# Automatic Medical Text Simplification: Challenges of Data Quality and Curation

**Chandrayee Basu**[1]
**Rosni Vasu**[2], **Michihiro Yasunaga**[1],
**Sohyeong Kim**[1], **Qian Yang**[3]
[1]Stanford University
cbasu@stanford.edu
[2]University of Zurich
[3]Cornell University

## Abstract

Health Literacy is the degree to which individuals can comprehend basic health information needed to make appropriate health decisions. The topmost reason for low health literacy is the vocabulary gap between providers and patients. Automatic medical text simplification can contribute to improving health literacy by assisting providers with patient-friendly communication, improving health data search, and making online medical texts more accessible. It is, however, extremely challenging to curate quality corpus for this natural language processing (NLP) task. In this position paper, we observe that, despite recent research efforts, existing open corpora for medical text simplification are poor in quality and size. In order to match the progress in general text simplification and style transfer, we must leverage careful crowd-sourcing. We discuss the challenges of naive crowd-sourcing. We propose that careful crowd-sourcing for medical text simplification is possible, when combined with automatic data labeling, a well-designed expert-layman collaboration framework, and context-dependent crowd-sourcing instructions.

Low health literacy has been associated with non-adherence to treatment plans and regimens, poor patient self-care, lack of timely communication of health issues, and increased risk of hospitalization and mortality (King 2010). Simplification of medical documents, of online communications like email messages and patient instructions can go a long way to mitigate health literacy challenges. While the consumer versions of medical journals, news articles, and a few trusted websites (NIA 2018; Savery et al. 2020) are written by trained experts, they are by no means exhaustive. Automated approaches are necessary to keep pace with the rapidly growing body of biomedical literature. In this work, we evaluate some of the open corpora that power automated text simplification in the medical domain.

We define text simplification, following Siddharthan (2014), as the process of reducing the linguistic complexity of a text, while still retaining the original information content and meaning. A domain-specific expert text undergoes various kinds of transformations to reach the final simple form. Research in automatic non-medical text simplification has been burgeoning, with the introduction of large parallel corpora (Zhu, Bernhard, and Gurevych 2010; Woodsend and Lapata 2011; Coster and Kauchak 2011; Xu, Callison-Burch, and Napoles 2015; Paetzold and Specia 2017). Creation of multi-references enabled models that can learn different kinds of textual transformations separately, viz. *lexical changes* (e.g. paraphrasing), *syntactic modifications* (e.g. reordering of concepts, splitting texts, reducing sentence length etc.) and *compression* (e.g. deleting peripheral information irrelevant to the target domain) (Alva-Manchego et al. 2020).

References are gold standard human generated simplifications, used to validate model outputs. The success of the automatic text simplification and style transfer hinges on large amounts of crowd-sourced multiple references. However, crowd-sourcing even a single set of references for medical texts is challenging. It requires the recruitment of a specific sub-population with a certain degree of domain expertise. For example, Nye et al. (2018) described an elaborate process of recruiting MDs and medical experts from Upwork, for PICO data annotation. Naturally, we observe a dearth of high-quality parallel training corpus in medical AI. Furthermore, text simplification task has additional challenges. Only the expert knows what content of the domain-specific text is relevant to the laymen, whereas the laymen or medical writers trained to translate medical texts can judge the quality and accessibility of the simplified versions.

In this work, we make the following contributions:

- identify the open-source datasets for medical text simplification
- characterize the datasets by their quantity, quality, diversity, and representativeness
- identify challenges of scaling high-quality corpus generation for medical text simplification

**Assumptions**: We treat summarization as a subset of text simplification. We only consider corpora that represent composite textual transformations (*simple text is derived after a combination of syntactic, semantic, thematic, and lexical transformations of the expert text*) (Lyu et al. 2021) for further analysis.

## Datasets for Medical Text Simplification

Datasets for medical text simplification support two kinds of document simplification: *sentence-level* and *paragraph-level*. We focus on sentence-level and short paragraph-level simplification. After an elaborate search, we found three datasets in English for medical text simplification: two parallel corpora SIMPWIKI (Van den Bercken, Sips, and Lofi 2019) and PARASIMP (Devaraj et al. 2021), and one non-parallel corpus MSD (Cao et al. 2020).

