scientific data

OPEN

DATA DESCRIPTOR

A dataset for plain language adaptation of biomedical abstracts

Kush Attal ✉, Brian Ondov & Dina Demner-Fushman

Though exponentially growing health-related literature has been made available to a broad audience online, the language of scientific articles can be difficult for the general public to understand. Therefore, adapting this expert-level language into plain language versions is necessary for the public to reliably comprehend the vast health-related literature. Deep Learning algorithms for automatic adaptation are a possible solution; however, gold standard datasets are needed for proper evaluation. Proposed datasets thus far consist of either pairs of comparable professional- and general public-facing documents or pairs of semantically similar sentences mined from such documents. This leads to a trade-off between imperfect alignments and small test sets. To address this issue, we created the Plain Language Adaptation of Biomedical Abstracts dataset. This dataset is the first manually adapted dataset that is both document- and sentence-aligned. The dataset contains 750 adapted abstracts, totaling 7643 sentence pairs. Along with describing the dataset, we benchmark automatic adaptation on the dataset with state-of-the-art Deep Learning approaches, setting baselines for future research.

Background & Summary

While reliable resources for health information conveyed in a plain language format exist, such as the MedlinePlus website from the National Library of Medicine (NLM)[1], these resources do not provide all the necessary information for every health-related situation or rapidly changing state of knowledge arising from novel scientific investigations or global events like pandemics. In addition, the language used in other health-related articles can be too difficult for patients and the general public to comprehend[2], which has a major impact on health outcomes[3]. While work in simplifying text exists, the unique language of biomedical text warrants a distinct subtask similar to machine translation, termed adaptation[4]. Adapting natural language involves creating a simplified version that maintains the most important details from a complex source. Adaptations are a common tool for teachers to use to improve comprehension of content for English language learners[5].

A standard internet search will return multiple scientific articles that correspond to a patient's query; however, without extensive clinical and/or biological knowledge, the user may not be able to comprehend the scientific language and content[6]. There are articles with verified, plain language summaries for health information, such as the articles with corresponding plain language summaries created by medical health organization Cochrane[7]. However, creating manual summaries and adaptations for every article addressing every user's queries is not possible. Thus, an automatic adaptation generated for material responding to a user's query is very relevant, especially for patients without clinical knowledge.

Though plain language thesauri and other knowledge bases have enabled rule-based systems that substitute difficult terms for more common ones, human editing is needed to account for grammar, context, and ambiguity[8]. Deep Learning may offer a solution for fully automated adaptation. Advances in architectures, hardware, and available data have led neural methods to achieve state-of-the-art results in many linguistic tasks, including Machine Translation[9] and Text Simplification[10]. Neural methods, however, require large numbers of training examples, as well as benchmark datasets to allow iterative progress[11].

Parallel datasets for Text Simplification have been assembled by searching for semantically similar sentences across comparable document pairs, for example articles on the same subject in both Wikipedia and Simple English Wikipedia (or Vikidia, an encyclopedia for children in several languages)[12–15]. Since Wikipedia contains some articles on biomedical topics, it has been proposed to extract subsets of these datasets for use in this domain[16–19]. However, since these sentence pairs exist in different contexts, they are often not semantically identical, having undergone sentence-level operations like splitting or merging. Sentence pairs pulled out of context may also use anaphora on one side of a pair but not the other. This can confuse models during training

Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. ✉e-mail: attalk21@alumni.wlu.edu

and expect impossible replacements during testing. Further, Simple English Wikipedia often still contains complex medical terms on the simple side[16,20,21]. Parallel sentences have also been mined from dedicated biomedical sources. Cao *et al*. have expert annotators pinpoint highly similar passages, usually consisting of one or two sentences from each passage, from Merck Manuals, an online website containing numerous articles on medical and health topics created for both professional and general public groups[22]. In addition, Pattisapu *et al*. have expert annotators identify highly similar pairs from scientific articles and corresponding health blogs describing them[23]. Though human filtering makes the pairs in both these datasets much closer to being semantically identical, at less than 1,000 pairs each, they are too small for training and even less ideal for evaluation[24]. Sakakini *et al*. manually translate a somewhat larger set (4,554) of instructions for patients from clinical notes[25]. However, this corpus covers a very specific case within the clinical domain, which itself constitutes a separate sublanguage from biomedical literature[26].

Since recent models can handle larger paragraphs, comparable corpora have also been suggested as training or benchmark datasets for adapting biomedical text. These corpora consist of pairs of paragraphs or documents that are on the same topic and make roughly the same points, but are not sentence-aligned. Devaraj *et al*. present a paragraph level corpus derived from Cochrane review abstracts and their Plain Language Summaries, using heuristics to combine subsections with similar content across the pairs. However, these heuristics do not guarantee identical content[27]. This dataset is also not sentence-aligned, which limits the architectures that can take advantage of it and results in restriction of documents to those with no more than 1024 tokens. Other datasets include comparable corpora or are created at the paragraph-level and omit relevant details from the original article[27]. To the best of our knowledge, no datasets provide manual, sentence-level adaptations of the scientific abstracts[28]. Thus, there is still a need for a high-quality, sentence-level gold standard dataset for the adaptation of general biomedical text.

To address this need, we have developed the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset. PLABA contains 750 abstracts from PubMed (10 on each of 75 topics) and expert-created adaptations at the sentence-level. Annotators were chosen from the NLM and an external company and given abstracts within their respective expertise to adapt. Human adaptation allows us to ensure the parallel nature of the corpus down to sentence-level granularity, but still while using the surrounding context of the entire document to guide each translation. We deliberately construct this dataset so it can serve as a gold standard on several levels:

1. Document level simplification. Documents are simplified in total, each by at least one annotator, who is instructed to carry over all content relevant for general public understanding of the professional document. This allows the corpus to be used as a gold standard for systems that operate at the document level.
2. Sentence level simplification. Unlike automatic alignments, these pairings are ensured to be parallel for the purpose of simplification. Semantically, they will differ only in (1) content removed from the professional register because the annotator deemed it unimportant for general public understanding, and (2) explanation or elaboration added to the general public register to aid understanding. Since annotators were instructed to keep content within sentence boundaries (or in split sentences), there are no issues with fragments of other thoughts spilled over from neighboring sentences on one side of the pair.
3. Sentence-level operations and splitting. Though rare in translation between languages, sentence-level operations (e.g. merging, deletion, and splitting) are common in simplification[29]. Splitting is often used to simplify syntax and reduce sentence length. Occasionally sentences may be dropped from the general public register altogether (deletion). For consistency and simplicity of annotation, we do not allow merging, creating a one-to-many relationship at the sentence level.

