M3: Multi-level dataset for Multi-document summarization of Medical studies

Yulia Otmakhova[1], Karin Verspoor[2,1], Timothy Baldwin[1,3],
Antonio Jimeno Yepes[2], Jey Han Lau[1]
[1]The University of Melbourne, [2]RMIT University, [3]MBZUAI
yotmakhova@student.unimelb.edu.au, karin.verspoor@rmit.edu.au, tb@ldwin.net,
antonio.jose.jimeno.yepes@rmit.edu.au, jeyhan.lau@gmail.com

Abstract

We present M3 (Multi-level dataset for Multi-document summarisation of Medical studies), a benchmark dataset for evaluating the quality of summarisation systems in the biomedical domain. The dataset contains sets of multiple input documents and target summaries of three levels of complexity: documents, sentences, and propositions. The dataset also includes several levels of annotation, including biomedical entities, direction, and strength of relations between them, and the discourse relationships between the input documents ("contradiction" or "agreement"). We showcase usage scenarios of the dataset by testing 10 generic and domain-specific summarisation models in a zero-shot setting, and introduce a probing task based on counterfactuals to test if models are aware of the direction and strength of the conclusions generated from input studies.

1 Introduction

Increased interest in summarisation has led to the development of many datasets and models, including in the biomedical domain, where summaries are required to support clinical decision making. Nevertheless, we still lack deep understanding of what these models should be capable of to support user-relevant summarisation. The approach to summarisation is largely based on the newswire domain, extracting the most salient and diverse details about particular entities, and recombining them. Yet multi-document summarisation in scientific domains poses unique challenges. For example, when summarising a set of clinical trials about drugs of a particular class, the general efficacy of that class and the certainty of that evidence is relevant, while a text full of potentially conflicting details about individual drugs is less helpful (Elhadad and McKeown, 2001). This requires the model to aggregate entities, decide on the relationships between them, and determine the strength of conclusions, which remains a distant goal (DeYoung et al., 2021).

Existing summarisation systems are not up to such complex tasks, yet limited tools exist to determine where and why they are failing. The M3 (Multi-level dataset for Multi-document summarisation of Medical studies) dataset aims to enable this, as follows: (1) We include inputs and targets of different levels of granularity such as documents, sentences and propositions, to help understand if a model is attending to correct text spans. (2) We subdivide multi-aspect evidence sentences into propositions (aspects) to measure the difficulty of aggregation across several sets of entities and relations. (3) We annotate important biomedical entities to enable their spans to guide summarisation, and to facilitate evaluation of generated summaries based on inclusion of correct entities. (4) We annotate conclusions for their direction (polarity) and strength (modality) to evaluate if a model is able to correctly aggregate relationships between entities and weight primary evidence, and also provide their spans to assist linguistic exploration of these aspects. (5) We label discourse relations between source documents as agreement or contradiction to see how it affects findings. (6) Lastly, we provide a set of counterfactual target summaries which are different from the correct ones in polarity or modality, to test the robustness of models.[1] We use the proposed dataset to conduct an empirical comparison of 10 state-of-the-art summarisation models and analyse their errors.

2 Related works

Current datasets for biomedical multi-document summarisation (MDS) can broadly be divided into two groups: large datasets based on abstracts of systematic reviews and their underlying primary studies (Wallace et al., 2021; DeYoung et al., 2021), and smaller datasets based on less complex inputs,

---

[1]The dataset and the scripts for working with it are provided at https://github.com/julia-nixie/m3.

targeting expert-written summaries aimed at a general audience (Mollá et al., 2016; Shah et al., 2021).

While systematic reviews are natural targets for summarisation of biomedical evidence and provide sufficient data to train neural summarisation models, they can be problematic to use. Source texts are long, complex scientific documents containing significant extraneous information. For example, DeYoung et al. (2021) use all clinical trials cited by a review as inputs for summarisation, though many citations are for related studies or background information rather than for primary studies directly relevant to the review scope. Though Wallace et al. (2021) avoid this by using only primary studies as inputs, the problem is not completely solved. As we show below, the conclusions of systematic reviews often cannot be directly summarised from abstracts of underlying studies. The abstracts of primary studies may contain details not directly relevant to the conclusions of systematic reviews, such as details on methods, that need to be ignored.

In contrast, the inputs of smaller-scale MDS datasets such as Mollá et al. (2016) and Shah et al. (2021) are shorter and more narrowly scoped, making them more appropriate sources for summarisation. However, as summaries are targeted at a more general audience and refer to evidence hand-picked by experts rather than a large set of primary studies, they are less relevant for the practical purpose of summarising and updating evidence for medical professionals.

Our dataset M3 tries to combine the advantages of both approaches: though it is based on systematic reviews and primary studies, we include only relevant studies as input and extract highly informative pieces of evidence from both sources and targets for summarisation. It is also not limited to Cochrane reviews like the data of Wallace et al. (2021), making it more varied in style.