Next, we delve deeper into how these datasets are created and the potential artifacts of the data collection and annotation processes.

### Artifacts of Corpus Curation

In the absence of reliable crowd-sourcing of medical texts, researchers resort to crawling medical websites. The expert texts are sampled from the online articles and checked posthoc for adequate corpus representativeness. The layman texts are retrieved from the layman or consumer versions of the professional articles, based on the alignment of section titles and text content. The alignment is either checked manually for a small fraction of the corpus or automatically derived using different algorithms. Only a few of the automatically aligned pairs are validated by the experts. Automatic alignment is not always reasonable (Alva-Manchego, Scarton, and Specia 2020). Random sampling of expert texts from larger articles and unreliable automatic retrieval can lead to text pieces that are not stand-alone (Choi et al. 2021). We found that the process of expert verification is insufficient for quality data curation and could still lead to pairs lacking correspondence. On the other end, models trained using highly aligned text pairs may exhibit limited generalizability.

A more recent trend is to generate large volumes of non-parallel corpus, obviating validation of automatically aligned pairs. This follows similar approaches in non-medical text style transfer (Shen et al. 2017; He and McAuley 2016; Madaan et al. 2020). Some researchers distinguish between text simplification and text style transfer tasks. We consider text simplification as a sub-domain of text style transfer where the goal is to transform text from the expert style to the layman style.

### Datasets

Van Den Bercken (Van den Bercken, Sips, and Lofi 2019) contributed the very first publicly available medical text simplification corpus, which we refer to as SIMPWIKI, similar to (Cao et al. 2020). The authors created three subsets, *fully-aligned expert*: medical subset of Wikipedia data from Hwang et al. (2015), gleaned using QuickUMLS (Soldaini and Goharian 2016) for NER and later validated by experts, *partly-aligned expert* and *fully-aligned automatic*: texts from Wikipedia and Simple Wikipedia aligned using BLEU score (Papineni et al. 2002). Fully aligned text pairs have strong one-on-one correspondence, partly aligned simple texts cover the expert text entirely, but have additional facts. This dataset has 9212 expert-layman pairs. The texts are $\leq 128$ tokens long.

MSD is a non-parallel corpus derived from Merck Manuals, a trusted health reference for 100 years, with a wide range of medical topics. For each topic, the manual contains a consumer version and an expert version of the text, making it an ideal candidate for a text simplification corpus curation. This dataset offers wide coverage of medical topics and medical PICO elements (Cao et al. 2020). The authors scraped raw consumer and professional texts from the MSD website, split them into sentences, identified parallel groups by matching document titles and subsection titles, and picked linked sentences from the matched sections of the articles. The resulting text pairs were validated by non-native English speakers. The annotators used native language translations to speed up annotations. The text pairs are also annotated with UMLS concepts (Bodenreider 2004) for domain knowledge. MSD data has 130,349 expert texts, 114,674 layman texts in the non-parallel training set, and 675 expert-layman pairs for validation. The texts are $\leq 245$ tokens long.

We also considered a paragraph-level simplification corpus (Devaraj et al. 2021). The corpus consists of technical abstracts of biomedical systematic reviews and corresponding plain language summaries (PLS) from Cochrane Database of Systematic Reviews (McIlwain et al. 2014). The PLS are written in simple English. They usually represent the key essence of the abstracts and are structured heterogeneously (Kadic et al. 2016). We decided to exclude this corpus from our analysis due to the abstractive summary nature of the layman versions.

The size of parallel corpora is extremely small compared to those for non-medical text simplification, where the median corpus size is 154K (Alva-Manchego, Scarton, and Specia 2020).