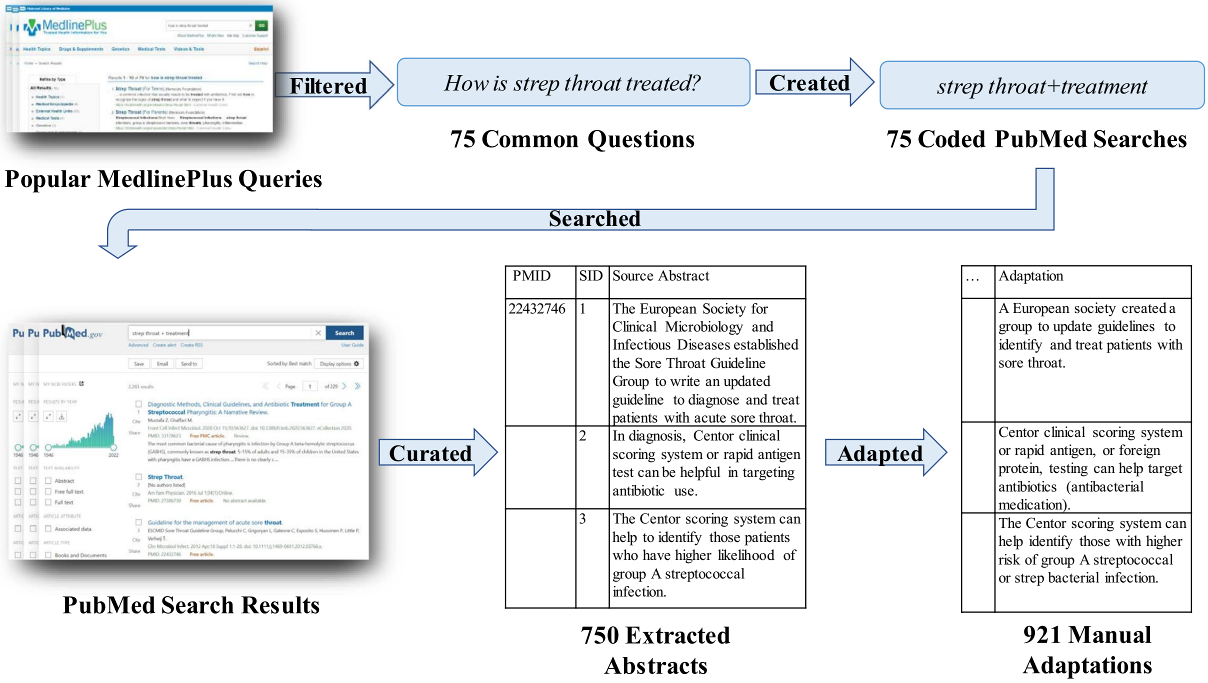
The PLABA dataset should further enable the development of systems that automatically adapt relevant medical texts for patients without prior medical knowledge. In addition to releasing PLABA, we have evaluated state-of-the-art deep learning approaches on this dataset to set benchmarks for future researchers.

Methods

The PLABA dataset includes 75 health-related questions asked by MedlinePlus users, 750 PubMed abstracts from relevant scientific articles, and corresponding human created adaptations of the abstracts. The questions in PLABA are among the most popular topics from MedlinePlus, ranging from topics like COVID-19 symptoms to genetic conditions like cystic fibrosis[1].

To gather the PubMed abstracts in PLABA, we first filtered questions from MedlinePlus logs based on the frequency of general public queries. Then, a medical informatics expert verified the relevance of and lack of accessible resources to answer each question and chose 75 questions total. For each question, the expert coded its focus (COVID-19, cystic fibrosis, compression devices, etc.) and question type (general information, treatment, prognosis, etc.) to use as keywords in a PubMed search[30]. Then, the expert selected 10 abstracts from PubMed retrieval results that appropriately addressed the topic of the question, as seen in Fig. 1.

To create the corresponding adaptations for each abstract in PLABA, medical informatics experts worked with source abstracts separated into individual sentences to create corresponding adaptations across all 75 questions. Adaptation guidelines allowed annotators to split long source sentences and ignore source sentences that were not relevant to the general public. Each source sentence corresponds to no, one, or multiple sentences in the adaptation. Creating these adaptations involved syntactic, lexical and semantic simplifications, which were developed in the context of the entire abstract. Examples taken from the dataset can be seen in Table 1. Specific examples of adaptation guidelines are demonstrated in Fig. 2 and included:



PMID	SID	Source Abstract
22432746	1	The European Society for Clinical Microbiology and Infectious Diseases established the Sore Throat Guideline Group to write an updated guideline to diagnose and treat patients with acute sore throat.
	2	In diagnosis, Centor clinical scoring system or rapid antigen test can be helpful in targeting antibiotic use.
	3	The Centor scoring system can help to identify those patients who have higher likelihood of group A streptococcal infection.

…	Adaptation
	A European society created a group to update guidelines to identify and treat patients with sore throat.
	Centor clinical scoring system or rapid antigen, or foreign protein, testing can help target antibiotics (antibacterial medication).
	The Centor scoring system can help identify those with higher risk of group A streptococcal or strep bacterial infection.

Fig. 1 Overview representing how questions and PubMed abstracts for the dataset were searched and chosen for annotators to adapt. PMID refers to the PubMed ID from which the example originates from. SID refers to the sentence ID or number of the example sentence from the source abstract.

