Section-level Simplification of Biomedical Abstracts

Jan Bakker and **Jaap Kamps**
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
{j.bakker, kamps}@uva.nl

Abstract

Cochrane produces systematic reviews whose abstracts are divided into seven standard sections. However, the plain language summaries (PLS) of Cochrane reviews do not adhere to the same structure, which has prevented researchers from training simplification models on paired abstract and PLS sections. In this work, we devise a two-step method to automatically divide PLS of Cochrane reviews into the same sections in which abstracts are divided. In the first step, we align each sentence in a PLS to a section in the parallel abstract if they cover similar content. In the second step, we classify the remaining sentences into sections based on the content of the PLS and what we learned from the first step. We manually divide 22 PLS into sections to evaluate our method. Upon execution of our method, we obtain the COCHRANE-SECTIONS dataset, which consists of paired abstract and PLS sections in English for a total of 7.7K Cochrane reviews. Thus, our work yields references for the section-level simplification of biomedical abstracts.

1 Introduction

Cochrane is an international organization dedicated to producing reliable health evidence in the form of systematic reviews.[1] A *systematic review* attempts to identify, appraise and synthesize all empirical evidence from relevant studies to answer a specific (health) question.[2] Systematic reviews produced by Cochrane contain an abstract, which should be targeted primarily at healthcare decision makers rather than researchers (Cumpston et al., 2023), but frequently includes jargon and other types of complex language. To make their findings more accessible, Cochrane reviews are also accompanied by a plain language summary (PLS).

Abstract section

Orthognathic surgery (OS) is a term that refers to many elective surgical techniques to correct facial deformity; the associated malocclusion and functional disorders related to the stomatognathic system. Whilst such surgery is classed as "clean-contaminated", the usefulness of and the most appropriate regimen for antibiotic prophylaxis in these patients are still debated.

PLS section

Many people undergo surgery of the jaws each year to correct malformations. Whilst a risk of infection following surgery has been noted, no agreement has been reached regarding how useful antibiotics are for infection prevention and what type and dose of antibiotic should be used.

Figure 1: A pair of *Background* sections from COCHRANE-SECTIONS.

Plain language summaries of Cochrane reviews are written directly from the full review; they are not simplified versions of the technical abstracts. Even so, the quality, quantity, and availability of these abstract-PLS pairs have motivated NLP researchers to use them for the training and evaluation of text simplification models (Guo et al., 2021; Devaraj et al., 2021; Joseph et al., 2023; Bakker and Kamps, 2024). The objective of such models is to transform (part of) a technical abstract into (part of) the corresponding PLS – the underlying idea being that if they perform well on this task, they should also be able to make other sources in biomedicine more accessible.

The technical abstract of a Cochrane review is usually divided into seven standard sections, from *Background* to *Authors' conclusions*. PLS are structured heterogeneously, which has prevented the pairing of abstract sections with corresponding PLS sections. Instead, researchers have trained and evaluated their models on the full abstract-PLS pairs (Guo et al., 2021), on paired subsets describing only the studies and results (Devaraj et al., 2021; Bakker and Kamps, 2024) and on aligned sentence pairs (Joseph et al., 2023). These approaches have

[1] https://www.cochrane.org/about-us

[2] https://www.cochranelibrary.com/about/about-cochrane-reviews

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13819–13833
November 4-9, 2025 ©2025 Association for Computational Linguistics

clear drawbacks compared to section-level simplification.

In this work, we devise a two-step method to automatically divide plain language summaries of Cochrane reviews into the same sections in which abstracts are divided, so that we can use the resulting PLS sections as references for simplification of the parallel abstract sections. In the first step, we align each sentence in a PLS to a section in the parallel abstract if they cover similar content; we leverage the neural CRF aligner of Jiang et al. (2020) for this purpose. In the second step, we classify the remaining sentences into sections based on the content of the PLS and what we learned from the first step. Moreover, we manually divide 22 PLS into sections to evaluate our method. Upon execution of our method, we obtain the COCHRANE-SECTIONS dataset, which consists of paired abstract and PLS sections in English for a total of 7.7K Cochrane reviews. An example is shown in Figure 1. Thus, our work yields references for the section-level simplification of biomedical abstracts.

2 Related work

In this section, we compare our work to existing work on the creation of corpora for the simplification of biomedical texts. We cover the creation of datasets that are derived from the abstracts and PLS of Cochrane reviews and those that are not. We also discuss related work on sentence alignment in text simplification, section-level text simplification, and sequential sentence classification.

2.1 Cochrane datasets

The Cochrane Database of Systematic Reviews[3] (CDSR) comprises thousands of systematic reviews of research in health care and health policy. Each review contains both a technical abstract and a PLS, which are written in English but often translated into multiple languages. These abstract-PLS pairs are freely available, and various NLP researchers have used them for the training and evaluation of text simplification models.

2.1.1 Descriptions of existing work

Grabar and Cardon (2018) derived 3.8K abstract-PLS pairs in French from the CDSR. They included them in the CLEAR corpus and let annotators align comparable sentences between a subset of 13 pairs.

Guo et al. (2021) introduced the task of generating PLS for Cochrane reviews based on their technical abstracts. They derived 7.8K abstract-PLS pairs in English from the CDSR and experimented with fine-tuning several document summarization models on them.

Devaraj et al. (2021) also compiled a corpus of source-reference pairs in English from the CDSR, but they used only the *Main results* and *Authors' conclusions* sections of each technical abstract as the source. They applied substring matching to the PLS to approximate the location of the first section, paragraph, or sentence (depending on its structure) describing the studies and results, and kept everything from that point onward as the reference.

Joseph et al. (2023) asked annotators to align similar sentences in 101 English abstract-PLS pairs from the CDSR. They then trained and evaluated a neural CRF model (Jiang et al., 2020) on these manually aligned instances and used it to automatically align sentences in all remaining 7.7K pairs. They also took advantage of the fact that these texts are often translated into multiple languages, so that they could train multilingual sentence simplification models on their aligned data. They shared their data in the MULTICOCHRANE corpus.

Lastly, Bakker and Kamps (2024) created the Cochrane-auto corpus. They automatically aligned similar sentences in an updated version of the corpus from Devaraj et al. (2021), and concatenated aligned PLS sentences to provide references not only at the sentence-level but also at the paragraph- and document-level.

2.1.2 Comparison to our work

In this work, we create a dataset of paired abstract and PLS sections. Thereby, we strike a balance between the document-level approach of Guo et al. (2021) on the one hand and the sentence-level approach of Joseph et al. (2023) on the other. We provide any model trained on our dataset with more context than a single sentence, but we do not require it to learn how to simplify a full abstract at once. Although the same can be said about the approaches of Devaraj et al. (2021) and Bakker and Kamps (2024), we argue that simplifying not only the results and conclusions sections, but at least the background section as well, is key to making the findings of a systematic review more understandable. We also expect that the types of simplifications learned by models trained on our dataset will better generalize. We are the first to provide refer-

[3] https://www.cochranelibrary.com/cdsr/reviews

ences for the simplification of just the *Background* section, and its contents are the least specific to the structure of a systematic review. Thus, models trained to simplify this section should also be able to make other sources in biomedicine more accessible.