## Automatic Dataset Quality Assessment

We assessed MSD and SIMPWIKI for their overall quality, diversity and representativeness. We define these terms as follows: *Quality* : grammatical correctness, average readability score, adherence to domain specific styles, *Diversity* : coverage of various transformations that text simplifications entail in medical domain (*different from diversity of language generation* (Ippolito et al. 2019)), and *Representativeness* : coverage of various medical sub-domains (e.g. gynecology, neurology, cardiology) and topics (for e.g., symptoms, signs, treatments).

### Metrics

We measured the above features separately for parallel and non-parallel corpora.

**Quality:** For *grammatical correctness*, we used the average acceptability score returned by textattack's RoBERTa-based classifier for CoLA (Morris et al. 2020; Warstadt, Singh, and Bowman 2019; HuggingFace 2021). We computed the *readability* of the two corpora in terms of Flesch-Kincaid Reading Ease, Flesch-Kincaid Grade level (Kincaid et al. 1975), and Automated Readability Index (ARI) (Senter and Smith 1967), similar to Li and Nenkova

(2015); Devaraj et al. (2021). We used classifiability, relative lexical complexity, and elaboration as metrics of *domain-specific styles*. We measured classifiability by the test accuracy of a trained attribute model (Yang et al. 2018; Subramanian et al. 2018; Prabhumoye et al. 2018). Following Siddharthan (2014), we expect good quality simplified corpus to contain sufficient elaborations of technical concepts and jargon and fewer low-frequency words. We trained a 1D CNN attribute model (Kim 2014) with GPT2 embedding (Radford et al. 2019) for computing classifiability. We reported how much elaborations are present in the simple texts of the corpora using thresholded cosine similarity between Sentence-BERT embeddings (Zhong et al. 2020; Reimers and Gurevych 2019) of the text pairs. We embedded each sentence of the simple text and the expert text and computed pairwise alignments. We used Sentence-BERT because it is tuned on several corpora, including SciDocs (Cohan et al. 2020) to embed sentences and short paragraphs and performed better than competing models on several downstream tasks. That said, wherever possible, we avoided language-model based metrics due to a mismatch between medical and model training datasets.

**Diversity:** We argue that quality corpus for text simplification should be *diverse* enough to accommodate various textual transformations, that domain-specific simplifications entail. These transformations could be lexical, semantic, and syntactic. Lexical transformations refer to substitution of complex terms or phrases by more accessible ones and could also include elaborations (extensions) or explanations (intentions). Syntactic transformations are more style dependent like formality change, voice change, tense change etc. We measured semantic diversity of the MSD validation data and the entire SIMPWIKI corpus using Sentence-BERT based corpus alignment.

We measured lexical and syntactic transformations using *referenceless quality features* like Levenshtein similarity, the proportion of words added, deleted or kept, compression ratio, lexical complexity ratio etc., from the EASSE library (Martin et al. 2018, 2019; Alva-Manchego et al. 2020).

**Representativeness:** We also checked which of the two corpora covers a wider range of medical subdomains and topics. Cao et al. (2020) already measured the representativeness for MSD by the distributions of the PICO elements (slightly different from the PICO elements in Nye et al. (2018)) and medical subdomains.

SIMPWIKI being a subset of Wikipedia articles relevant to medical topics, we referred to Shafee et al. (2017), for its representativeness. There are 30,000 articles on medical topics in Wikipedia. The articles are rated for quality and importance by editors. The top-rated articles are on tuberculosis and pneumonia. High-importance includes common diseases and treatments. Mid-importance encompasses conditions, tests, drugs, anatomy and symptoms. The remaining low-importance articles include niche or peripheral medical topics such as laws, physicians and rare conditions.

## Results

**Quality** Approximately 90 % of the expert texts in MSD were acceptable and $>$ 97 % of the MSD layman texts and SIMPWIKI were acceptable by the CoLA model. This means $\approx$360 texts, each in expert and layman versions within SIMPWIKI corpus, were not acceptable. 10 % of the expert texts (11460 texts) in MSD had low acceptability score, possibly because of unique vocabulary and sentence structures, and incomplete references.