3 Dataset structure

The benchmark dataset we present is based on biomedical literature for major eye diseases such as glaucoma and macular degeneration. We chose to focus the dataset around a particular topic to improve reliability by making it easier for the annotators and evaluators to understand the subject matter. This approach also helps to verify if models, which are usually trained on multi-topic datasets, can perform well in a specific domain with more limited vocabulary, testing whether they can learn to per-

	Glaucoma	AMD	DME	Total
Retrieved	515	507	107	1129
Included	203	177	71	451

Table 1: Number of retrieved and included reviews

form summarisation operations rather than rely on vocabulary distribution (Li et al., 2018). On the other hand, our corpus is very diverse in terms of types of documents, comprising diagnostics, management, prognosis, cost effectiveness, genetics, risk factors, etc. studies in addition to the treatment trials which are usually the focus of research aimed at the synthesis of biomedical evidence.

3.1 Document level

The primary aim of this dataset is to support the evaluation of multi-document summarisation systems. Hence at the top level it consists of the abstracts of systematic reviews (summarisation targets) and the abstracts of primary studies they were based on (input documents). We used the Entrez API[2] to retrieve systematic reviews from PubMed[3] for the following eye diseases: glaucoma, age-related macular degeneration (AMD), and diabetic macular oedema (DME). After retrieving the reviews, we manually checked their abstracts and excluded reviews which were not directly relevant to the query, older versions of other reviews, and reviews with full text not accessible to us.[4] The results for the retrieved and included reviews are outlined in Table 1.

The dataset includes review abstracts, and their full texts where available. The Conclusions of each abstract serves as the summarisation target, following Wallace et al. (2021). As the input documents for the summaries, we used the abstracts of primary studies included in each systematic review. Since the majority of citations in a review are not there to introduce the clinical trials underlying the review but rather to support an argument, provide background information or outline the previous research, we located the references to primary studies which were actually used to arrive at conclusions. This was done based on the full text: automatically, when there was a separate section for the included studies as in Cochrane reviews, or by manually se-

---

[2] biopython.org/docs/1.75/api/Bio.Entrez.html
[3] www.ncbi.nlm.nih.gov/pmc/
[4] Though the corpus includes abstracts of reviews, full texts were needed to find citations of included primary studies.

lecting the relevant citations. Overall, we found and retrieved 6365 related primary studies, around 14 studies per review.

3.2 Sentence level

The abstracts of systematic reviews have a very different structure and content from the abstracts of primary studies. For example, the Methods section of a systematic review describes the process of searching for and selecting primary studies, while the Methods section of a primary study explains the conditions and process of a clinical trial. Directly summarising the input primary studies with a systematic review as a target is nearly impossible. Therefore, to ensure fairness of evaluation, we select parts of both inputs and targets which contain summarisable pieces of evidence or conclusions related to the clinical question of the review.

For each of the systematic reviews, we selected one or more evidence-bearing sentences. These are primarily located in Outcomes sections but sometimes in Results. Then, we manually checked their underlying primary studies to identify sentences supporting the claims we selected. If a sentence from a systematic review did not have support in the abstracts of primary studies, we excluded it. When selecting the sentences from the primary study, we tried to find the minimal sentence-level span (usually one, but occasionally two sentences) which contains the most information to support the findings, but also is the most similar to it. For the most part, such evidence snippets were the first sentence in the Outcomes section of structured abstracts, and either the last or penultimate sentence in the unstructured ones. Sometimes the Outcomes sentences are too generic and we used a more specific sentence from the Results part.

We attempted to map the input sentences to targets based on cosine similarity of their embeddings. We tested 15 encoders on a sample of human-matched sentences, but the matching accuracy of even the best-performing one was not enough to do it automatically – PubmedBERT (Gu et al., 2020) trained on full texts, with an accuracy of 0.522. This reflects the well-known problem of surface-form divergence and variety, especially prominent in the biomedical domain, which is rife with complex synonymy, and acronyms and elisions (Plaza et al., 2011). However, there were even more challenging issues which made the mapping of sentences difficult even for human annotators. It re-

quired some additional information and reasoning operations, such as aggregation, separation, comparison, and inference. For example, if a systematic review concludes that a risk factor affects Asian but not Caucasian patients, we had to separate the underlying clinical trials into two groups according to patients' origin and map their sentences to the relevant conclusions.

In the process of mapping we had to remove a large number of items that were found to be unsuitable for summarisation. Mainly this was due to the fact that systematic reviews imply meta-analysis of results, involving aggregating and dissecting data in different ways to what is reported in primary studies. For example, a systematic review may compare two treatments, but in the primary studies they were compared with a placebo rather than each other; it can be based on the tabular or underlying data from the primary studies rather than on the text of their abstracts; it may require a numerical aggregation of data (such as number of disease cases) across studies or involve calculating correlation or other statistics for variables. We also excluded systematic reviews where the number of underlying studies was too large to be reliably mapped.