2.2 Biomedical text simplification datasets

There are also other datasets for the training and evaluation of biomedical text simplification models; here, we discuss some datasets of English texts. On the one hand, there are datasets that contain abstracts of research papers from the biomedical domain paired with manual simplifications (Ermakova et al., 2022; Attal et al., 2023). These datasets are very valuable, but limited in size and restricted to sentence-level simplifications. On the other hand, there are corpora that comprise paired abstracts and lay summaries of studies published in biomedical journals like PNAS and PLOS (Guo et al., 2024). They are large enough to train text simplification models on, but the abstracts and lay summaries in these corpora are generally written with less care than those in the CDSR – simply because performing these studies requires less effort than undertaking a systematic review.

2.3 Sentence alignment in text simplification

To date, most research on automatic text simplification has focused on sentence-level inputs, mainly due to the availability of large-scale sentence-aligned training data (Alva-Manchego et al., 2020). Jiang et al. (2020) invented a neural CRF alignment model to extract such training data from parallel texts. When provided with a complex-simple text pair as input, their model automatically aligns each sentence in the simple text to either one or zero corresponding sentences in the complex text. In doing so, it leverages the similar order of content between parallel texts and utilizes a fine-tuned BERT model to capture the semantic similarity between complex-simple sentence pairs. The authors fine-tuned their models on manually aligned article pairs from Newsela (Xu et al., 2015) and Wikipedia (Zhang and Lapata, 2017). Their pretrained models remain popular in the field of text simplification, because the in-domain fine-tuning on complex-simple text pairs makes them especially suitable for sentence alignment tasks in this field (Joseph et al., 2023; Maddela and Alva-Manchego, 2025).

2.4 Section-level text simplification

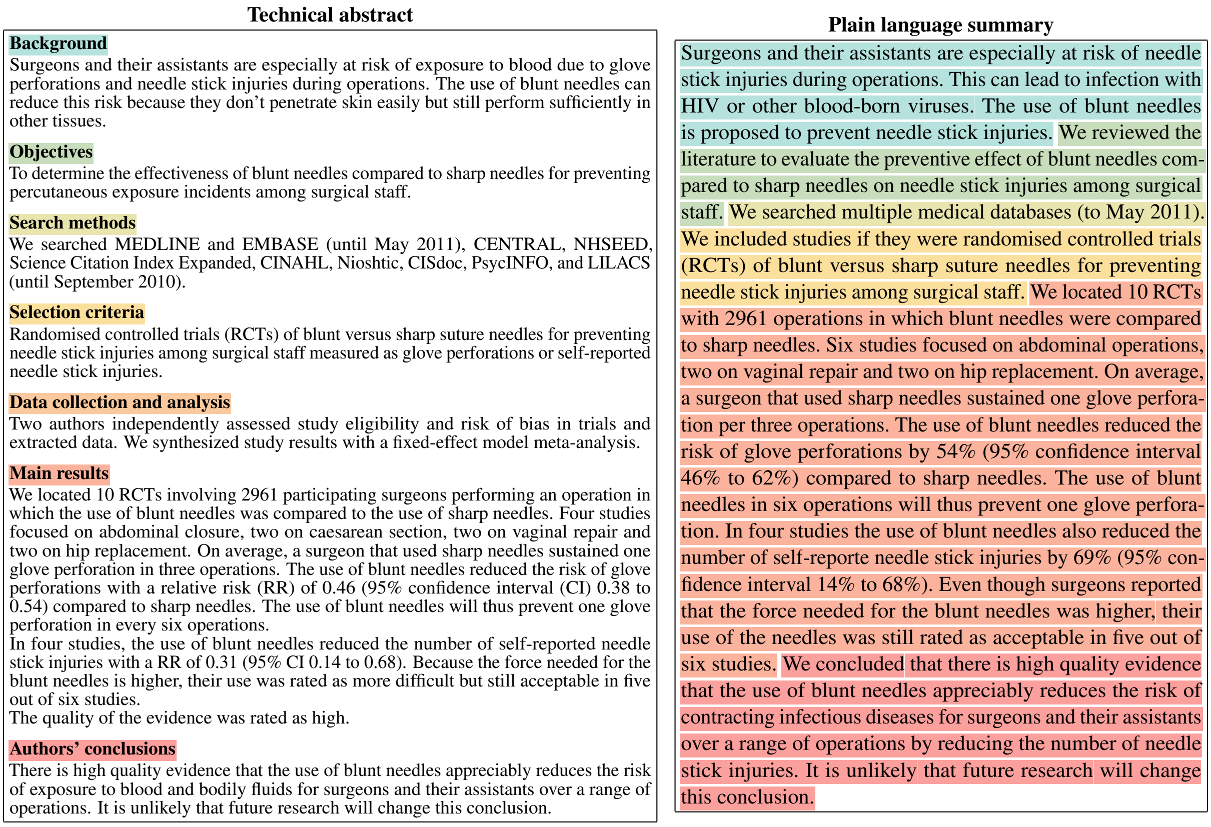
Despite the focus on sentence-level inputs, the simplification of documents cannot be tackled by naively simplifying each sentence in isolation, as this approach fails to preserve the discourse structure of the document and account for the possibility of multi-sentence operations (Alva-Manchego et al., 2019). At the same time, modern sequence-to-sequence models like the Transformer (Vaswani et al., 2017) struggle with document-level inputs due to the quadratic attention complexity, which also leads to increased memory and computation requirements. Therefore, in this work, we lay the foundation for a section-level approach to document simplification.

Our work is empirically motivated by the work of Cripwell et al. (2023), who demonstrated that text simplification models that take as input either a sentence or a whole document underperform compared to models that operate at the paragraph-level. Most recently, Fang et al. (2025) proposed a progressive document simplification method based on ChatGPT. The first step of their method is to divide the document into sections; in the next steps, each section is simplified independently. The authors showed that their method significantly outperforms the baseline methods that simplify the entire document in one step.

2.5 Sequential sentence classification

We divide PLS into sections by labeling each sentence with a section header. The value of such high-level sectioning of scholarly texts has been widely studied in NLP, with the ‘argumentative zoning’ of Teufel et al. (2009) as an influential example. They asked annotators to label the sentences in scientific papers by their rhetorical status.

For the biomedical domain, Dernoncourt and Lee (2017) created a corpus with 200K abstracts of randomized controlled trials, which are divided into five standardized sections (*Background*, *Objective*, *Method*, *Results* and *Conclusion*). Kim et al. (2011) asked annotators to label all sentences in the abstracts of 1,000 systematic reviews with one of six categories (*Background*, *Population*, *Intervention*, *Outcome*, *Study Design* and *Other*). Various researchers have used these and similar datasets to train and evaluate models for sequential sentence classification: the task of assigning each sentence in a text a label, such as a section header, that indicates its role in the text (Jin and Szolovits, 2018;



Technical abstract

Background
Surgeons and their assistants are especially at risk of exposure to blood due to glove perforations and needle stick injuries during operations. The use of blunt needles can reduce this risk because they don't penetrate skin easily but still perform sufficiently in other tissues.

Objectives
To determine the effectiveness of blunt needles compared to sharp needles for preventing percutaneous exposure incidents among surgical staff.

Search methods
We searched MEDLINE and EMBASE (until May 2011), CENTRAL, NHSEED, Science Citation Index Expanded, CINAHL, Nioshtic, CISdoc, PsycINFO, and LILACS (until September 2010).