See Table. 2 for the readability scores. We found discrepancies with the readability scores reported in Cao et al. (2020). Paired t-test shows that the expert and the simple texts in both MSD and SIMPWIKI have statistically significant differences in readability, measured by Flesch Reading Score, Flesch Kincaid Grade, and Automated Readability Index ($p < 0.001$). The minimum readability of medical texts compared to general English corpora is low (Minimum Flesch Kincaid grade level is 11.9), also observed by Devaraj et al. (2021).

Lexical complexity, computed using the EASSE package (Martin et al. 2019), represents the word rank score distribution of the corpus. While the mean complexity of MSD is not very different between expert and layman versions, a much lower standard deviation confirms that expert texts have more rare words. SIMPWIKI has more common words in both expert and layman versions than MSD, and the complexity varies across the corpus. We measured percentage of simple texts that potentially contain elaborations, both for MSD, and separately for differently aligned pairs of SIMPWIKI. We found high proportion of elaborations in MSD based on our coarse approach, which is desirable. However, further human validations are required to confirm the relevance of these elaborations.

We trained two different attribute models for classifiability check. We did not notice a significant difference in the test accuracy of the two corpora. Note that the training data size was significantly larger for MSD. The accuracy was 0.88 and 0.81 for MSD and SIMPWIKI respectively.

## Diversity

We computed several *referenceless* text quality metrics using the EASSE library (Martin et al. 2019). We made some modifications to output mean, standard deviation, and standard error of the metrics. We used these automatic metrics as a proxy for simplification-related transformations. An average compression ratio of $>$ 1 in MSD points to more elaborations and explanations (potentially irrelevant facts). A higher standard deviation of compression ratio indicates more diversity in transformations. Higher additions in MSD indicate more domain specific words (possibly more common words) being introduced in the simpler versions. Overall, we observe that MSD represents more textual transformations than SIMPWIKI.

# Human Data Quality Assessment

In the previous section, we used automatic metrics to evaluate the approximate quality and diversity of the corpora

Table 1: Transformation Diversity Metrics.

| Metrics | Layman to Expert Ratio | |
|---|---|---|
| | **MSD** | **SIMPWIKI** |
| Compression Ratio | $1.257 \pm 0.9$ | $0.907 \pm 0.46$ |
| Levenshtein Similarity | $0.519 \pm 0.166$ | $0.641 \pm 0.219$ |
| Exact copies | 0.029 | 0.07 |
| Additions proportion | $0.526 \pm 0.254$ | $0.304 \pm 0.251$ |
| Deletions proportion | $0.439 \pm 0.244$ | $0.421 \pm 0.286$ |
| Added words | $20.135 \pm 18.144$ | $7.951 \pm 8.941$ |
| Deleted words | $17.181 \pm 20.504$ | $12.16 \pm 11.89$ |
| Kept words | $11.914 \pm 9.881$ | $12.529 \pm 9.5$ |
| Corpus alignment | $0.428 \pm 0.226$ | $0.832 \pm 0.161$ (auto_full)<br>$0.862 \pm 0.125$ (exp_full)<br>$0.597 \pm 0.163$ (exp_part) |

Table 2: Quality metrics.

| Metric | MSD Test | | SIMPWIKI | |
|---|---|---|---|---|
| | **Expert** | **Layman** | **Expert** | **Layman** |
| Acceptability score | 0.907 | 0.976 | 0.977 | 0.965 |
| Flesch Reading Ease | $17.44 \pm 32.25$ | $37.116 \pm 28.19$ | $30.07 \pm 28.34$ | $41.47 \pm 29.12$ |
| Flesch Kincaid Grade level | $15.2 \pm 5.4$ | $12.6 \pm 5.7$ | $14.4 \pm 5.5$ | $11.9 \pm 5.1$ |
| ARI | $15.6 \pm 6.4$ | $13 \pm 6.9$ | $15.1 \pm 6.7$ | $12.4 \pm 6.2$ |
| Lexical complexity | $9.17 \pm 0.087$ | $9 \pm 0.792$ | $8.842 \pm 0.79$ | $8.695 \pm 0.867$ |
| Elaboration | | 27.4 | | 4.6 (auto_full)<br>0.8 (exp_full)<br>0.8 (exp_part) |

for medical text simplification. We found that MSD potentially is more diverse, but also has lower acceptability because of the sheer scale of the data and unique vocabulary. The expert texts in MSD require a higher minimum reading grade. While this corpus seems to contain more elaborations in the validation set, compared to SIMPWIKI, the elaborations cannot be explicitly learnt from the non-parallel training data. All of the above points to the need for further data collection and quality human annotation.