Overall, we were able to annotate 233 clusters of systematic reviews and primary studies, which consisted of 318 target evidence snippets from reviews mapped with 1840 source evidence snippets from 1258 clinical trials.

3.3 Proposition level

In the most fine-grained layer, we annotate the corpus on the level of propositions (claims). We choose this rather than elementary discourse units (EDUs: Mann and Thompson (1988)) following the practice adopted in translation, where the basic unit is a proposition (idea) rather than a span of a particular granularity. This differentiates M3 from those datasets where discourse relationships are encoded on the EDU level, such as SciDTB (Yang and Li, 2018) and BioDTB (Prasad et al., 2011). We are more interested in the interaction of ideas between documents rather than in the relations between sentence units within the document.

As M3 consists of answers to clinical questions, we define the units (propositions) based on the PICO scheme (Richardson et al., 1995). In particular, each unit has a unique combination of *Patient/population*, *Intervention*, *Comparator* (alternative treatment or placebo), and *Outcome*. If a

target sentence includes several sets of PICO elements, we annotate them separately and map only those primary studies from the sentence-level corpus that directly correspond to them. Consider, for the example, the following target sentence:

> *The pooled evidence confirmed that, compared with ranibizumab, bevacizumab was associated with equivalent effects on visual acuity at 1 year and with a higher risk of systemic serious adverse events.*

At the proposition level, we annotate two PICO tuples: {**P**: [*neovascular age-related macular degeneration*], **I**: *bevacizumab*, **C**: *ranibizumab*, **O**: *visual acuity at 1 year* } and a second with the same **P**, **I**, and **C** but with a distinct **O**: *risk of systemic serious adverse events*. Apart from clearly capturing that this sentence answers two clinical questions, it allows us to specify different relations between the PICO elements, and ensures that each proposition is mapped to directly relevant supporting evidence.

If a PICO element is not explicitly mentioned in the annotated proposition, it is filled in using the overall context of the abstract but enclosed in brackets to differentiate from explicitly mentioned spans. This ensures fairness of evaluation, as generated summaries should explicitly specify entities implied in the target summary. Comparator entities are left blank if a placebo control is used.

In addition to the entities mentioned above, we annotate the relation between the Intervention and Outcome with regards to Comparator in terms of its semantic orientation (polarity) and degree of certainty (modality), adopting the annotation scheme introduced by Lehman et al. (2019) and further refined by Otmakhova et al. (2022). In particular, we specify the polarity of the relation, that is, whether the Intervention has a *positive* effect on the outcome, a *negative* effect, or has *no effect*. If a comparator is present, the polarity signifies positive or negative difference with it rather than an absolute effect, i.e. a treatment can have a positive effect on the outcome, but be non-superior to the comparator and thus annotated as *no effect*.

Following the scheme proposed by Otmakhova et al. (2022), we also annotate the propositions in terms of their modality, or certainty of claim (*strong*, *moderate*, *weak* or *no evidence*). We assign labels for modality and mark corresponding spans.

Target propositions from the review and source sentences from primary studies are grouped together based on their shared PICO elements into

Level	Inputs	Targets	Inputs/target	Compr.
Documents	451	6365	14.1	89.7
Sentences	318	1840	5.8	8.3
Propositions	369	1899	5.1	6.2

Table 2: Dataset statistics. *Compr.* denotes the average compression rate between inputs and targets.

clusters, therefore annotating PICO only at the target level. On the other hand, as polarity and modality can be different for each of the inputs and for the target (heterogeneous results of differing certainty), we annotate them individually for each proposition. At this level, the dataset includes 369 target propositions mapped with 1899 input propositions, each explicitly annotated for polarity and modality, and grouped together based on their PICO elements.

Our fine-grained proposition mapping and annotation allows to test the capabilities of a summarisation system in a sandboxed scenario with highly relevant inputs and precise targets, as well as facilitating evaluation of the factual correctness of summarisation. In particular, the explicit annotations of PICO, modality and polarity can be used as golden labels when assessing the corresponding aspects of the generated summaries

The statistics of the resulting dataset for each of the levels described above are provided in Table 2. Appendix A contains some examples which show the structure of inputs and targets for these levels.

3.4 Reliability of annotation

To ensure the reliability of M3, we ask three external volunteers to annotate a subset of samples. In particular, they were asked to choose polarity and modality labels for 40 summaries on the proposition level, and to match conclusions of 30 summaries with evidence sentences from primary studies. The details of the annotation tasks together with the annotation instructions and examples as well as details of inter-annotator agreement are provided in Appendix B. Overall, we achieve a high agreement for the classification tasks and a substantial one for semantic similarity matching.

4 Evaluation tasks

4.1 Summarisation

We next turn to testing the performance of state-of-the-art multi-document summarisation models with M3.