Selection criteria
Randomised controlled trials (RCTs) of blunt versus sharp suture needles for preventing needle stick injuries among surgical staff measured as glove perforations or self-reported needle stick injuries.

Data collection and analysis
Two authors independently assessed study eligibility and risk of bias in trials and extracted data. We synthesized study results with a fixed-effect model meta-analysis.

Main results
We located 10 RCTs involving 2961 participating surgeons performing an operation in which the use of blunt needles was compared to the use of sharp needles. Four studies focused on abdominal closure, two on caesarean section, two on vaginal repair and two on hip replacement. On average, a surgeon that used sharp needles sustained one glove perforation in three operations. The use of blunt needles reduced the risk of glove perforations with a relative risk (RR) of 0.46 (95% confidence interval (CI) 0.38 to 0.54) compared to sharp needles. The use of blunt needles will thus prevent one glove perforation in every six operations.
In four studies, the use of blunt needles reduced the number of self-reported needle stick injuries with a RR of 0.31 (95% CI 0.14 to 0.68). Because the force needed for the blunt needles is higher, their use was rated as more difficult but still acceptable in five out of six studies.
The quality of the evidence was rated as high.

Authors' conclusions
There is high quality evidence that the use of blunt needles appreciably reduces the risk of exposure to blood and bodily fluids for surgeons and their assistants over a range of operations. It is unlikely that future research will change this conclusion.

Plain language summary

Surgeons and their assistants are especially at risk of needle stick injuries during operations. This can lead to infection with HIV or other blood-born viruses. The use of blunt needles is proposed to prevent needle stick injuries. We reviewed the literature to evaluate the preventive effect of blunt needles compared to sharp needles on needle stick injuries among surgical staff. We searched multiple medical databases (to May 2011). We included studies if they were randomised controlled trials (RCTs) of blunt versus sharp suture needles for preventing needle stick injuries among surgical staff. We located 10 RCTs with 2961 operations in which blunt needles were compared to sharp needles. Six studies focused on abdominal operations, two on vaginal repair and two on hip replacement. On average, a surgeon that used sharp needles sustained one glove perforation per three operations. The use of blunt needles reduced the risk of glove perforations by 54% (95% confidence interval 46% to 62%) compared to sharp needles. The use of blunt needles in six operations will thus prevent one glove perforation. In four studies the use of blunt needles also reduced the number of self-reporte needle stick injuries by 69% (95% confidence interval 14% to 68%). Even though surgeons reported that the force needed for the blunt needles was higher, their use of the needles was still rated as acceptable in five out of six studies. We concluded that there is high quality evidence that the use of blunt needles appreciably reduces the risk of contracting infectious diseases for surgeons and their assistants over a range of operations by reducing the number of needle stick injuries. It is unlikely that future research will change this conclusion.

Figure 2: The shortest abstract in the test set, along with the corresponding PLS (Saarto et al., 2011). The colors indicate the labels that we assign to each sentence in the PLS; our manually annotated and automatically predicted labels are the same.

Cohan et al., 2019; Lan et al., 2024).

Our work differs from earlier work on the topic in that we are given pairs of comparable texts, one of which is structured into standard sections (the abstract) while the other is not (the PLS). We leverage the similarities between these text pairs to divide each unstructured text into the same sections that the parallel structured text is divided into.

3 Data

In this section, we describe the structure of the abstracts, the plain language summaries, and the characteristics of the data that we work with. An illustrative example of an abstract-PLS pair from the CDSR is shown in Figure 2.

3.1 Outline of a technical abstract

The authors of a Cochrane review normally write the abstract as follows (Page et al., 2021). First, they justify their review in the context of existing knowledge (*Background*) and declare its objectives (*Objectives*). Then they specify the strategies used to identify relevant studies (*Search methods*), the criteria used to determine which studies are included in the review (*Selection criteria*), and the methods used to collect and synthesize data from included studies (*Data collection and analysis*). Next, the authors describe the characteristics, results, and limitations of the included studies, as well as the data synthesis results (*Main results*). Lastly, the authors provide a general interpretation of the review outcomes (*Authors' conclusions*).

3.2 Plain language summaries

The plain language summary (PLS) of a Cochrane review is usually written by the review authors. It should clearly convey the questions and key findings of the review, using plain language that can be understood by most readers without university education (McIlwain et al., 2013). As such, it should avoid technical terms and jargon where possible and explain them otherwise. Cochrane first published their standards for reporting PLS in 2013. However, as demonstrated by Jelicic Kadic et al. (2016), PLS of Cochrane reviews are in practice heterogeneous with low adherence to these standards. In 2022, Cochrane introduced a new template and new guidelines, following a one-year pilot

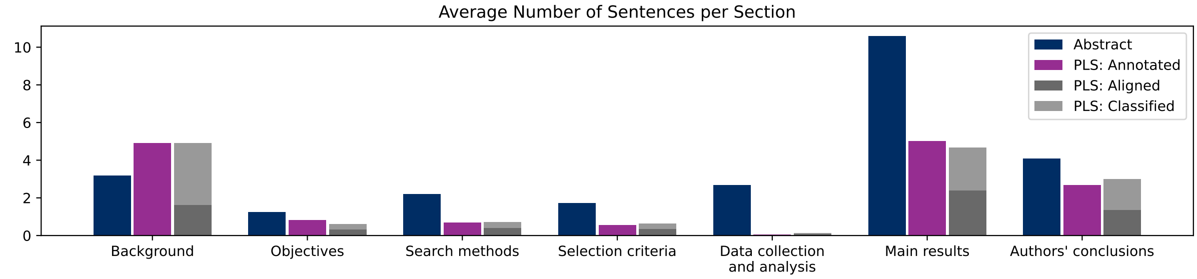


Figure 3: A barplot that shows the distribution of sentences over sections in the test set. Blue = abstract sentences; Purple = PLS sentences according to our annotations; Gray = PLS sentences according to our predictions.

study to improve the quality of their PLS (Pitcher et al., 2022).

3.3 Data characteristics

We aim to divide the PLS of Cochrane reviews into the same seven sections in which their technical abstracts are divided. To do so, we make use of the same dataset that Joseph et al. (2023) used to create the MULTICOCHRANE corpus. It consists of 7,703 abstract-PLS pairs[4] in English, derived from the Cochrane Database of Systematic Reviews by Devaraj et al. (2021) around March 2020.

On average, the abstracts in this dataset contain more sentences (25.0 $>$ 14.8), and more tokens per sentence (32.4 $>$ 26.4), than the PLS. The dark blue bars in Figure 3 show the average number of sentences per abstract section in the test split. All PLS in our dataset were written before the introduction of the updated reporting guidelines in 2022. Consequently, 44% of our PLS are split into sections with a variety of headings, while the other 56% consist of one or more paragraphs.

We split our dataset in the same way as the creators of the MULTICOCHRANE corpus. They divided their manually aligned subset, which they called MC-CLEAN, into train/validation/test splits of 74/5/22 pairs. Subsequently, the authors trained an alignment model on these data and used it to automatically align similar sentences in the remaining 7,602 abstract-PLS pairs. They called this subset MC-NOISY; we will refer to it as the auto split.