## Crowd-sourcing

In many NLP tasks, it is customary to complement automatic model validations with human evaluations. A large body of work has been dedicated to analyse and correct the mismatch between human judgement and automatic evaluation. Researchers found that both metrics (Banerjee and Lavie 2005; Zhang et al. 2019; Ma et al. 2019) and artifacts of data collection (Freitag, Grangier, and Caswell 2020) can be responsible for the mismatch. One solution to ensure data diversity is to crowd-source multiple references (Freitag, Grangier, and Caswell 2020). Lyu et al. (2021); Alva-Manchego et al. (2020) released a text simplification multi-reference corpus annotated with various simplification transformations. Newsela corpus for general text simplification was annotated for different grades of education (Xu, Callison-Burch, and Napoles 2015). Multi-references will also be useful in the medical domain for personalization (Paetzold and Specia 2016; Su et al. 2021).

To assess whether crowd-sourcing is a valid option for quality check and multi-reference generation of medical texts, we conducted a test internally, between two coauthors of this paper. Both the authors had high school biology in English. One author consumes medical information weekly from scientific articles, popular science news and blogs, and communicates with medical practitioner online. Another author uses google search infrequently for medical symptoms lookup only. We sampled 60 sentences from MSD: 20 with longer simple texts, 20 with longer expert texts and 20 where simple and expert texts have similar number of tokens. We asked each author to indicate agreement on several statements covering content preservation, coverage, textual simplicity, concept simplicity and fluency of the simple text, for e.g.

- The simple sentence explains all the unknown concepts adequately
- The simple sentence removes all redundancy and covers only the key point in the reference sentence
- I cannot think of an alternative way to simplify it

Average Krippendorff's alpha (Krippendorff 2011) across 10 quality questions, between the two authors, was $0.299 \pm$

0.048. The results show high disagreement between the authors, questioning the plausibility of reliable human evaluation and crowd-sourcing of medical texts. However, in the absence of crowd-sourcing, we cannot generate diverse enough data to train and validate models with good generalizability.

### Can layman assess the simplification quality and provide alternative references?

To test this question, we conducted a pilot study with two users, where we iterated on a few different designs of layman evaluation of MSD validation data. The users had high school biology in English, but minimal experience of consuming medical information online. We found that the users were unmotivated to read the entire expert text, because of the jargon, resulting in an inability to judge the quality of the simplification. More importantly, some of the ratings changed, after the texts were explained to the users. A prominent artifact of data scraping and automatic alignment was the change in the *subject* of the text, which confused the evaluation. For e.g. in this text pair: **expert:** *In adults , BMI , defined as weight ( kg ) divided by the square of the height ( m2 ) , is used to screen for overweight or obesity ( see table Body Mass Index ( BMI ) ) : Overweight = 25 to 29.9 kg/m2 ; Obesity = $\geq$ 30 kg/m2* **simple:** *Obesity is diagnosed by determining the BMI.*
BMI is the subject in the former and obesity is the subject in the latter. When asked if the users were confident that they could rewrite the simplification better, we got an unanimous yes.

We concluded that only experts have the ability to comprehend which sections of the expert texts are useful for laymen. Only laymen and trained writers can validate whether the simple versions are readable and meaningful. In other words, scaling up human evaluation and annotation, in this case, calls for well-designed collaboration between experts and laymen.