4.1.1 Zero-shot performance of a domain-specific model across M3 levels

We begin with a model specifically developed for the biomedical domain (DeYoung et al., 2021). We use the BART (Lewis et al., 2020) model checkpoint[5] to generate summaries without any additional fine-tuning on M3. To accommodate for the long length of input texts, we encode inputs individually and then concatenate into a single representation, following DeYoung et al. (2021). A feature of this model is that each input is also prepended by the Background section of the summary, to provide general information and guide the generation. This essentially serves as an extended prompt. However, the model showed an overly strong reliance on this prompt and a tendency to copy from it rather than generate from inputs (Otmakhova et al., 2022). Hence we exclude it in three of our experiments and use a short prompt consisting of relevant PICO elements in the last experiment. Moreover, in contrast to their work, we do not mark PICO entities and conclusion sentences specifically. Thus, the aim of this set of experiments is to check the performance of the model without relying on prompting or decoration, and to see how it reacts to different types of inputs in such more realistic scenarios.

We generate four sets of summaries based on input texts of different levels (document, sentence, and proposition level). For the proposition level we either use no prompt or concatenate the annotated PICO elements into a single string and prepend it to each input. We randomly sample 50 generated documents for each of these sets and manually evaluate their consistency with the annotated target summary in terms of PICO elements, polarity and modality. We also compare them to the target summaries using automatic metrics ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019).

Table 3 shows the percentage of generated summaries containing the same PICO elements, polarity or modality of the findings as the target summary, as well as the percentage of generated summaries that were correct or wrong in term of these aspects. In general, the number of correct summaries increases when the model is supplied with exact snippets of text needed to generate them, showing the importance of extracting highly-relevant pieces of evidence from multiple documents. Interestingly, the performance for PICO elements goes down at the proposition level. It is

likely difficult for the model to pick out the correct set of entities if several are present in the inputs. Naïve prompting with the set of relevant PICO elements did not help resolve this problem and introduced another, where the model would copy all or some PICO elements as a single entity and use them to fill PICO slots in the generated summaries. That is, instead of focusing generation of PICO elements and relationships between them based on the prompt, the model copies the prompt literally. For example, for the prompt *Gingko biloba vision* for Intervention *Gingko biloba* and Outcome *vision* the model generated the following summary, placing the full prompt in the Intervention slot:

> *The meta- analysis showed that **Gingko biloba vision** improved after treatment with Ginkgo biloba extract.*

Worryingly, though almost all summaries generated with prompting suffer from such excessive copying of biomedical entities and are thus meaningless, their ROUGE scores show substantial improvements due to multiple lexical overlaps.

4.1.2 Zero-shot performance of summarisation models at proposition level

We report performance of current summarisation models on level 3 (propositions) of M3. Comparison at this level is motivated in two ways. First, the availability of "oracle" evidence statements as inputs to the summarisarion model allows us to dissect extractive and abstractive operations and focus on evaluating the latter. Second, the inputs are relatively short compared to whole abstracts, which is important for fair comparison as most of the current summarisation models still have limitations in terms of number of input tokens.

As our task here is to showcase M3 by comparing models and highlighting their unique behaviour on it rather than select or propose the one with the best performance, we either use publicly available fine-tuned checkpoints or fine-tune pre-trained models using similar hyperparameters to those suggested by their authors. That is, we do not focus on hyperparameter search and our results might not reflect the best-achievable performance for a given model. We chose a variety of models to compare, including ones fine-tuned on both single-document and multi-document datasets from both biomedical and generic domains, including a

---

[5] github.com/allenai/ms2/

	PICO	Polarity	Modality	All correct	All wrong	R-1	R-2	R-L	BertScore
Documents	12%	44%	48%	2%	18%	24.6	5.7	18.0	85.3
Sentences	**36%**	54%	56%	10%	12%	24.1	5.9	18.8	85.7
Propositions	26%	58%	58%	**14%**	**10%**	23.7	5.8	18.6	85.6
Prompted propositions	6%	**62%**	**64%**	0%	**10%**	38.6	19.2	30.4	87.1

Table 3: Results of manual factuality evaluation vs. automatic evaluation for the in-domain model

a multi-document corpus of biomedical studies we refer to as Cochrane (Wallace et al., 2021):

- BART XSUM, a generic BART-large model (Lewis et al., 2020) fine-tuned on a single-document newswire dataset.[6]
- BART Cochrane, a generic BART-large model fine-tuned on Cochrane.
- BioBART Cochrane, a biomedical BART-large model (Yuan et al., 2022) fine-tuned on Cochrane.
- Pegasus XSUM, a Pegasus-large model (Zhang et al., 2020) fine-tuned on a single-document newswire corpus.[7]
- Pegasus Pubmed, a Pegasus-large model fine-tuned on a single-document dataset of biomedical articles (Pubmed).[8]
- BigBird Pegasus Pubmed, a BigBird Pegasus-large model (Zaheer et al., 2020) fine-tuned on the same Pubmed dataset.[9]
- Pegasus Cochrane, a Pegasus-large model fine-tuned on Cochrane.
- Primera, the first pre-trained model specifically designed for multi-document summarisation (Xiao et al., 2022).[10]
- Primera Cochrane, a Primera model fine-tuned on Cochrane.