4 Objective

We approach the task of dividing PLS into sections as a sequence labeling task, where we need to label each sentence in a PLS with exactly one section header from Section 3.1, such that all sentences with the same label together, in their original order, form that section. Our objective is to perform this task in such a way that the resulting PLS sections constitute the best possible references for simplification of the parallel abstract sections.

Concretely, this means that we should try to label each sentence in a PLS with the header of the section in the parallel abstract that covers the same content – although likely in more technical terms. This will not always be possible, as the PLS might include elaborations (e.g., the second sentence in Figure 2) and additional information from the review. There will also be cases where a PLS sentence contains content that is distributed over, or discussed in, multiple abstract sections.[5] In such cases, we can assign only one label to that sentence, and this is a limitation of our approach.

If a PLS sentence can be aligned to a section in the parallel abstract based on its meaning, we should label it with the header of that section. If its content is not present in the technical abstract, we should instead determine the label of a PLS sentence based on how well it fits each of the section descriptions given in Section 3.1, as well as the labels of the surrounding sentences. In particular, when one sentence refers to another, we should probably assign both sentences the same label to form coherent PLS sections. Lastly, if a sentence can be aligned to multiple abstract sections, we should align it to the section that covers the largest part of its content, in the most similar terms. Our complete guidelines can be found in Appendix D.

[4] We excluded 52 instances in which the technical abstract does not adhere to the standard format.

[5] For instance, the second to last PLS sentence in Figure 2 contains content that is distributed over the conclusions (*We concluded ... of operations*) and results (*by reducing ... stick injuries*) sections of the technical abstract.

5 Annotation

Based on our annotation guidelines, we manually annotate the sentences in every PLS in our test set with a section header. Our main annotator is the first author of this paper: a Dutch PhD student with a professional command of English. They have no medical background, but a thorough understanding of the task (as they were the one to come up with it in the first place). In addition, they tried their best to perform the task as well as possible; the colors in Figure 2 indicate the labels that they assigned to each sentence in the illustrative PLS. We use the section header labels assigned by our main annotator for all analysis and evaluation. As such, we refer to them as our annotations.

To demonstrate the reliability of our annotations, we also have the second author of this paper label all 323 sentences in our test set. They follow the same guidelines, but spend less time on the task. Inter-annotator agreement measured by Cohen's kappa is 0.82; inter-annotator accuracy is 87.3%. Furthermore, we asked a scientist who works at Cochrane to annotate a subset of 138 sentences for 10 abstract-PLS pairs. Their agreement with the main annotator is 0.80 in terms of Cohen's kappa, and none of the disagreements seem to be dependent on medical interpretation.

In general, we find that it is relatively easy to determine the right label for most sentences. However, we also find that the real alignments between abstracts and PLS may not always reside at the sentence-level. The fact that we can only assign one label to each sentence in a PLS makes it difficult to annotate sentences whose content is discussed in, or distributed over, multiple sections in the parallel abstract. In particular, we find that it can be difficult to choose between the *Main Results* and *Authors' conclusions* labels when the corresponding abstract sections both discuss results that are discussed only once in the PLS. This finding is reflected in the annotation results: although inter-annotator agreement is high, the majority of disagreements between our annotators involves these two labels. We should thus be careful in interpreting our manual annotations as ground truths.

Ultimately, we divide each PLS in our test set into sections by grouping sentences that our main annotator labeled with the same header. The purple bars in Figure 3 represent the average number of sentences per PLS section, according to our annotations. They reveal that in practice, plain language summaries cover mainly the background, results and conclusions of a Cochrane review. Furthermore, while PLS are generally shorter than abstracts, Figure 3 shows that the length difference varies per section; it even shows that the *Background* sections in our annotated PLS on average span more sentences than the *Background* sections in the parallel abstracts. This makes sense intuitively, as PLS of Cochrane reviews should be understandable for people who do not have enough background knowledge to read the technical abstract.

6 Method

So far, we have divided 22 PLS into sections through manual annotation. While this is great for analysis and evaluation purposes, we cannot do the same for all 7.7K PLS in our dataset. Therefore, in this section, we outline a two-step approach towards automatically sectioning these PLS. We also describe how we evaluate our method.

6.1 Step 1: Alignment

Recall that our objective is to divide every PLS into the best possible references for section-level simplification of the parallel abstract. As discussed in Section 4, this implies that we should try to label each sentence in a PLS with the header of the section in the parallel abstract that covers the same content. We implement this idea in the first step of our method. That is, we align semantically similar sentences between abstract-PLS pairs and subsequently label each aligned PLS sentence with the header of the abstract section (that contains the sentence) to which it is aligned.

6.1.1 Alignment model

To generate our sentence alignments, we use the neural CRF model proposed by Jiang et al. (2020). More specifically, we leverage the neural CRF model that they fine-tuned on complex-simple article pairs from Wikipedia and that Joseph et al. (2023) further fine-tuned to create the MULTI-COCHRANE corpus. They trained and evaluated this model on MC-CLEAN: 101 manually aligned pairs of abstracts (complex texts) and PLS (simple texts) of Cochrane reviews. Thus, the alignment model that we use was trained in-domain and has already been shown to surpass generalized alignment approaches on our target data.

6.1.2 Labeling

We apply the alignment model to all abstract-PLS pairs in the test set. As a result, we obtain automatic alignments between sentences in these parallel abstracts and PLS that are semantically similar. Rather than directly training simplification models on such complex-simple sentence pairs, we aim to encapsulate each aligned sentence pair within a section pair. Therefore, we label each aligned PLS sentence with the header of the abstract section that contains the sentence to which it is aligned. Since the precision of the alignment model is relatively high, the labels that we assign will be mostly accurate. In fact, the accuracy of our labels will be even higher than the precision of the alignment model, because we could align a PLS sentence to any sentence within the correct abstract section for its label to be correct.

Although we could also apply the alignment model to the auto set, this is exactly what Joseph et al. (2023) did when they created MC-NOISY. Therefore, we simply base our labels for the auto set on the automatic alignments that they shared, in the same way as before. Similarly, we do not generate automatic alignments for the train and validation splits, because we have access to the manual alignments in MC-CLEAN. However, PLS sentences in MC-CLEAN may be aligned to multiple sentences from different sections. In these cases, we make an exception and allow these sentences to be labeled with multiple section headers. In the end, this means that the sentence will be copied into multiple PLS sections.

6.2 Step 2: Classification

The second step is to classify all unaligned PLS sentences into sections. For this purpose, we treat each PLS as a sequence of sentences whose label we aim to predict. In the first step, we already labeled aligned sentences based on manual and automatic alignments. In our second step, we treat these labels as ground truths. That is, we train a classifier to simultaneously predict the section header labels of all sentences in a PLS, and we calculate its loss based on the labels that we obtained in the first step. After training, we perform inference on every PLS in our dataset and assign all unaligned sentences the label predicted by our best classifier.

6.2.1 Classification models

We tackle the sequence labeling task by first encoding each sentence in a PLS with a Sentence Transformer (Reimers and Gurevych, 2019). Then we provide our sequence of 384-dimensional sentence embeddings to a one-layer BiLSTM (Huang et al., 2015) that transforms it into $2 \cdot 48$ hidden states per sentence. Finally, these hidden states are passed to a linear layer, which outputs one of seven classes. We train the weights of our BiLSTM and linear layer, while keeping the sentence embeddings frozen.