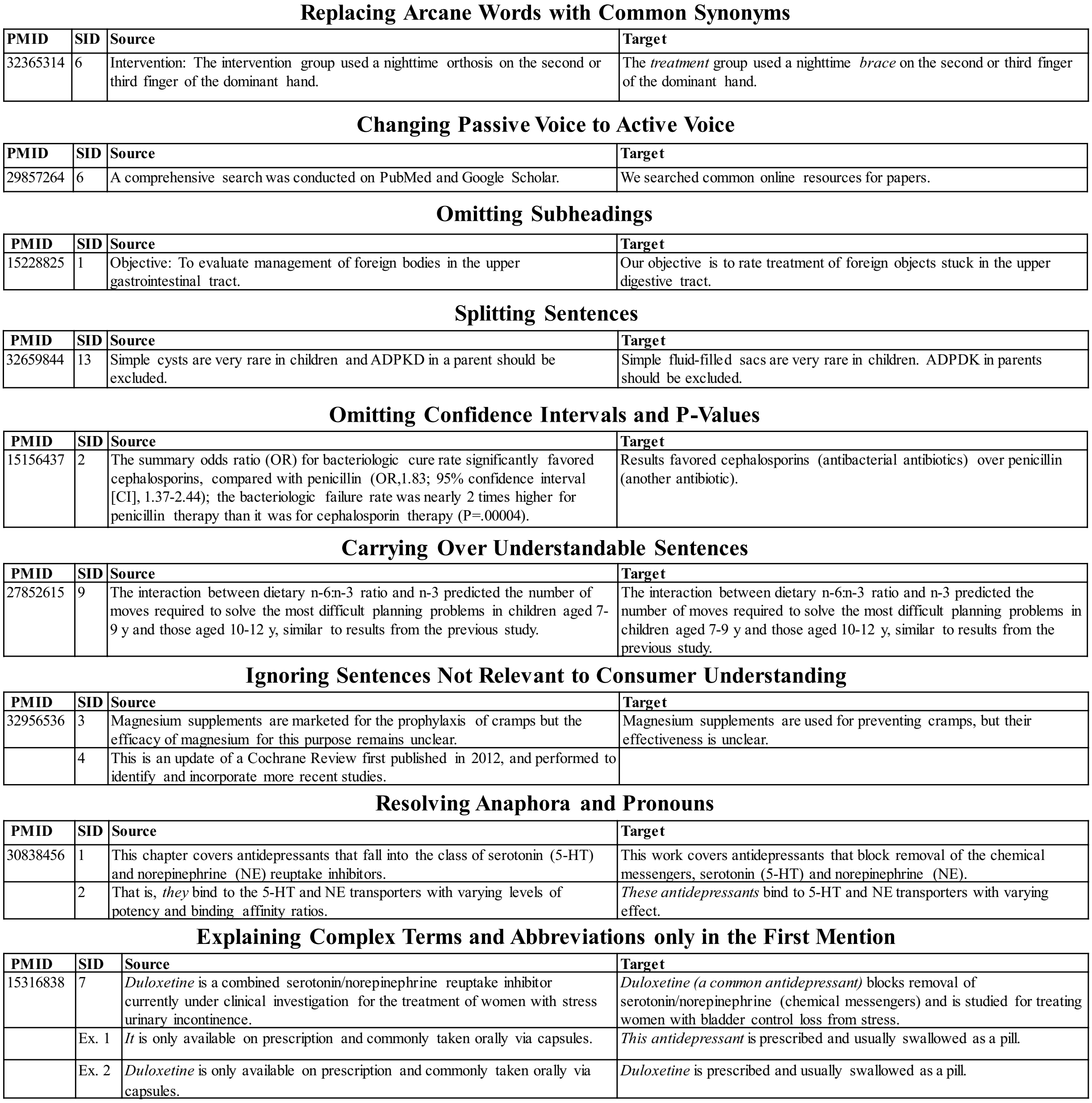
Data Type	Text
Question	*How is strep throat treated?*
Abstract	Most patients who seek medical attention for sore throat are concerned about streptococcal tonsillopharyngitis, but fewer than 10% of adults and 30% of children actually have a streptococcal infection. Group A beta-hemolytic streptococci (GAS) are most often responsible for bacterial tonsillopharyngitis, although Neisseria gonorrhea, Arcanobacterium haemolyticum (formerly Corynebacterium haemolyticum), Chlamydia pneumoniae (TWAR agent), and Mycoplasma pneumoniae have also been suggested as possible, infrequent, sporadic […]
Adaptation	Most people who go to a doctor for sore throat are worried they have a strep throat and tonsil infection, but fewer than 10% of adults and 30% of children actually have a strep infection. Group A strep bacteria are the most common cause of bacterial strep throat and tonsil infection, but other bacteria known to cause sexually-transmitted gonorrhea or chlamydia, or head, neck, and lung infections occasionally might […]
Question	*How do you assist an unconscious victim who is already vomiting?*
Abstract	The tongue is the most common cause of upper airway obstruction, a situation seen most often in patients who are comatose or who have suffered cardiopulmonary arrest. Other common causes of upper airway obstruction include edema of the oropharynx and larynx, trauma, foreign body, and […]
Adaptation	The tongue is the most common cause of blocked upper airways, seen most often in people in comas or cardiac arrest (abrupt heart stop). Other common causes of blocked upper airways include swelling of the middle part of the throat and voice box, injury, objects that shouldn't be swallowed, and […]

Table 1. Examples of questions, abstracts, and adaptations in PLABA.

- Replacing arcane words like "orthosis" with common synonyms like "brace"
- Changing sentence structure from passive voice to active voice
- Omitting or incorporating subheadings at the beginning of sentences (e.g., "Aim:", "Purpose:")
- Splitting long, complex sentences into shorter, simpler sentences
- Omitting confidence intervals and other statistical values
- Carrying over understandable sentences from the source with no changes into the adaptation
- Ignoring sentences that are not relevant to a patient's understanding of the text
- Resolving anaphora and pronouns with specific nouns
- Explaining complex terms and abbreviations with explanatory clauses when first mentioned

Data Records

We archived the dataset with Open Science Framework (OSF) at https://osf.io/rnpmf/https://osf.io/rnpmf/[31]. The dataset is saved in JSON format and organized or "keyed" by question ID. Each key is a question ID that contains a corresponding nested JSON object. This nested object contains the actual question, a single-letter key denoting if the question is a clinical question or biological question, and contains the abstracts and



Replacing Arcane Words with Common Synonyms

PMID	SID	Source	Target
32365314	6	Intervention: The intervention group used a nighttime orthosis on the second or third finger of the dominant hand.	The *treatment* group used a nighttime *brace* on the second or third finger of the dominant hand.

Changing Passive Voice to Active Voice

PMID	SID	Source	Target
29857264	6	A comprehensive search was conducted on PubMed and Google Scholar.	We searched common online resources for papers.

Omitting Subheadings

PMID	SID	Source	Target
15228825	1	Objective: To evaluate management of foreign bodies in the upper gastrointestinal tract.	Our objective is to rate treatment of foreign objects stuck in the upper digestive tract.

Splitting Sentences

PMID	SID	Source	Target
32659844	13	Simple cysts are very rare in children and ADPKD in a parent should be excluded.	Simple fluid-filled sacs are very rare in children. ADPDK in parents should be excluded.

Omitting Confidence Intervals and P-Values

PMID	SID	Source	Target
15156437	2	The summary odds ratio (OR) for bacteriologic cure rate significantly favored cephalosporins, compared with penicillin (OR,1.83; 95% confidence interval [CI], 1.37-2.44); the bacteriologic failure rate was nearly 2 times higher for penicillin therapy than it was for cephalosporin therapy (P=.00004).	Results favored cephalosporins (antibacterial antibiotics) over penicillin (another antibiotic).

Carrying Over Understandable Sentences

PMID	SID	Source	Target
27852615	9	The interaction between dietary n-6:n-3 ratio and n-3 predicted the number of moves required to solve the most difficult planning problems in children aged 7-9 y and those aged 10-12 y, similar to results from the previous study.	The interaction between dietary n-6:n-3 ratio and n-3 predicted the number of moves required to solve the most difficult planning problems in children aged 7-9 y and those aged 10-12 y, similar to results from the previous study.

