle, United States. Association for Computational Linguistics.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

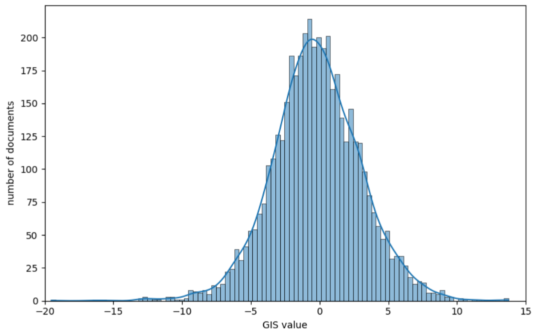
Lixing Zhu, Runcong Zhao, Gabriele Pergola, and Yulan He. 2023. Disentangling aspect and stance via a Siamese autoencoder for aspect clustering of vaccination opinions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1827–1842, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Original GIS distribution on Cochrane dataset



(a) GIS distribution for ABS



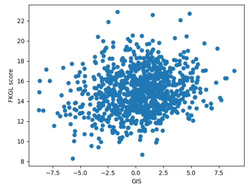
(b) GIS distribution for PLS

Figure A1: GIS distribution histogram and KDE for Cochrane text simplification dataset

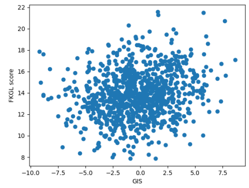
The GIS distributions of ABS and PLS are jointly shown in Figure A1. Both distributions resemble Gaussian distributions, since all indices in GIS were transformed into z-scores, which were subsequently summed up with coefficients to GIS.

A.2 GIS correlation with other TS metrics

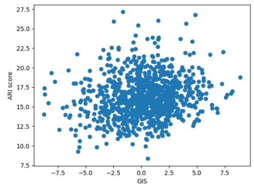
This is initially illustrated in Figure A2, where ABS and PLS from a subset of Cochrane simplification dataset (1000 samples) are plotted on corresponding scatter plots, with GIS on the vertical axis and the respective text simplification metric on the horizontal axis. Here we sampled 1000 documents from the development set due to the difficulty to visualize the original large amount of data. If there were a correlation, the points would roughly form a line, however this is not observed in any of the plots.



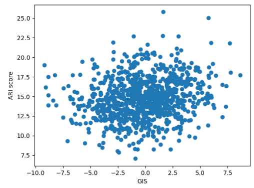
(a) GIS vs. FKGL for ABS



(b) GIS vs. FKGL for PLS



(c) GIS vs. ARI for ABS



(d) GIS vs. ARI for PLS

Figure A2: Scatter plot for GIS and other metrics on Cochrane simplification dataset, on ABS and PLS documents separately

A.3 GisPy and Other Text Simplification Metrics

In this section, we demonstrate that there is no overlap between the aspects evaluated by GIS and other automatic text simplification metrics (FKGL and ARI), with highlighting the unique advantages of using GIS for this task.

Documents	GIS vs. FKGL	GIS vs. ARI
ABS	0.17	0.14
PLS	0.18	0.18

Table A1: Pearson correlation coefficients between GIS and FKGL, and between GIS and ARI

To validate the above argument, we calculated the Pearson correlation coefficients between the reference-less metrics shown in A1, where the results are all close to zero: for between GIS and KFGL, the numbers are 0.17 for ABS texts and 0.18 for PLS texts; for between GIS and ARI, the coefficient is 0.14 for ABS texts, and 0.18 for PLS texts. Note that the Pearson correlation coefficient would suggest no linear correlation if the value between two distributions is close to 0. These results further prove that GIS is uncorrelated with these metrics.