More precisely, we explore two options that trade off accuracy and scale. The first option is to use the manual alignment labels; these are the section header labels that we derived from the alignments in MC-CLEAN. In this case, we train our classifier on the train set (74 PLS) and validate it on the validation set (5 PLS). The second option is to use the automatic alignment labels that we derived from MC-NOISY as well; in this case, we train our classifier on the auto set (7,602 PLS) and validate it on the combined train and validation sets. In both cases, we also explore the option of using just a linear layer that maps each sentence embedding to a class label directly. Training details are specified in Appendix B.

6.3 Evaluation

We evaluate our method by computing the accuracy and weighted F1 score of the predicted labels against our manual annotations on the test set. Compared to accuracy, the weighted F1 score is more sensitive to performance on minority classes and thus better accounts for class imbalance (the class distribution in our test set is shown by the purple bars in Figure 3). We also experiment with predicting section header labels using large language models (LLMs) in a zero-shot fashion.[6] We provide each paired abstract and PLS in our test set as input to the LLM, together with our annotation guidelines and additional output instructions (Appendix D). We then report on the performance of the highest-scoring LLM. The complete evaluation of these experiments, including an efficiency analysis, can be found in Appendix E.

[6] We conduct experiments with four open-source instruction-tuned LLMs: Gemma3-27B-it, Qwen2.5-32B-Instruct, Llama3.3-70B-Instruct and Qwen2.5-72B-Instruct.

	Correct	Total	Accuracy	F1 score
Step 1	128	162	79.0	79.2
Step 2	141	161	87.6	87.3
Overall	269	323	83.3	83.4

Table 1: Number of correct predictions, total predictions, accuracy and weighted F1 score of our method on the annotated test set.

7 Results

In this section, we present and discuss our results. We start with the results of the first and second steps of our method, continue with the overall results, and end with the creation of our final dataset.

7.1 Alignment results

In the first step, we automatically aligned each sentence in a PLS to one or zero sentences in the parallel abstract based on their semantic similarity. As a result, we obtained 162 alignments for the test set. Note that 161 out of 323 PLS sentences were left unaligned, which demonstrates that PLS are indeed no direct simplifications of the technical abstracts. We subsequently labeled each aligned PLS sentence with the header of the abstract section that contains the sentence to which it is aligned; the dark gray bars in Figure 3 show how these labels are distributed. As can be seen in the upper row of Table 1, the predicted labels match our manual annotations for 128 out of 162 aligned sentences.

7.2 Classification results

Table 2 shows the accuracy and weighted F1 score of our section classifiers (and an LLM) when evaluating their predictions against all our annotations on the test set. It can be seen that the use of a BiLSTM significantly increases performance. This indicates that the answer to the question of which section a sentence should belong to depends not only on the meaning of that sentence, but also on its context and its location within the PLS. Training on the automatic alignment labels rather than the manual alignment labels also significantly improves performance. Thus, scale is more important in the scale-precision trade-off. This is probably due to the fact that the precision of the automatic alignment model is still relatively high. Based on these results, we select the BiLSTM + linear layer model trained on the automatic alignment labels for the second step of our final method.

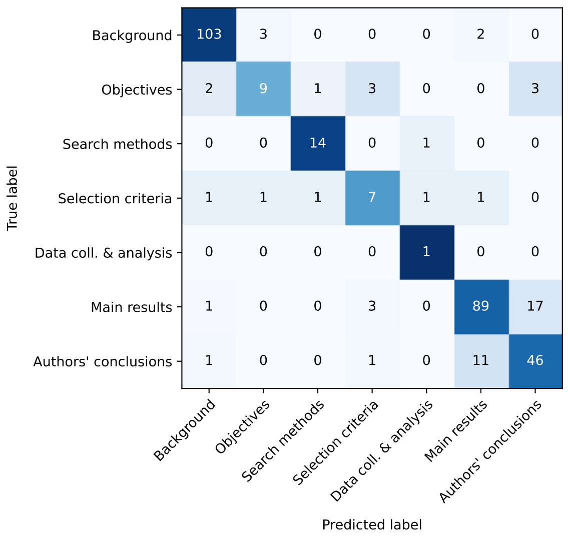
Model	Train set	Acc.	F1
Linear layer	Manual	72.1	71.3
Linear layer	Auto	78.0	77.1
BiLSTM + lin. layer	Manual	79.3	77.3
BiLSTM + lin. layer	Auto	85.4	85.1
LLM	-	84.2	83.8

Table 2: Section classification performance on the annotated test set.

7.3 Overall results

Table 1 summarizes the results of our final method on the test set. In the first step, we aligned 162 PLS sentences to an abstract section with an accuracy of 79.0%. In the second step, we classified the 161 remaining sentences into sections with an accuracy of 87.6%. This resulted in a total accuracy of 83.3% against our manual annotations and a weighted F1 score of 83.4. Meanwhile, the best-performing LLM, namely Qwen2.5-32B-Instruct, generated section labels with an accuracy of 84.2% and an F1 score of 83.8 based on our annotation guidelines. Hence, our two-step method performs competitively while it is significantly more efficient (see Appendix E). According to Table 2, performing only classification and no alignment leads to an even higher accuracy of 85.4% and a weighted F1 score of 85.1. Thus, the performance of our classifier (on the test set) is higher than that of the automatic alignment labels, even though it was trained to predict these labels (in the auto set).

At first, this result may seem counterintuitive. However, the task of the alignment model is to align each sentence in a PLS to one or zero sentences in the parallel abstract based on their semantic similarity. This is a difficult task, so it makes sense that our first step produces some incorrect alignments, which can result in incorrect labels. Meanwhile, our classifier has two advantages over the alignment model. First, its task of predicting the section header labels for each sentence in a PLS is easier, and it aligns better with our objective. Second, our classifier will learn to generalize and so it will be able to avoid some of the mistakes made by the alignment model. Our classifier also has an obvious disadvantage compared to the alignment model, namely that it does not have access to the parallel abstract. Even so, our results reveal that the classifier can outperform the automatic alignment labels without having access to the abstract.



True label \ Predicted label	Background	Objectives	Search methods	Selection criteria	Data coll. & analysis	Main results	Authors' conclusions
Background	103	3	0	0	0	2	0
Objectives	2	9	1	3	0	0	3
Search methods	0	0	14	0	1	0	0
Selection criteria	1	1	1	7	1	1	0
Data coll. & analysis	0	0	0	0	1	0	0
Main results	1	0	0	3	0	89	17
Authors' conclusions	1	0	0	1	0	11	46

Figure 4: A confusion matrix that compares our predicted and annotated labels for all PLS sentences in the test set.

Figure 4 displays a confusion matrix that compares the automatic labels produced by our method against our manual annotations on the test set. First of all, it shows that our method effectively identifies the *Background* sections in each PLS. As discussed in Section 2.1.2, this is an important contribution of our work. Second, our method may assign sentences the *Main results* label when we manually divided them into the *Authors' conclusions* section of the PLS, and vice versa. As explained in Section 5, a similar pattern can be observed when comparing the labels assigned by different annotators, because the corresponding abstract sections may both discuss results that are only discussed once in the PLS. Third, we observe that our method is least effective in identifying the *Objectives* and *Selection criteria* sections in a PLS. This may be because the sentences in the corresponding abstract sections do not always contain a subject and a verb, making them harder to align.