Ignoring Sentences Not Relevant to Consumer Understanding

PMID	SID	Source	Target
32956536	3	Magnesium supplements are marketed for the prophylaxis of cramps but the efficacy of magnesium for this purpose remains unclear.	Magnesium supplements are used for preventing cramps, but their effectiveness is unclear.
	4	This is an update of a Cochrane Review first published in 2012, and performed to identify and incorporate more recent studies.	

Resolving Anaphora and Pronouns

PMID	SID	Source	Target
30838456	1	This chapter covers antidepressants that fall into the class of serotonin (5-HT) and norepinephrine (NE) reuptake inhibitors.	This work covers antidepressants that block removal of the chemical messengers, serotonin (5-HT) and norepinephrine (NE).
	2	That is, *they* bind to the 5-HT and NE transporters with varying levels of potency and binding affinity ratios.	*These antidepressants* bind to 5-HT and NE transporters with varying effect.

Explaining Complex Terms and Abbreviations only in the First Mention

PMID	SID	Source	Target
15316838	7	*Duloxetine* is a combined serotonin/norepinephrine reuptake inhibitor currently under clinical investigation for the treatment of women with stress urinary incontinence.	*Duloxetine (a common antidepressant)* blocks removal of serotonin/norepinephrine (chemical messengers) and is studied for treating women with bladder control loss from stress.
	Ex. 1	*It* is only available on prescription and commonly taken orally via capsules.	*This antidepressant* is prescribed and usually swallowed as a pill.
	Ex. 2	*Duloxetine* is only available on prescription and commonly taken orally via capsules.	*Duloxetine* is prescribed and usually swallowed as a pill.

Fig. 2 Example of the guidelines set for annotators. PMID refers to the PubMed ID from which the example originates from. SID refers to the sentence ID or number of the example sentence from the source abstract. Target refers to the manual adaptation.

corresponding human adaptations grouped by the PubMed ID (PMID) of the abstract. Table 2 shows statistics of the abstracts and adaptations. An example of the data format for one record can be found in the README file in the OSF archive.

Technical Validation

We measured the level of complexity, the ability to train tools and how well the main points are preserved in the automatic adaptations trained on our data. We first introduce the metrics we used to measure text complexity followed by the metrics to measure text similarity and inter-annotator agreement between manually created adaptations. We use the same text similarity metrics to also compare automatically created adaptations to both the source abstracts and manually created adaptations.

Evaluation metrics. To measure text readability and compare the abstracts and manually created adaptations, we use the Flesch-Kincaid Grade Level (FKGL) test[32]. FKGL uses the average number of syllables per word and the average number of words per sentence to calculate the score. A higher FKGL score for a text indicates a higher reading comprehension level needed to understand the text.

Data Type	Count	Words		Sentences	
		Average	S.d.	Average	S.d.
Questions	75	10	6	1	0
Abstracts	750	240	95	10	4
Adaptations	921	244	95	12	5

Table 2. Average number of words and sentences per data type.

In addition, we use BLEU[33], ROUGE[34], SARI[4,35], and BERTScore[36], commonly used text and semantic similarity and simplification metrics, to measure inter-annotator agreement, compare abstracts to manually created adaptations, and evaluate the automatically created adaptations. BLEU and ROUGE look at spans of contiguous words (referred to as n-grams in Natural Language Processing or NLP) to evaluate a candidate adaptation against a reference adaptation. For instance, BLEU-4 measures how many of the contiguous sequences from one to four words in length in the candidate adaptation appear in the reference adaptation. However, BLEU is a measure of precision and penalizes candidates for adding incorrect n-grams. ROUGE is a measure of recall and penalizes candidate adaptations for missing n-grams. Similarly, BERTScore looks at subwords to evaluate a candidate sentence against a reference sentence, comparing each candidate subword against every reference subword using contextual word embeddings. While BERTScore gives values of precision, recall, and F1 (which averages precision and recall), we solely report F1 metrics. Since BLEU, ROUGE, and BERTScore are not specifically designed for simplification, we also use SARI, which also incorporates the source sentence in order to weight the various operations involved in simplification. While n-grams are still used, SARI balances (1) addition operations, in which n-grams of the candidate adaptation are shared with the reference adaptation but not the source, (2) deletion operations, in which n-grams appear in the source but neither the reference nor candidate, and (3) keep operations, in which n-grams are shared by all three. We report BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L (which measures the longest shared sub-sequence between a candidate and reference), BERTScore-F1, and SARI. All metrics can account for multiple possible reference adaptations.

Text readability. To verify that the human generated adaptations simplify the source abstracts, we calculated the FKGL readability scores for both the adaptations and abstracts. FKGL scores were lower for the adaptations compared to the abstracts ($p < 0.0001$, Kendall's tau). It is important to note that FKGL does not measure similarity or content preservation, so additional metrics like BLEU, ROUGE, and SARI are needed to address this concern.

Inter-annotator agreement. To measure inter-annotator agreement, we used adaptions from the most experienced annotator (who also helped define the guidelines) as reference adaptations. Agreement was measured for all abstracts that were adapted by this annotator and another annotator. For the inter-annotator agreement metrics of ROUGE-1, ROUGE-2, ROUGE-L, BLEU-4, and BERTScore-F1, the values ranged from 0.4025–0.5801, 0.1267–0.2983, 0.2591–0.4689, 0.0680–0.2410, and 0.8305–0.9476, respectively, for all adaptations that were done by the reference annotator and another annotator. As the ROUGE-1 results show, the other annotators included, on average, about half of the words that the reference annotator used. As expected, ROUGE-2 values are lower, on average, because as n-grams increase in n, there will be less similarity between adaptations since individuals may use different combinations of words when creating new text.

We also calculated the similarity between human adaptations and the source abstracts. Using the abstracts as candidates and adaptations as references since BLEU-4 can only match multiple references to a single candidate and not vice versa, the scores in Table 3 show the adaptations contain over half of the same words, a third of the same bi-grams, and a large portion of the same subwords as the source abstracts.