The results of the automatic evaluation of these models in terms of ROUGE scores (Lin, 2004) and BERTScore (Zhang et al., 2019) are presented in Table 4. We note a low discriminatory ability of BERTScore for this task, with scores in a very narrow range and coinciding for several models. In terms of ROUGE scores, it seems that fine-tuning on a multi-document in-domain dataset (Cochrane) helps, as nearly all Cochrane models (except for Primera) achieve higher ROUGE scores than their generic counterparts, and that the simplest model (BART-large) is able to generalize better than the

	R-1	R-2	R-L	BScore
BART XSUM	19.3	2.6	14.4	0.86
BART Cochrane	**24.6**	**5.5**	**19.2**	**0.88**
BioBART Cochrane	21.5	3.6	16.4	0.87
Pegasus XSUM	22.1	4.0	16.6	0.87
Pegasus Pubmed	11.7	1.8	8.8	0.81
BigBird Pubmed	10.4	1.4	7.8	0.81
Pegasus Cochrane	23.8	5.1	18.2	0.88
Primera	20.2	4.2	14.0	0.86
Primera Cochrane	18.8	2.3	14.5	0.87

Table 4: The results of automatic evaluation of pre-trained models on the proposition level of the dataset. "BScore" = BertScore

others. Interestingly, Pegasus models fine-tuned on a single-document biomedical dataset (Pubmed) are the worst performers. In general, scores presented here are substantially lower than those reported for the corresponding models on their original dataset, due to the difficulty of zero-shot learning on a highly abstractive dataset. What concerns us more is what these numbers truly represent, which we analyse by performing human evaluation of the generated summaries.

We analyse the same set of 30 randomly-chosen summaries for each of the models. In addition to the factuality rubrics used in the previous section we mark if there are any *hallucinations* (content which is inconsistent with the input documents) or *repetitions* in the generated text. As many of the generated summaries do not contain any evidence or claim, we also note the number of such *non-evidential* summaries to further refine the *Modality* rubric. The results across these dimensions are presented in Table 5. In general, the results of manual evaluation are not aligned with ROUGE scores: though models trained on the Cochrane corpus had the highest ROUGE scores, they also perform the worst in terms of generating completely wrong summaries. Thus, though the models learn to reproduce biomedical vocabulary learned by fine-tuning on an in-domain corpus, it does not necessarily lead to meaningful improvements. Below we discuss some patterns and typical errors these models make; examples of wrong and correct summaries generated

by the models can be found in Appendix C and Appendix D, respectively.

Surprisingly, some generic models (Pegasus XSUM, Primera) performed very well in terms of producing correct summaries and an absence of hallucinations. As we show below, this is achieved mostly by copying from the source. On the other hand, BART XSUM tended to hallucinate unrelated biomedical entities, leading to worse performance in terms of PICO and overall correctness. Interestingly, BART XSUM also tended to add irrelevant references such as *according to a study in the Journal of Ocular Medicine*, which we did not notice in the other models.

Hallucinations, however, are especially severe in models fine-tuned on the Pubmed corpus (Pegasus Pubmed and BigBird Pegasus Pubmed), producing abstract-style summaries with external information about the entities mentioned in the input documents. We hypothesise that since the data these models was trained on is substantially longer than the inputs and targets in M3, the models are trying to reproduce the style and length of documents they are accustomed to by "padding" the input texts with similar information from the model itself and repeating the content (both models also suffer from excessive repetitions).

In terms of models fine-tuned on the multi-document biomedical corpus (Cochrane), they seem to do better in terms of determining the direction of findings (*positive*, *negative*, or *no effect*) by aggregating often conflicting inputs. However, it is difficult to make strong conclusions here, as these models actually make claims in a very small number of cases, mostly generating non-evidential summaries such as *There is not enough evidence to support or refute . . . .* It appears that when the multi-document models are certain enough to make a conclusion, they choose the direction better than models fine-tuned on single-document sources.

Though the original Primera model had the best performance across almost all categories, upon further analysis this proves to be due to excessive copying from source documents. The model mostly performed extractive summarisation by stitching together whole sentences from inputs, relating to an objective in Primera that pushes the model to identify and extract unique information from input documents rather than aggregating and recombining similar propositions. Thus, when the inputs are non-contradictory and there is no need to aggregate

biomedical entities of different levels, Primera is able to make correct conclusions by copying one or more important input sentences. However, whenever there is a need for more complex operations, copying is not enough to make a correct conclusion, and Primera fails.