The dark gray bars in Figure 3 represent PLS sentences that were aligned to each section, and the light gray bars represent unaligned PLS sentences that were classified into these sections. They show that the ratio of classified versus aligned sentences is especially high for the *Background* section, due to the presence of additional background information in the PLS. They also show that the average numbers of automatically labeled and manually annotated sentences per PLS section are comparable. Furthermore, our annotations for the illustrative PLS in Figure 2 turn out to be precisely the same labels that we predict using our automatic method. In general, all of our results reveal that our method works relatively well.

7.4 COCHRANE-SECTIONS

Ultimately, we apply our method not just to the test set but to all 7.7K abstract-PLS pairs in our dataset. That is, we automatically (auto split) or semiautomatically (train and validation splits) divide each PLS into sections. We pair these sections with the parallel abstract sections and compile these pairs in COCHRANE-SECTIONS. By making our dataset available under a GPL-3.0 license (see Appendix A), we enable the training of section-level simplification models on these data. We recommend using the auto split for training and using the combined train and validation splits for validation. Moreover, we include our manually sectioned PLS as references in the test set. Not only does this enable researchers to evaluate their section-level text simplification models on a clean test set: it also enables them to evaluate better methods for the automatic sectioning of PLS against our manual annotations.

8 Conclusion

In conclusion, we devised a two-step method to automatically divide plain language summaries of Cochrane reviews into the same sections in which abstracts are divided. In the first step, we aligned each sentence in a PLS to a section in the parallel abstract if they covered similar content. In the second step, we classified the remaining sentences into sections based on what we learned from the first step. Moreover, we manually divided 22 PLS into sections and found that our automatic method divides approximately 83% of sentences into the same sections. Ultimately, we applied our method to all PLS to create the COCHRANE-SECTIONS dataset, which consists of paired abstract and PLS sections for a total of 7.7K Cochrane reviews. In future work, we will train simplification models on our dataset and evaluate their capacity to make biomedical texts more accessible.

9 Limitations

The focus of this work is on the simplification of biomedical abstracts in English. However, the abstracts and PLS of Cochrane reviews are often translated into multiple languages, and there exists a natural 1-1 sentence alignment between English and translated versions. Joseph et al. (2023) exploited this alignment to create the MULTICOCHRANE corpus; the same could be done to construct a multilingual version of COCHRANE-SECTIONS.

Although our work enables section-level simplification, the division of plain language summaries into the same sections in which abstracts are divided is not a natural one. Some PLS sentences cover content that is discussed in, or distributed over, multiple abstract sections. The division of PLS into sections also affects coherence, as sentences might refer to sentences from another section. Because PLS are written directly from the full review, they may contain information that cannot be aligned to the abstract. Besides, while we align similar sentences, the real alignments between abstracts and PLS may not reside at the sentence level. All of these limitations are reflected in the disagreements between the annotators; an example of such disagreements can be found in Appendix F. Future work could explore a multi-label setting, whereby one PLS sentence can automatically be assigned to multiple sections, as well as the division of PLS sentences into smaller units of information, to alleviate some of these limitations.

Our automatic method is imperfect, and we observe both alignment and classification errors; two examples are presented in Appendix C. Even so, a small qualitative analysis shows that many predicted labels that do not match the annotated label may still be justifiable rather than errorneous, in the same way that the disagreements between annotators can be justified based on the limitations mentioned above. In our interpretation of the results, we therefore focus on significant differences in accuracy and efficiency.

Although the section structure used is relatively generic, systematic reviews are to some extent structured differently from other studies. This can limit the generalization capacities of simplification models trained to simplify their abstracts. Nevertheless, we are the first to enable the training of simplification models on just the *Background* section, and this section is the least specific to the structure of a systematic review. Like all generative models, simplification models trained on our dataset may suffer from hallucinations, and so their outputs should not be used without manual inspection.

Acknowledgments

We thank René Spijker (Cochrane Netherlands, Amsterdam UMC) and Giovanni Cina (Amsterdam UMC) for their help. Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Jan Bakker is funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Statement on the use of Generative AI

During the preparation of this work, the authors used Writeful to check spelling and grammar. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings* *of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Miranda Cumpston, Toby Lasserson, Ella Flemyng, and Matthew Page. 2023. Chapter III: Reporting the review. In Julian Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew Page, and Vivian Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions version 6.5*. Cochrane.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 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*, pages 4972–4984, Online. Association for Computational Linguistics.

Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Élise Mathurin, and Patrice Bellot. 2022. Overview of the CLEF 2022 simpletext lab: Automatic simplification of scientific texts. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 470–494. Springer.

Dengzhao Fang, Jipeng Qiang, Yi Zhu, Yunhao Yuan, Wei Li, and Yan Liu. 2025. Progressive document-level text simplification via large language models. *Preprint*, arXiv:2501.03857.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:160–168.

Louise Hartley, Nadine Flowers, Myeong Soo Lee, Edzard Ernst, and Karen Rees. 2014. Tai chi for primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews*, (4).

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv*, abs/1508.01991.

Antonia Jelicic Kadic, Mahir Fidahic, Milan Vujcic, Frano Saric, Ivana Propadalo, Ivana Marelja, Svjetlana Dosenovic, and Livia Puljak. 2016. Cochrane plain language summaries are highly heterogeneous with low adherence to the standards. *BMC Medical Research Methodology*, 16(1):61.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

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

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(2):S5.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.

Catherine McIlwain, Nancy Santesso, Silvana Simi, Maryann Napoli, Toby Lasserson, Emma Welsh, Rachel Churchill, Tamara Rader, Jackie Chandler, David Tovey, Lorne Becker, Gill Gyte, and Annelise Synnot. 2013. Standards for the reporting of plain language summaries in new Cochrane intervention reviews (PLEACS). Cochrane.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M

Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Nicole Pitcher, Denise Mitchell, and Carolyn Hughes. 2022. Guidance for writing a cochrane plain language summary. In Julian Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew Page, and Vivian Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions version 6.5*. Cochrane.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Annika Saarto, Jos H Verbeek, Marie-Claude Lavoie, and Manisha Pahwa. 2011. Blunt versus sharp suture needles for preventing percutaneous exposure incidents in surgical staff. *Cochrane Database of Systematic Reviews*, (11).

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A Data, code and models

We share all our data, code, and pretrained models on GitHub: `https://github.com/JanB100/cochrane-sections`

B Training details for classification models

We encode sentences using the all-MiniLM-L12-v2 Sentence Transformer[7] and use the standard BiLSTM implementation from PyTorch.[8] We tune the learning rates of our classification models in {0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0} based on the validation loss. The best learning rate per model is shown in Table 3. We use a batch size of 1 and implement early stopping after five epochs. All models are trained on one NVIDIA A100 GPU; training times range from 1 to 30 minutes.

Model	Train set	Learning rate
Linear layer	Manual	2.0
Linear layer	Auto	0.2
BiLSTM + lin. layer	Manual	1.0
BiLSTM + lin. layer	Auto	0.5

Table 3: Tuned learning rate per classifier.