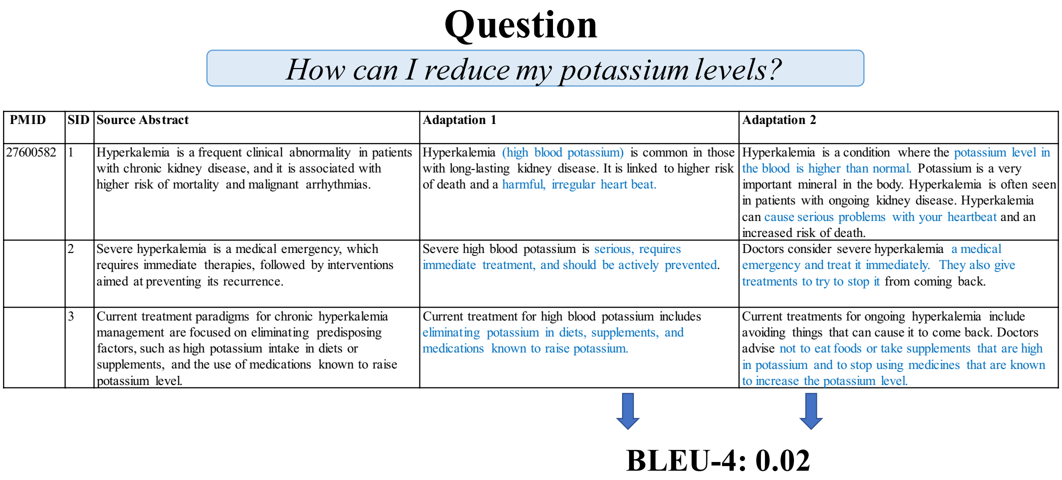
While ROUGE and BLEU are metrics for text similarity and BERTScore measures semantic similarity, they do not necessarily measure correctness. Even if a pair of adaptations have a low ROUGE, BLEU, or BERTScore score, both could be accurate restatements of the source abstract as seen in Fig. 3. While the BLEU-4 score can be low, both adaptations can relevantly describe the topic in response to the example question. The differences between the adaptations can be attributed to synonyms and differences in explanatory content. While BLEU and ROUGE are useful for measuring lexical similarity, calculating differences between adaptations like these is more nuanced. To address this issue, researchers are actively developing new metrics[37].

Experimental benchmarking. To benchmark the PLABA dataset and show its use in evaluating automatically generated adaptations, we used a variety of state-of-the-art deep learning algorithms listed below:

Text-to-text transfer transformer (T5). T5[38] is a transformer-based[39] encoder-decoder model with a bidirectional encoder setup similar to BERT[40] and an autoregressive decoder that is similar to the encoder except with a standard attention mechanism. Instead of training the model on a single task, T5 is pre-trained on a vast amount of data and on many unsupervised and supervised objectives, including token and span masking, classification, reading comprehension, translation, and summarization. The common feature of every objective is that the task can be treated as a language-generation task, in which the model learns to generate the proper textual output in response to the textual prompt included in the input sequence. As with other models, pre-training has been shown to achieve state-of-the-art results on many NLP tasks[37,38,41]. When the T5 model is fine-tuned on a specific dataset for a specific task, the task's objective (e.g., translate from English to French, summarize, etc.)

Comparison Type	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	BERTScore-F1
Adaptations vs Abstracts	0.58	0.36	0.50	0.39	0.91

Table 3. ROUGE-1, ROUGE-2, ROUGE-L, BLEU-4, and BERTScore-F1 using human adaptations as references and abstracts as candidates.



Question

How can I reduce my potassium levels?

PMID	SID	Source Abstract	Adaptation 1	Adaptation 2
27600582	1	Hyperkalemia is a frequent clinical abnormality in patients with chronic kidney disease, and it is associated with higher risk of mortality and malignant arrhythmias.	Hyperkalemia (high blood potassium) is common in those with long-lasting kidney disease. It is linked to higher risk of death and a harmful, irregular heart beat.	Hyperkalemia is a condition where the potassium level in the blood is higher than normal. Potassium is a very important mineral in the body. Hyperkalemia is often seen in patients with ongoing kidney disease. Hyperkalemia can cause serious problems with your heartbeat and an increased risk of death.
	2	Severe hyperkalemia is a medical emergency, which requires immediate therapies, followed by interventions aimed at preventing its recurrence.	Severe high blood potassium is serious, requires immediate treatment, and should be actively prevented.	Doctors consider severe hyperkalemia a medical emergency and treat it immediately. They also give treatments to try to stop it from coming back.
	3	Current treatment paradigms for chronic hyperkalemia management are focused on eliminating predisposing factors, such as high potassium intake in diets or supplements, and the use of medications known to raise potassium level.	Current treatment for high blood potassium includes eliminating potassium in diets, supplements, and medications known to raise potassium.	Current treatments for ongoing hyperkalemia include avoiding things that can cause it to come back. Doctors advise not to eat foods or take supplements that are high in potassium and to stop using medicines that are known to increase the potassium level.

BLEU-4: 0.02

Fig. 3 Example of the low BLEU-4 score between human adaptations from two different annotators created from the same source abstract and answering the same question. PMID refers to the PubMed ID from which the example originates from. SID refers to the sentence ID or number of the example sentence from the source abstract. Colored text in an adaptation represents parts of the adaptation that strongly differ from the other adaptation.

is prepended with a colon to the input text as a prompt to guide the T5 model during training and testing. In our experiments, we use the T5-Base model with the prompt "summarize:" since it is the closest prompt to the task of plain language adaptation that the T5 model was pre-trained on. We also show the performance of a T5 model not fine-tuned on our training data (T5-No-Fine-Tune) to compare it to a T5 model fine-tuned on PLABA to demonstrate the importance of training models on our dataset given recent developments in out-of-box or zero-shot settings[42,43].

Pre-training with extracted gap-sentences for abstractive summarization sequence-to-sequence (PEGASUS). PEGASUS[44] is another transformer-based encoder-decoder model; however, unlike T5, PEGASUS is pre-trained on a unique self-supervised objective. With this objective, entire sentences are masked from a document and collected as the output sequence for the remaining sentences of the document. In other words, PEGASUS is designed for abstractive summarization and similar tasks, achieving human performance on multiple datasets. In our experiments, we use the PEGASUS-Large model.

Bidirectional autoregressive transformer (BART). BART[45] is another transformer-based encoder-decoder that is pre-trained with a different objective. Instead of training the model directly on data with a text-to-text objective or summarization-specific objective, BART was pre-trained on tasks such as token deletion and masking, text-infilling, and sentence permutation. These tasks were developed to improve the model's ability to understand the content of text before summarizing or translating it. After this pre-training, BART can be fine-tuned for downstream tasks of summarization or translation with a more specific dataset to output higher quality text. These datasets include the CNN Daily Mail[46] dataset, a large news article dataset designed for summarization tasks. In our experiments, we use the BART-Base model and BART-Large model fine-tuned on the CNN Daily Mail dataset (BART-Large-CNN).