Though Primera is the extreme example, many of the models copy large chunks of text (even whole sentences) from the input texts. To estimate the amount of copying and thus extractiveness in what are supposed to be abstractive pretrained models, we calculate lexical overlap of generated summaries with the source documents using ROUGE scores. We also report the difference in ROUGE scores when comparing the summaries to inputs vs targets (see Table 6). Not surprisingly, the Cochrane models tend to abstract more from the input documents and rely less on copying than the single-document models.

4.2 Probing for awareness of modality and direction

Apart from polarity and modality labels facilitating manual evaluation of factual correctness of summaries, we also provide counterfactual targets to assist automatic evaluation. Unlike usual counterfactual tasks where we check how a change in inputs affects the targets, we modify the target summaries so that their content cannot be inferred from the inputs. The underlying idea is that given the same inputs, the model should be able to assign more probability to the correct rather than counterfactual generation. We manually corrupt the target propositions in the following five ways, making sure to modify only one aspect at a time:

- **Negation**: flip the negation of the predicate (affirmative into negative and vice versa) and change other words to preserve grammaticality and coherence, such as changing *some* into *any*, or *and* into *but* (Pullum and Huddleston, 2002).
- **Antonymy**: flip the polarity of conclusions by substituting cue words with an antonym (e.g. *increases* → *decreases*, *less* → *more*).
- **No effect**: change the propositions with a specific polarity of findings (*positive* or *negative*) into a *no effect* statement, or vice versa.
- **No evidence**: change specific conclusions into non-evidential sentences (*There is no evidence ...*), and vice versa.
- **Modality**: hedge sentences with moderate

	PICO↑	Modal.↑	Polarity↑	All correct↑	All wrong↓	Halluc.↓	Repet.↓	No evid.↓
BART XSUM	13.3	46.7	61.1	6.7	36.7	83.3	0.0	43.3
BART Cochrane	43.3	20.0	**100.0**	10.0	43.3	10.0	3.3	70.0
BioBART Cochrane	46.7	30.0	66.7	20.0	43.3	16.7	**0.0**	60.0
Pegasus XSUM	33.3	66.7	50.0	20.0	26.7	20.0	**0.0**	30.0
Pegasus Pubmed	16.7	**73.3**	54.2	13.3	26.7	93.3	56.7	20.0
BigBird Pubmed	6.7	53.3	52.9	6.7	43.3	96.7	36.7	46.7
Pegasus Cochrane	26.7	10.0	25.0	0.0	60.0	40.0	**0.0**	83.3
Primera	**53.3**	**73.3**	60.0	**36.7**	**13.3**	6.7	**0.0**	**0.0**
Primera Cochrane	20.0	3.3	66.7	0.0	73.3	70.0	**0.0**	86.7

Table 5: Results of manual evaluation of pre-trained models (%); ↑: higher is better, ↓: lower is better

	R-1	R-2	R-L
BART XSUM	20.4 (+1.1)	7.1 (+4.5)	14.8 (+0.5)
BART Cochrane	25.3 (+0.7)	16.2 (+10.7)	21.6 (+2.4)
BioBART Cochrane	19.8 (−1.7)	9.4 (+5.8)	16.2 (−0.2)
Pegasus XSUM	24.1 (+2.0)	13.4 (+9.4)	19.4 (+2.8)
Pegasus Pubmed	29.8 (+18.1)	11.4 (+9.6)	19.9 (+11.2)
BigBird Pubmed	25.1 (+14.7)	7.2 (+5.8)	16.0 (+8.2)
Pegasus Cochrane	18.7 (−5.2)	6.7 (+1.7)	14.0 (−4.2)
Primera	74.5 (+54.3)	73.7 (+69.5)	73.1 (+59.1)
Primera Cochrane	14.1 (−4.8)	3.0 (+0.7)	10.3 (−4.2)

Table 6: ROUGE scores of *generated summaries* vs. input texts. Differences of the scores of targets vs. inputs are in brackets; higher differences indicate the model copies more from the source and summaries are less abstractive.

	Negation		Antonyms		No effect		No evidence		Modality	
	Δ loss	Acc	Δ loss	Acc	Δ loss	Acc	Δ loss	Acc	Δ loss	Acc
BART XSUM	0.092±0.19	0.73	0.126±0.18	0.83	0.084±0.18	0.73	−0.188±0.34	0.25	0.051±0.17	0.65
BART Cochrane	0.046±0.17	0.65	0.102±0.15	0.80	0.035±0.15	0.63	−0.306±0.29	0.09	0.021±0.13	0.56
BioBART Cochrane	0.091±0.18	0.74	0.138±0.16	0.87	0.081±0.17	0.73	−0.128±0.29	0.32	0.032±0.14	0.60
Pegasus XSUM	−0.012±0.17	0.50	0.039±0.25	0.63	−0.008±0.17	0.51	−0.377±0.61	0.23	0.002±0.20	0.50
Pegasus Pubmed	0.023±0.24	0.53	0.124±0.24	0.75	0.019±0.24	0.53	−0.714±2.00	0.41	0.042±0.31	0.56
BigBird Pubmed	0.034±0.34	0.56	0.143±0.26	0.75	0.030±0.33	0.56	−0.080±0.96	0.56	0.043±0.32	0.56
Pegasus Cochrane	0.078±0.18	0.73	0.144±0.21	0.83	0.068±0.18	0.71	−0.163±0.33	0.30	0.032±0.18	0.54
Primera	0.064±0.21	0.68	0.123±0.18	0.83	0.051±0.18	0.67	−0.194±0.36	0.25	0.047±0.18	0.61
Primera Cochrane	0.070±0.19	0.68	0.106±0.15	0.82	0.059±0.18	0.66	−0.186±0.32	0.27	0.034±0.14	0.59