C Error analysis

Table 4 presents two examples of aligned PLS sentences from the test set for which either the alignment label or the classification label is correct, while the other label is incorrect. The first sentence describes the need for future trials on the review subject, so we annotated it with the *Author's conclusions* label. Based on the meaning, context, and location of that sentence within the PLS, our classifier does the same. However, the alignment model finds a sentence in the *Selection criteria* section of the abstract stating that the authors looked for this kind of trials. It aligns these sentences despite the fact that they have a different meaning. For the second sentence, our classifier is unable to correctly determine whether the sentence should belong to the results or conclusions section. We manually annotated it with the *Author's conclusions* label, because the exact same sentence is present in the conclusions section of the parallel abstract. The alignment model aligns these sentences, leading to a correct alignment label.

[7] `https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2`

[8] `https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html`

PLS sentence	Alignment label	Classification label
We need more large multicentre randomised controlled trials of commonly-used psychological therapies in older adolescents and adults with anorexia nervosa.	*Selection criteria*	*Authors' conclusions*
Postoperative complications such as infection and fever, as well as cost and time to work resumption were less in the aspiration and sclerotherapy group; however the recurrence rate was higher.	*Authors' conclusions*	*Main results*

Table 4: Examples of aligned PLS sentences from the test set for which either the alignment label or the classification label is correct. The annotated label is *Authors' conclusions* for both sentences.

D Annotation guidelines

Task description You are given the abstract and plain language summary (PLS) of a systematic review in the biomedical domain. Your task is to divide the PLS into the same sections in which the abstract is divided. You will label every sentence in the PLS with exactly one of seven section headers, so that all sentences with the same label together, in their original order, form the corresponding section. The objective is to perform this task in such a way that the resulting PLS sections constitute the best possible references for the section-level simplification of the abstract.

Annotation guidelines To determine the label of any sentence in a PLS, ask yourself: Is there a section in the abstract that covers the same content?

A. Yes, there is exactly one such section. –> Label the sentence with the header of that section.

B. Yes, there are multiple such sections. –> Select the header of the abstract section that covers the largest part of the content of the given sentence. If there are multiple sections that cover an equally large part of its content, select the label of the section that does so in the most similar terms.

C. No, there are no such sections. –> Determine which of the section descriptions below best fits the sentence, and label it with the corresponding header. If the sentence fits multiple descriptions equally well, select from the remaining options the label that leads to the most coherent PLS sections. In particular, if the given sentence refers to another sentence that you label / labeled with one of these options, you should assign the given sentence the same label.

Section descriptions See Section 3.1.

D.1 Additional LLM instructions

Output instructions Return a json dictionary that contains the indices of the PLS sentences as keys and the section header labels that you assign to these sentences as values. Do not provide any explanation.

E Comparison with LLMs

We experiment with using open-source instruction-tuned LLMs in a zero-shot fashion. Thereby, we provide abstract-PLS pairs as input to the LLM, together with our annotation guidelines and additional output instructions. We measure performance on the annotated test set and compute the efficiency of all models on the combined train and test sets, which together span 96 abstract-PLS pairs. Experiments are conducted on one to four H100 GPUs to ensure that even the longest inputs can be processed by the largest language model. We use a batch size of 8 for LLMs with 70B to 72B parameters and a batch size of 32 for those with 27B and 32B parameters. Unfortunately, we cannot leverage smaller language models since they are incapable of following complicated instructions.

E.1 Performance

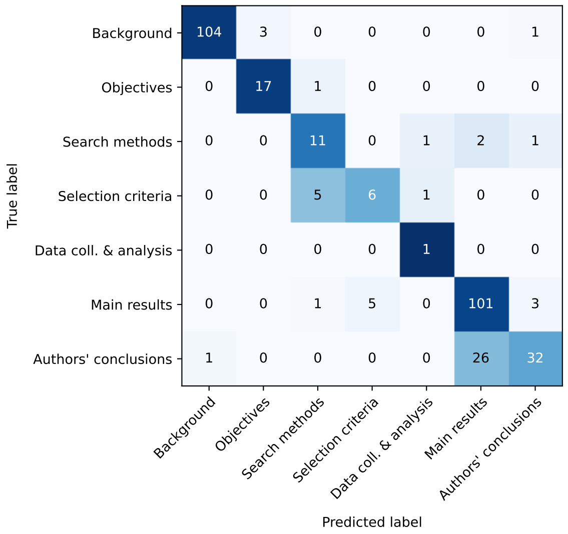
Table 5 presents the accuracy and weighted F1 score for each LLM when evaluating their predictions against our manual annotations on the test set. Performance is comparable across LLMs and is also comparable to that of our two-step method.

Model	Accuracy	F1 score
Gemma3-27B-it	81.7	82.3
Qwen2.5-32B-Instruct	84.2	83.8
Llama3.3-70B-Instruct	83.6	83.1
Qwen2.5-72B-Instruct	82.4	82.3

Table 5: Section classification performance of open-source LLMs on the annotated test set.

Model	Execution time (s)	×	#GPUs	=	GPU Time (s)
Aligner	0.80		1		0.80
Classifier	0.01		1		0.01
Gemma3-27B-it	1.84		2		3.68
Qwen2.5-32B-Instruct	2.55		2		5.11
Llama3.3-70B-Instruct	5.28		4		10.57
Qwen2.5-72B-Instruct	5.69		4		11.38

Table 6: Average GPU time per input.



	Background	Objectives	Search methods	Selection criteria	Data coll. & analysis	Main results	Authors' conclusions
Background	104	3	0	0	0	0	1
Objectives	0	17	1	0	0	0	0
Search methods	0	0	11	0	1	2	1
Selection criteria	0	0	5	6	1	0	0
Data coll. & analysis	0	0	0	0	1	0	0
Main results	0	0	1	5	0	101	3
Authors' conclusions	1	0	0	0	0	26	32

Figure 5: A confusion matrix that compares our annotated labels to those generated by Qwen2.5-32B-Instruct for all PLS sentences in the test set.

Figure 5 displays a confusion matrix that compares the section header labels generated by Qwen2.5-32B-Instruct to our manual annotations. Although the distributions of the generated and annotated labels differ to some extent, these differences mainly involve related sections, with the LLM frequently assigning sentences to the *Search methods* and *Main results* sections when we manually assigned them to the *Selection criteria* and *Authors' conclusions*, respectively, but not the other way around. Again, we find that label differences may be justifiable. Compared to our two-step method, the LLM more accurately identifies PLS sentences that describe the review objectives.

E.2 Efficiency

Table 6 shows the efficiency of our models, based on the number of H100 GPUs needed and the execution time per input. It can be observed that large language models have a high computational footprint: the average GPU time per abstract-PLS pair when using an LLM ranges from 3.68 to 11.38 seconds. This is already significantly longer than the average GPU time of our two-step method.

Still, the GPU time measured for the aligner is relatively high, namely 0.80 seconds per input. This is because Jiang et al. (2020) implemented their alignment model in such a way that it can process only one text pair at a time. Even so, the BERT-based aligner model is small enough (220M parameters) that it could be made far more efficient by parallelizing the code. In this work, we leveraged the alignment labels generated by Joseph et al. (2023), so that in practice we only had to apply our classifier to the full dataset. This model even smaller (34M parameters) and highly efficient, with an average GPU time of 0.01 seconds per PLS.