T Zero plus plus (T0PP). T0PP[47] is a variation of the original T5 encoder-decoder model created for zero-shot performance, or out-of-box performance on certain tasks and datasets without prior fine-tuning or training. To develop this zero-shot model, T0PP was trained on a subset of tasks (e.g., sentiment analysis, question answering) and evaluated on a different subset of tasks (e.g., natural language inference). In our experiments, we use the T0PP model with 3 billion parameters without fine-tuning on our dataset and with the same prompt "summarize:" as the T5 models to maintain consistency across prompt-based models.

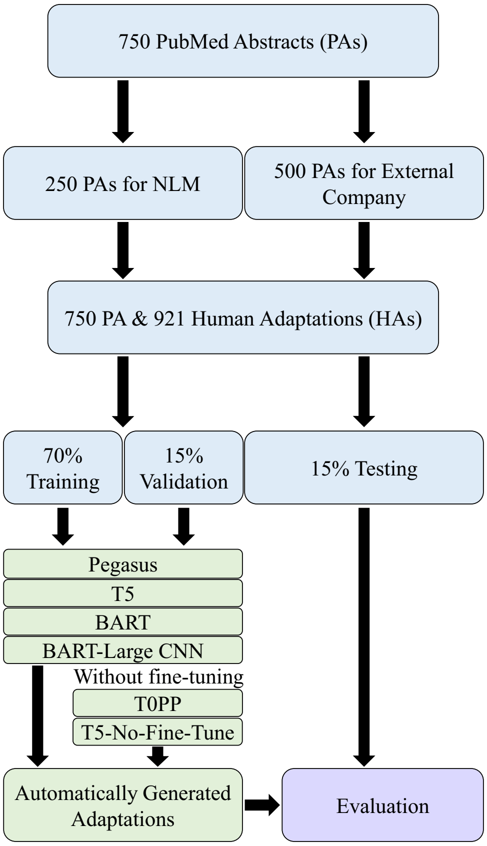


Fig. 4 Overview representing how PubMed abstracts and human adaptations are split for training and testing models.

	FKGL	
Data Type	**Average**	**S.d.**
Abstracts	15.78	8.06
Adaptations	12.04	2.39
T5	13.67/13.81/13.83	2.70/2.50/2.93
PEGASUS	13.77/13.80/13.84	2.80/2.98/2.80
BART-Base	13.52/13.06/13.37	2.88/2.52/2.59
BART-Large-CNN	13.43/12.74/13.19	2.54/2.64/2.66
T5-No-Fine-Tune	23.50	5.00
T0PP	15.22	3.52

Table 4. FKGL scores for automatically generated adaptations. The three results for models trained across three seeds are separated with slashes.

Algorithm Type	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	BERTScore-F1	SARI
T5	0.56/0.56/0.56	0.30/0.30/0.30	0.42/0.42/0.42	0.28/0.28/0.28	0.90/0.90/0.90	0.33/0.32/0.34
PEGASUS	0.57/0.57/0.57	0.32/0.32/0.31	0.43/0.44/0.43	0.31/0.30/0.30	0.90/0.90/0.90	0.30/0.31/0.32
BART-Base	0.56/0.56/0.56	0.30/0.31/0.31	0.42/0.42/0.43	0.29/0.29/0.29	0.90/0.90/0.90	0.29/0.34/0.33
BART-Large-CNN	0.57/0.56/0.58	0.32/0.30/0.31	0.43/0.41/0.42	0.30/0.28/0.29	0.90/0.90/0.90	0.34/0.30/0.26
T5-No-Fine-Tune	0.24	0.13	0.19	0.01	0.86	0.26
T0PP	0.22	0.12	0.17	0.01	0.86	0.25

Table 5. Automatically generated adaptations compared to human adaptations and (only for SARI) source abstracts. The three results for models trained across three seeds are separated with slashes.

Algorithm Type	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	BERTScore-F1
T5	0.77/0.77/0.77	0.61/0.61/0.62	0.73/0.73/0.74	0.54/0.55/0.55	0.94/0.94/0.94
PEGASUS	0.79/0.79/0.79	0.64/0.65/0.64	0.75/0.75/0.74	0.61/0.61/0.60	0.95/0.95/0.95
BART-Base	0.73/0.75/0.76	0.56/0.59/0.62	0.68/0.72/0.73	0.53/0.54/0.57	0.94/0.94/0.94
BART-Large-CNN	0.78/0.73/0.73	0.63/0.56/0.56	0.74/0.68/0.69	0.58/0.52/0.50	0.95/0.93/0.93
T5-No-Fine-Tune	0.38	0.31	0.34	0.02	0.88
T0PP	0.34	0.28	0.30	0.03	0.88

Table 6. Automatically generated adaptations compared to source abstracts. The three results for models trained across three seeds are separated with slashes.

Experimental setup. For our experiments, all deep learning models except for T0PP and T5-No-Fine-Tune were trained using the abstracts and adaptations in the PLABA dataset. Each PubMed abstract is used as the source document, and the human generated adaptations are used as the references. The dataset was divided such that 70% was used for training, 15% for validation, and 15% for testing. In addition, the stratified split was performed such that all abstracts and adaptations of each question were grouped and exclusively contained in the training, validation, or testing set. We utilized the pre-trained models from Hugging Face[48], and each trained model was fine-tuned with the AdamW optimizer and the default learning rate of 5e-5 for 20 epochs using V100X GPUs (32 GB VRAM) on a shared cluster. Maximum input sequence length was set to 512 tokens except for the BART models, in which the maximum was set to 1024. Validation loss was measured every epoch, and the checkpoint model with the lowest validation loss was used for test set evaluation. Each trained model was also randomly seeded with 3 different sets of initial parameters to assess model performance variability. In addition, the inputs and output of the models will vary between training and testing. If a model is being trained, its two inputs per training step will be the source abstract and its respective human generated adaptations. The output is the model's automatically generated adaptation, which will be compared to the human generated adaptation to evaluate how close the output is to that input. The model is rewarded for how similar the output is to the gold-standard human generated adaptation. While training occurs with the training dataset, the model is periodically evaluated with the validation set to monitor performance during training. If a model is being tested, its input will just be the source abstract, while its output continues to be the model's automatically generated adaptation. All metrics (except SARI) will compare the output to the human generated adaptations to calculate the score. For SARI, this metric will compare the output to the human generated adaptations and source abstract to generate a score. While trained models will first be trained on the training and validation sets and then tested on the test set, zero-shot models like T0PP and T5-No-Fine-Tune will skip training and immediately be tested on the test set. A visual overview of the experiments can be seen in Fig. 4.