Table 7: Awareness of the corruptions of the target. "Δ loss" is the difference and s.d. in mean NLL loss.

strength of claim using such words as *might*, and remove markers of weak and strong modality to moderate the claims.

We follow several strategies to ensure that the modified targets are not unnatural or too dissimilar from the originals. Specifically, we avoid changes to syntactic structure, limit changes to maximum 5 words, and coordinate changes with other parts of the sentence to preserve grammar and cohesion. The resulting dataset contains 234 summaries with inverted negation, 196 with antonyms for polarity, 234 targets with a specific polarity changed to/from *no effect*, 167 conclusions changed to/from *no evidence*, and 200 summaries with moderate modality flips. Examples of created counterfactual summaries appear in Table 8.

The counterfactual summaries are used to probe the ability of models to detect corruptions, and thus their awareness of the polarity and strength of aggregated evidence. In this pilot study, we use the negative log-likelihood loss used to train the models as a simple measure to compare the likelihood of generating correct and corrupted summaries. We calculate the difference in NLL loss between the counterfactual and correct summaries (Δ loss), and the relative number of times the model has a strictly lower loss for the correct as compared to the corrupted summary ("Acc"). Results are presented in Table 7. Intuitively, a positive Δ loss means that the model assigns more probability to the correct target than to the modified one, while a loss around zero means the model cannot differentiate them.

Target (PMID 18242575)	Heavy alcohol consumption (more than three standard drinks per day) is associated with an increased risk of early AMD.
Negation	Heavy alcohol consumption (more than three standard drinks per day) is **not** associated with an increased risk of early AMD.
Antonym)	Heavy alcohol consumption (more than three standard drinks per day) is associated with **a reduced** risk of early AMD.
No effect	Heavy alcohol consumption (more than three standard drinks per day) is **not** associated with an increased risk of early AMD.
No evidence	**There is no evidence that** heavy alcohol consumption (more than three standard drinks per day) is associated with an increased risk of early AMD.
Modality	Heavy alcohol consumption (more than three standard drinks per day) **may be** associated with an increased risk of early AMD.

Table 8: Examples of **counterfactual** summaries created by changing the target modality or polarity of the same target summary. The tokens that were modified are in **bold**.

The standard deviation reflects the stability of results across samples, indicating whether the model is biased to particular language features.

The results show that some aspects of factuality are difficult for models to recognize: the majority of models struggle to detect changes in negation, modality, or polarity of findings (no effect). In contrast, the pre-trained models are quite good at choosing correct antonyms, especially those fine-tuned on the in-domain Pubmed dataset. Interestingly, all models assigned less loss to summaries with changed evidentiality. Most summaries in M3 contain specific findings while counterfactual summaries are non-evidential, supporting the conclusion that models tend to generate sentences citing lack of evidence, and not making any claims at all.

5   Conclusions

In this paper we presented a new dataset for multi-document biomedical summarisation which is designed to be used as a benchmark for evaluation and comparison of summarisation models. To assist this task, we provide inputs and targets of different granularity, human annotations in terms of several aspects, and also sets of counterfactual target summaries. We show some ways to use M3 by evaluating 10 summarisation models, and introduce a baseline probing task for determining awareness of polarity and modality of findings. We hope that the abundance of different annotations we provide will help the community to develop new ways to test and evaluate summarisation systems in the biomedical and scientific domain.

6   Limitations

Though we did our best to retrieve and include all relevant citations for the systematic reviews, in rare cases (for older studies or papers in languages other than English) their abstracts were not available online. Moreover, as M3 is almost entirely reliant on meticulous human annotation, it requires substantial effort to construct. Its current size is suitable only for evaluation or few-shot learning rather than

training and fine-tuning of summarisation models.

7   Ethical considerations

Multi-document summarisation is of critical importance for the biomedical domain, where it can allow medical practitioners to keep up with the current evidence and thus improve patient care. However, errors in this domain are particularly dangerous as the summarised evidence may directly inform treatment of a patient. We must therefore be careful to deeply understand what we are trying to achieve. We hope that our work will help to mitigate concerns with their use in practice by shedding some light on the state of these systems.