### Expert-layman collaboration

We delineated various potential formats of expert-layman collaborations. The experts could be MD and biomedical students, physicians and nurses directly, or they could be models of expert behavior. The simplest approach would be to show definitions of the UMLS concepts. We found that these concepts are not always accessible for layman. Other researchers have used Google's "define:" to improve readability of medical texts (Elhadad 2006). Some potential expert-layman collaboration could look like the following: Show examples of text pairs rated by experts, their rationale behind rating and their corrections to unacceptable simplification, ask experts to generate a question from the expert text and ask layman to answer the question after reading the simple version of the text. The expert generated question is automatically based on the key content of the expert text. The layman should understand the content of the simple text to answer this question. We could also use limited expert annotated data to model expert behavior in terms of extracting key concepts from texts, identifying concepts that need elaborations and so on. This model can be leveraged to improve layman evaluations.

## Discussion

Automatic medical text simplification can contribute to improving health literacy by assisting providers with patient-friendly communication, improving health data search, and making online medical texts more accessible. However, it is challenging to create large annotated and parallel corpus for this task, unlike for non-medical texts. In this paper, we identified the existing corpora for training automatic text simplification models, and analyzed their quality and diversity using several automatic metrics. We found that taking snapshots from expert and consumer articles that are not aligned could lead to poor quality parallel corpus. We also assessed the potential of leveraging crowd-sourcing for large-scale model evaluation and data annotation for this task. We found that laymen evaluate the medical texts very differently, depending upon their exposure to medical information. We proposed some crowd-sourcing solutions that could use expert-layman collaboration. In future, we plan to explore such collaborative data curation and annotation, in practice. Another exciting research avenue would be to train controllable simplification models that can interface with and learn from these two stakeholders.

## References


Alva-Manchego, F.; Martin, L.; Bordes, A.; Scarton, C.; Sagot, B.; and Specia, L. 2020. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4668–4679. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.424. URL https://aclanthology.org/2020.acl-main.424.

Alva-Manchego, F.; Scarton, C.; and Specia, L. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics* 46(1): 135–187.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32(suppl_1): D267–D270.

Cao, Y.; Shui, R.; Pan, L.; Kan, M.-Y.; Liu, Z.; and Chua, T.-S. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. *arXiv preprint arXiv:2005.00701* .

Choi, E.; Palomaki, J.; Lamm, M.; Kwiatkowski, T.; Das, D.; and Collins, M. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics* 9: 447–461.

Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; and Weld, D. S. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180* .

Coster, W.; and Kauchak, D. 2011. Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 665–669. Portland, Oregon, USA: Association for Computational Linguistics. URL https://aclanthology.org/P11-2117.

Devaraj, A.; Marshall, I.; Wallace, B.; and Li, J. J. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4972–4984. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.395. URL https://aclanthology.org/2021.naacl-main.395.

Elhadad, N. 2006. Comprehending technical texts: Predicting and defining unfamiliar terms. In *AMIA annual symposium proceedings*, volume 2006, 239. American Medical Informatics Association.

Freitag, M.; Grangier, D.; and Caswell, I. 2020. BLEU might be Guilty but References are not Innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 61–71. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.5. URL https://aclanthology.org/2020.emnlp-main.5.

He, R.; and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, 507–517.

HuggingFace. 2021. The AI community building the future. URL https://huggingface.co/.

Hwang, W.; Hajishirzi, H.; Ostendorf, M.; and Wu, W. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 211–217.

Ippolito, D.; Kriz, R.; Kustikova, M.; Sedoc, J.; and Callison-Burch, C. 2019. Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362* .

Kadic, A. J.; Fidahic, M.; Vujcic, M.; Saric, F.; Propadalo, I.; Marelja, I.; Dosenovic, S.; and Puljak, L. 2016. Cochrane plain language summaries are highly heterogeneous with low adherence to the standards. *BMC medical research methodology* 16(1): 1–4.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. *CoRR* abs/1408.5882. URL http://arxiv.org/abs/1408.5882.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

King, A. 2010. Poor health literacy: a'hidden'risk factor. *Nature Reviews Cardiology* 7(9): 473–474.

Krippendorff, K. 2011. Computing Krippendorff's alpha-reliability .

Li, J. J.; and Nenkova, A. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Lyu, Y.; Liang, P. P.; Pham, H.; Hovy, E.; Póczos, B.; Salakhutdinov, R.; and Morency, L.-P. 2021. StylePTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer. *arXiv preprint arXiv:2104.05196* .