Results

Table 4 shows the FKGL scores between the automatically generated adaptations, all of which were significantly lower than the abstracts except from T5-No-Fine-Tune and significantly higher than the manually crafted adaptations except from T5-No-Fine-Tune ($p < 0.05$, Kendall's tau). Table 5 shows the comparison between the automatically generated adaptations and the human generated adaptations with ROUGE and BLEU and the comparison between the automatically generated adaptations, human generated adaptations, and source abstracts with SARI. Table 6 shows the comparison between the automatically generated adaptations and the source abstracts with ROUGE and BLEU. It is interesting to note that the automatically generated adaptations from the trained models and T0PP are more readable than the abstracts but less readable than the human generated adaptations according to FKGL scores. However, the T5 variant without fine-tuning generated adaptations less readable than even the source abstracts. Thus, the dataset gives the models sufficient training data to develop outputs that outperform the source abstracts in terms of readability. Regarding SARI, the trained models tend to perform comparably in terms of simplification. In terms of ROUGE, BLEU, and BERTScore, the automatically generated adaptations tend to share more n-grams and subwords with the source abstracts rather than the human generated adaptations. This relationship is potentially because the abstracts tend to be shorter than the adaptations, as seen in Table 2. This may make it easier for the automatically generated adaptations to share more contiguous word sequences with the abstracts relative to the human generated adaptations. In addition, the choice of metrics used for evaluation will influence the reported performance of a model. However, across all metrics in Tables 4, 5, both zero-shot models T5-No-Fine-Tune and T0PP performed significantly worse compared to the trained models ($p < 0.0001$, Wilcoxon signed-rank test).

An example of the automatically generated adaptations from each model in response to the same abstract is shown in Table 7. The generated adaptations from the zero-shot models show visibly fewer sentences, less details, and less explanations than generated adaptations from the trained models. These demonstrate that the PLABA dataset, in addition to being a high-quality test set, is useful for training generative deep learning models with the objective of text adaptation of scientific articles. Since there are no existing manually crafted datasets for this objective, PLABA can be a valuable dataset for benchmarking future research in this domain.

Algorithm Type	Generated Text
Question	*How to treat a bakers cyst?*
AbstractAbstract	Purpose: Osteoarthritis of the knee is a common disease, often associated with a Baker's cyst. Besides osteoarthritis, also other joint pathologies of the knee can be […] A decrease in volume of more than 25% compared to the baseline could be achieved for 75% of the patients in the short-term and 79% of the patients in the long-term follow up. Conclusion: Radiotherapy of knee osteoarthritis is an effective treatment that decreases the volume of a Baker's cyst. Most patients respond to the treatment. Whether radiotherapy is an effective treatment for Baker's cyst without associated osteoarthritis has to be further examined.
PEGASUS	Osteoarthritis of the knee is a common disease, often associated with a Baker's cyst. Besides osteoarthritis, also other joint diseases (diseases that cause pain and swelling in the joints) can be […] A decrease in volume more than 25% compared to the baseline could be achieved for 75% of patients. Most patients respond to treatment. Radiation therapy of knee osteoarthritis (damage to cartilage and bone) is effective that decreases the number of affected knees. However, whether radiotherapy is the best treatment is still being studied.
T5	Osteoarthritis of the knee (knee arthritis) is a common disease. It is often associated with the Baker's cyst (a cyst in the back of your knee). Besides arthritis, also other joint diseases of knees […] A decrease of more than 25% compared to the baseline could be achieved for 75% of patients in short term and 79% of people in long term follow up. Conclusion: Radiation therapy is an efficient treatment that decreases the amount of swollen area. Most patients respond to this treatment. Whether radiation is effective for Baker'cyst without associated arthritis must be further examined.
BART-Base	Osteoarthritis of the knee is a common disease, often associated with a Baker's cyst (a small, bone-related lump). Also, other joints in the leg can be causative […] A decrease of more than 25% compared to the baseline could be achieved for 75% of patients at the beginning and 79% (long-lasting) after the treatment. This conclusion is that radiation therapy of knee osteosarcoma is an important treatment that decreases the amount and severity of an osteoporosis cysteostasis. Most patients respond to it. However, it is still unclear how effective this treatment is.
BART-Large-CNN	Osteoarthritis of the knee is a common disease, often associated with a Baker's cyst, a growth of swollen tissues in the joints. Besides bone pain, other problems with the ligament that runs between the thigh bone and knee joint can cause […] A decrease in volume of more than 25% compared to the baseline could be achieved for 75% of patients in both the short term and the long term. In conclusion, radiation therapy of knee bone arthritis is an helpful treatment that decreases the size of an acchly Cyst. Most patients respond to treatment. Whether radiation is the best treatment to treat a Bakty for a kyst without the associated pain caused by bone damage is further examined.
T5-No-Fine-Tune	radiotherapy is an effective treatment for osteoarthritis, with an anti-inflammatory effect. the excessive production of synovia usually is associated with intraarticular inflammation. a prospective trial was carried out, including 20 knees receiving radio therapy for knee arthritis.
T0PP	Low-dose radiotherapy of knee osteoarthritis is an effective treatment that decreases the volume of a Baker's cyst. Most patients respond to the treatment.

Table 7. Examples of adaptations created by PEGASUS, T5, BART-Base, BART-Large-CNN, T5-No-Fine-Tune, T0PP.

Usage Notes

We have added instructions in the README file of our OSF repository that show how to use the PLABA dataset. Pre-processing the dataset and evaluating adaptation algorithms on it can be located in the code scripts at our GitHub repository given below. To reproduce the experimental results, users can download the data from the OSF repository, download the code scripts from the GitHub repository, and run the code scripts on their machine to train and benchmark the models with the same results.

Code availability

Code scripts to pre-process PLABA, reproduce the benchmark results of the experiments, and train and test additional models can be found at https://doi.org/10.5281/zenodo.7429310, a Zenodo DOI[49] containing a static release of our GitHub repository.