F Example with disagreements

Figure 6 shows one last example of an abstract-PLS pair from the test set. The colors indicate the labels that the main annotator assigned to each sentence in the PLS. Note that the PLS sentence which we (correctly) assigned to the *Objectives* section still discusses content that is distributed over multiple abstract sections. As mentioned above, this is one of the limitations in the way we divide PLS into reference sections. Furthermore, in the given example, the main author divided three sentences into the *Main results* section and three sentences into the *Authors' conclusions*. In doing so, they aimed to create the best reference for the simplification of the conclusions section in the abstract. However, the second author annotated all five sentences under *Key results* with the *Main results* label and labeled only the last sentence with *Authors' conclusions*. Therewith, they arguably created a better reference for the simplification of the results section. Indeed, both divisions are justifiable.

Technical abstract

Background

Stress and a sedentary lifestyle are major determinants of cardiovascular disease (CVD). As tai chi involves exercise and can help in stress reduction, it may be effective in the primary prevention of CVD.

Objectives

To determine the effectiveness of tai chi for the primary prevention of CVD.

Search methods

We searched the following electronic databases: the Cochrane Central Register of Controlled Trials (CENTRAL) (Issue 11, 2013); MEDLINE (Ovid) (1946 to November week 3, 2013); EMBASE Classic + EMBASE (Ovid) (1947 to 6 December 2013); Web of Science (Thomson Reuters) (1970 to 6 December 2013); PsycINFO (Ovid) (1806 to December week 1, 2013); Database of Abstracts of Reviews of Effects (DARE); Health Technology Assessment Database and Health Economics Evaluations Database (Issue 4, 2013). We also searched the Allied and complementary Medicine Database (AMED) and OpenGrey (inception to October 2012) and several Asian databases. We searched trial registers and reference lists of reviews for further studies. We applied no language restrictions.

Selection criteria

Randomised controlled trials of tai chi lasting at least three months involving healthy adults or adults at high risk of CVD. The comparison group was no intervention or minimal intervention. The outcomes of interest were CVD clinical events and CVD risk factors. We excluded trials involving multifactorial lifestyle interventions or focusing on weight loss to avoid confounding.

Data collection and analysis

Two review authors independently selected trials for inclusion, abstracted the data and assessed the risk of bias.

Main results

We identified 13 small trials (1520 participants randomised) and three ongoing trials. All studies had at least one domain with unclear risk of bias, and some studies were at high risk of bias for allocation concealment (one study) and selective reporting (two studies). Duration and style of tai chi differed between trials. Seven studies recruited 903 healthy participants, the other studies recruited people with borderline hypertension or hypertension, elderly people at high risk of falling, and people with hypertension with liver and kidney yin deficiency syndrome.

No studies reported on cardiovascular mortality, all-cause mortality or non-fatal events as most studies were short term (all studies had follow-up of one year or less). There was also considerable heterogeneity between studies, which meant that it was not possible to combine studies statistically for cardiovascular risk (I2 statistic for systolic blood pressure (SBP) was 96%, for diastolic blood pressure (DBP) 96%, for total cholesterol 96%, low-density lipoprotein-cholesterol (LDL-C) 95%, high-density lipoprotein-cholesterol (HDL-C) 98%, triglycerides 75%). Nine trials measured blood pressure, six individual trials found reductions in SBP (reductions ranged from -22.0 mmHg (95% confidence interval (CI) -26.3 to -17.7) to -11.5 mmHg (95% CI -21.5 to -1.46)), two trials found no clear evidence of a difference (however, CIs were wide and an increase or decrease in SBP cannot be ruled out), and one trial found an increase in SBP with tai chi (increase 5.2 mmHg, 95% CI 3.73 to 6.67). A similar pattern was seen for DBP: three trials found a reduction in DBP (reductions ranged from -12.2 mmHg (95% CI -15.8 to -8.7) to -4.43 mmHg (95% CI -7.14 to -1.72)) and three trials found no clear evidence of a difference, however again with wide CIs. Three trials reported lipid levels and two found reductions in total cholesterol, LDL-C and triglycerides (total cholesterol reductions ranged from -1.30 mmol/L (95% CI -1.57 to -1.03) to -0.50 mmol/L (95% CI -0.74 to -0.26): LDL-C reductions ranged from -0.76 mmol/L (95% CI -0.93 to -0.59) to -0.59 mmol/L (95% CI -0.80 to -0.38): triglyceride reductions ranged from -0.46 mmol/L (95% CI -0.62 to -0.30) to -0.37 mmol/L (95% CI -0.67 to-0.07)) and increased HDL-C with the intervention (HDL-C increases ranged from 0.61 mmol/L (95% CI 0.51 to 0.71) to 0.16 mmol/L (95% CI 0.02 to 0.30)), while the third study found no clear evidence of a difference between groups on lipid levels. Quality of life was measured in one trial: tai chi improved quality of life at three months. None of the included trials reported on adverse events, costs or occurrence of type 2 diabetes.

Authors' conclusions

There are currently no long-term trials examining tai chi for the primary prevention of CVD. Due to the limited evidence available currently no conclusions can be drawn as to the effectiveness of tai chi on CVD risk factors. There was some suggestion of beneficial effects of tai chi on CVD risk factors but this was not consistent across all studies. There was considerable heterogeneity between the studies included in this review and studies were small and at some risk of bias. Results of the ongoing trials will add to the evidence base but additional longer-term, high-quality trials are needed.

Plain language summary

Background

Cardiovascular diseases (CVD) are a group of conditions that affect the heart and blood vessels that are a worldwide health burden. However, it is thought that CVD risk can be lowered by changing a number of modifiable behaviours including increasing levels of exercise, and relaxation to reduce stress levels, and both of these comprise tai chi. This review assessed the effectiveness of tai chi interventions for healthy adults and adults at high risk of CVD at reducing cardiovascular death, all-cause death, non-fatal endpoints (such as heart attacks, strokes and angina) and CVD risk factors.

Study characteristics

We searched scientific databases for randomised controlled trials (clinical trials where people are allocated at random to one of two or more treatments) looking at the effects of tai chi on adults at high risk of developing CVD. We did not included people who had already had CVD (e.g. heart attacks and strokes). The evidence is current to December 2013.

Key results

We found 13 trials, none of them were large enough or of long enough duration to examine the effects of tai chi on reducing cardiovascular deaths or non-fatal endpoints. There were variations in the duration and style of tai chi and the follow-up of the interventions ranged from three to 12 months. Due to the small number of short-term studies and the variability between them, we were unable to determine conclusively whether or not tai chi was beneficial at reducing cardiovascular risk in healthy adults and adults at increased risk of CVD, although beneficial effects for CVD risk factors were seen in some studies. None of the included studies reported on adverse events. Longer-term, high-quality trials are needed in order to determine the effectiveness of tai chi for CVD prevention.

Quality of the evidence

The results of this review should be treated with caution as the studies were small, of short duration and there was some risk of bias (where there was a risk of arriving at the wrong conclusions because of favouritism by the participants or researchers).

Figure 6: Another abstract in the test set, along with the corresponding PLS (Hartley et al., 2014). The colors indicate the labels that the main annotator assigned to each sentence in the PLS.