Ma, Q.; Wei, J.; Bojar, O.; and Graham, Y. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 62–90. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-5302. URL https://aclanthology.org/W19-5302.

Madaan, A.; Setlur, A.; Parekh, T.; Poczos, B.; Neubig, G.; Yang, Y.; Salakhutdinov, R.; Black, A. W.; and Prabhumoye, S. 2020. Politeness Transfer: A Tag and Generate Approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1869–1881. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.169. URL https://aclanthology.org/2020.acl-main.169.

Martin, L.; Humeau, S.; Mazaré, P.; Bordes, A.; de la Clergerie, É. V.; and Sagot, B. 2019. Reference-less Quality Estimation of Text Simplification Systems. *CoRR* abs/1901.10746. URL http://arxiv.org/abs/1901.10746.

Martin, L.; Humeau, S.; Mazaré, P.-E.; de La Clergerie, É.; Bordes, A.; and Sagot, B. 2018. Reference-less Quality Estimation of Text Simplification Systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, 29–38. Tilburg, the Netherlands: Association for Computational Linguistics. doi:10.18653/v1/W18-7005. URL https://aclanthology.org/W18-7005.

McIlwain, C.; Santesso, N.; Simi, S.; Napoli, M.; Lasserson, T.; Welsh, E.; Churchill, R.; Rader, T.; Chandler, J.; Tovey, D.; et al. 2014. Standards for the reporting of Plain Language Summaries in new Cochrane Intervention Reviews (PLEACS) .

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.

NIA, N. 2018. Online Health Information: Is It Reliable? URL https://www.nia.nih.gov/health/online-health-information-it-reliable.

Nye, B.; Li, J. J.; Patel, R.; Yang, Y.; Marshall, I. J.; Nenkova, A.; and Wallace, B. C. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, 197. NIH Public Access.

Paetzold, G.; and Specia, L. 2016. Anita: An Intelligent Text Adaptation Tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 79–83. Osaka, Japan: The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-2017.

Paetzold, G. H.; and Specia, L. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research* 60: 549–593.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000* .

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8): 9.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* .

Savery, M.; Abacha, A. B.; Gayen, S.; and Demner-Fushman, D. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data* 7(1): 1–9.

Senter, R.; and Smith, E. A. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Shafee, T.; Masukume, G.; Kipersztok, L.; Das, D.; Häggström, M.; and Heilman, J. 2017. Evolution of Wikipedia's medical content: past, present and future. *J Epidemiol Community Health* 71(11): 1122–1129.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. *arXiv preprint arXiv:1705.09655* .

Siddharthan, A. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165(2): 259–298.

Soldaini, L.; and Goharian, N. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, 1–4.

Su, L.; Duan, N.; Cui, E.; Ji, L.; Wu, C.; Luo, H.; Liu, Y.; Zhong, M.; Bharti, T.; and Sacheti, A. 2021. GEM: A General Evaluation Benchmark for Multimodal Tasks. *arXiv preprint arXiv:2106.09889* .

Subramanian, S.; Lample, G.; Smith, E. M.; Denoyer, L.; Ranzato, M.; and Boureau, Y.-L. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552* .

Van den Bercken, L.; Sips, R.-J.; and Lofi, C. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, 3286–3292.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7: 625–641.

Woodsend, K.; and Lapata, M. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 409–420. Edinburgh, Scotland, UK.: Association for Computational Linguistics. URL https://aclanthology.org/D11-1038.

Xu, W.; Callison-Burch, C.; and Napoles, C. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics* 3: 283–297.

Yang, Z.; Hu, Z.; Dyer, C.; Xing, E. P.; and Berg-Kirkpatrick, T. 2018. Unsupervised text style transfer using language models as discriminators. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7298–7309.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* .

Zhong, Y.; Jiang, C.; Xu, W.; and Li, J. J. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9709–9716.

Zhu, Z.; Bernhard, D.; and Gurevych, I. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1353–1361. Beijing, China: Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1152.