Received: 19 August 2022; Accepted: 19 December 2022;
Published online: 04 January 2023

References

1. MedlinePlus - Health Information from the National Library of Medicine.
2. Rosenberg, S. A. *et al.* Online patient information from radiation oncology departments is too complex for the general population. *Practical Radiation Oncology* **7**, 57–62, https://doi.org/10.1016/j.prro.2016.07.008 (2017).
3. Stableford, S. & Mettger, W. Plain language: a strategic response to the health literacy challenge. *Journal of public health policy* **28**, 71–93 (2007).
4. Xu, W., Napoles, C., Pavlick, E., Chen, Q. & Callison-Burch, C. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* **4**, 401–415, https://doi.org/10.1162/tacl_a_00107 (2016).
5. Carlo, M. S. *et al.* Closing the gap: Addressing the vocabulary needs of english-language learners in bilingual and mainstream classrooms. *Reading research quarterly* **39**, 188–215 (2004).
6. White, R. W. & Horvitz, E. Cyberchondria: Studies of the escalation of medical concerns in Web search. *ACM Trans. Inf. Syst.* **27**, 23:1–23:37, https://doi.org/10.1145/1629096.1629101 (2009).
7. Cochrane Handbook for Systematic Reviews of Interventions.
8. Kauchak, D. & Leroy, G. A web-based medical text simplification tool. In *53rd Annual Hawaii International Conference on System Sciences, HICSS 2020*, 3749–3757 (IEEE Computer Society, 2020).
9. Stahlberg, F. Neural machine translation: A review. *Journal of Artificial Intelligence Research* **69**, 343–418 (2020).
10. Al-Thanyyan, S. S. & Azmi, A. M. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)* **54**, 1–36 (2021).
11. Savery, M., Abacha, A. B., Gayen, S. & Demner-Fushman, D. Question-driven summarization of answers to consumer health questions. *Scientific Data* **7**, 1–9 (2020).
12. Jiang, C., Maddela, M., Lan, W., Zhong, Y. & Xu, W. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7943–7960 (2020).

13. Coster, W. & Kauchak, D. 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 (2011).
14. Hwang, W., Hajishirzi, H., Ostendorf, M. & Wu, W. 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, https://doi.org/10.3115/v1/N15-1022 (Association for Computational Linguistics, Denver, Colorado, 2015).
15. Zhu, Z., Bernhard, D. & Gurevych, I. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1353–1361 (2010).
16. Van, H., Kauchak, D. & Leroy, G. AutoMeTS: The Autocomplete for Medical Text Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1424–1434, https://doi.org/10.18653/v1/2020.coling-main.122 (International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020).
17. Van den Bercken, L., Sips, R.-J. & Lofi, C. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, 3286–3292 (2019).
18. Adduru, V. *et al*. Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In *KHD@ IJCAI* (2018).
19. Cardon, R. & Grabar, N. Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *RANLP 2019* (2019).
20. Xu, W., Callison-Burch, C. & Napoles, C. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics* **3**, 283–297, https://doi.org/10.1162/tacl_a_00139. Place: Cambridge, MA Publisher: MIT Press (2015).
21. Shardlow, M. & Nawaz, R. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 380–389, https://doi.org/10.18653/v1/P19-1037 (Association for Computational Linguistics, Florence, Italy, 2019).
22. Cao, Y. *et al*. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1061–1071 (2020).
23. Pattisapu, N., Prabhu, N., Bhati, S. & Varma, V. Leveraging Social Media for Medical Text Simplification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 851–860 (2020).
24. Štajner, S., Sheang, K. C. & Saggion, H. Sentence simplification capabilities of transfer-based models. *Proceedings of the AAAI Conference on Artificial Intelligence* (2022).
25. Sakakini, T. *et al*. Context-Aware Automatic Text Simplification of Health Materials in Low-Resource Domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, 115–126 (2020).
26. Friedman, C., Kra, P. & Rzhetsky, A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* **35**, 222–235, https://doi.org/10.1016/S1532-0464(03)00012-1 (2002).
27. Basu, C., Vasu, R., Yasunaga, M., Kim, S. & Yang, Q. Automatic medical text simplification: Challenges of data quality and curation. In *HUMAN@ AAAI Fall Symposium* (2021).
28. Ondov, B., Attal, K. & Demner-Fushman, D. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association* **29**, 1976–1988 (2022).
29. Frankenberg-Garcia, A. A corpus study of splitting and joining sentences in translation. *Corpora* **14**, 1–30 Publisher: Edinburgh University Press The Tun-Holyrood Road, 12 (2f) Jackson's Entry… (2019).
30. Deardorff, A., Masterton, K., Roberts, K., Kilicoglu, H. & Demner-Fushman, D. A protocol-driven approach to automatically finding authoritative answers to consumer health questions in online resources. *Journal of the Association for Information Science and Technology* **68**, 1724–1736, https://doi.org/10.1002/asi.23806 (2017).
31. Attal, K., Ondov, B. & Demner, D. A dataset for plain language adaptation of biomedical abstracts. OSF, https://doi.org/10.17605/OSF.IO/RNPMF (2022).
32. Flesch, R. A new readability yardstick. *Journal of Applied Psychology* **32**, 221–233, https://doi.org/10.1037/h0057532. Place: US Publisher: American Psychological Association (1948).
33. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318, https://doi.org/10.3115/1073083.1073135 (Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002).
34. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004).
35. Sun, R., Jin, H. & Wan, X. Document-Level Text Simplification: Dataset, Criteria and Baseline. *arXiv:2110.05071 [cs]*. ArXiv: 2110.05071 (2021).
36. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
37. Kryscinski, W., McCann, B., Xiong, C. & Socher, R. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346, https://doi.org/10.18653/v1/2020.emnlp-main.750 (Association for Computational Linguistics, Online, 2020).
38. Raffel, C. *et al*. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*. ArXiv: 1910.10683 (2020).
39. Vaswani, A. *et al*. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
40. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805 (2019).
41. Radford, A. *et al*. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
42. Goodwin, T. R., Savery, M. E. & Demner-Fushman, D. Flight of the pegasus? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *Proceedings of COLING. International Conference on Computational Linguistics*, vol. 2020, 5640 (NIH Public Access, 2020).
43. Goodwin, T. R., Savery, M. E. & Demner-Fushman, D. Towards zero-shot conditional summarization with adaptive multi-task fine-tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2020, 3215 (NIH Public Access, 2020).
44. Zhang, J., Zhao, Y., Saleh, M. & Liu, P. J. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv:1912.08777 [cs]* ArXiv: 1912.08777 (2020).
45. Lewis, M. *et al*. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880, https://doi.org/10.18653/v1/2020.acl-main.703 (Association for Computational Linguistics, Online, 2020).
46. Nallapati, R., Zhou, B., Santos, C. N. D., Gulcehre, C. & Xiang, B. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *arXiv:1602.06023 [cs]* ArXiv: 1602.06023 version: 5 (2016).
47. Sanh, V. *et al*. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* (2021).
48. Wolf, T. *et al*. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6 (Association for Computational Linguistics, Online, 2020).
49. Attal-Kush, attal-kush/PLABA: v1.0.0, *Zenodo*, https://doi.org/10.5281/ZENODO.7429310 (2022).

Acknowledgements

This work was supported by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Author contributions

K.A. created the code scripts for data pre-processing and deep learning experiments, contributed to adaptation guidelines, contributed to creating the manual adaptations, and wrote and edited the manuscript. B.O. contributed to adaptation guidelines and edited the manuscript. D.D.-F. conceived the project, edited the manuscript, and provided feedback at all stages of the study.

Funding

Open Access funding provided by the National Institutes of Health (NIH).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.