Acknowledgements

References

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MS^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513. Association for Computational Linguistics.

N Elhadad and Kathleen McKeown. 2001. Towards generating patient specific summaries of medical articles. In *Proceedings of NAACL-2001 Automatic Summarization Workshop*, Seattle, USA.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. arXiv:2007.15779.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treat-

ments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 474–479, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text — Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Diego Mollá, María Elena Santiago-Martínez, Abeed Sarker, and Cécile Paris. 2016. A corpus for research in text processing for evidence based medicine. *Language Resources and Evaluation*, 50(4):705–727.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

Laura Plaza, Alberto Díaz, and Pablo Gervás. 2011. A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence in Medicine*, 53(1):1–14.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):1–18.

Geoffrey K Pullum and Rodney Huddleston. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13.

Darsh J Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021. Nutri-bullets: Summarizing health studies by composing segments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13780–13788.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization. In *AMIA Annual Symposium Proceedings*, volume 2021, page 605. American Medical Informatics Association.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A   Dataset examples

In this section we provide the examples of data structure on the three levels of the dataset: *Documents*, *Sentences*, and *Propositions*. As the example of the target and inputs we use the systematic

3925

review with *PMID 16856103*, and its underlying primary studies *PMID 9795850* and *12913321*, correspondingly. Please note that for simplicity we use a review where there is only one conclusion, as the result of which the text of examples on Sentence and Proposition levels coincides; however, the example for the Proposition level has additional annotations as described in section B. Some long paragraphs on the Document level were abbreviated for display purposes. The examples for the Document, Sentence and Proposition levels are shown on Figures 1 respectively.

B   Annotation details

We evaluated the reliability of annotation and sentence mapping with the help of 3 volunteer annotators. For the classification task, they were asked to choose polarity and modality labels given particular PICO elements. For the sentence mapping task, as it was unfeasible to ask annotators to compare target sentences with all input sentences (there are on average 12 sentences in each input document), we asked them to choose the closest sentence out of four candidates, one of which was the sentence chosen by the main annotator, and the other three were sentences with the highest cosine similarity to the target based on PubmedBERT CLS token representation, which we found to be most accurate in predicting the correct matching. The four resulting candidates were shuffled. Some examples of annotation for the classification task are shown in Figure 4. Figure 5 presents annotation instructions and an annotated sample for the similarity matching task.

The results of inter-annotator agreement are provided in Table 9.

	Agreement	Gwet's AC1	Fleiss' κ
Polarity	85%	0.79	0.81
Modality	87.5%	0.86	0.85
Sent. similarity	81.7%	0.79	0.76

Table 9: Inter-annotator agreement on classification (Polarity and Modality) and evidence matching tasks.

C   Examples of wrong generated summaries

In Table 10 we provide examples of summaries generated by different pre-trained models for the same set of input propositions, together with the correct *Target* summary. Only the summary generated by Primera (and essentially copied from the

inputs) is fully correct.

D   Examples of correct generated summaries

As it was impossible to find a sample where the majority of generated summaries would be correct, we provide examples for different sets of inputs in Table 11. If a summary contains some irrelevant hallucinated details, we mark them in *italics*. If the generation is correct but not perfect, we provide comments in italics. In the evaluated sample Pegasus Cochrane and Primera Cochrane did not have completely correct generated summaries.

```
{
    "target_pmid":"16856103",
    "target_abstract": "BACKGROUND: Angle-closure glaucoma is characterized by obstruction to the
outflow of aqueous humor and consequent rise in intraocular pressure. The obstruction may result
from an anatomical predisposition of the eye or may be due to pathophysiologic processes in any
part of the eye. ... AUTHORS' CONCLUSIONS: There is no evidence from good quality randomized
trials or non-randomized studies of the effectiveness of lens extraction for chronic primary
angle-closure glaucoma.",
    "target_conclusions": "AUTHORS' CONCLUSIONS: There is no evidence from good quality randomized
trials or non-randomized studies of the effectiveness of lens extraction for chronic primary
angle-closure glaucoma.",
        "source_items":[
        {
            "source_pmid":"9795850",
            "source_abstract":"PURPOSE: To evaluate the long-term effects of extraction of
incipient cataracts or clear lenses on glaucoma control in patients with subacute or chronic
angle-closure glaucoma.... CONCLUSION: Drainage surgery in patients with angle-closure glaucoma
proved to be associated with multiple surgical interventions and deterioration in visual
function. The choice of first a cataract procedure with the option of a future trabeculectomy may
be a more attractive approach in patients with subacute or chronic angle-closure glaucoma than
trabeculectomy followed by an optional cataract procedure."
        },
        {
            "source_pmid": "12913321",
            "source_abstract": "PURPOSE: We studied the surgical outcomes of phacoemulsification
and intraocular lens (IOL) implantation for cataract and/or uncontrolled intraocular pressure
(IOP) in eyes with angle closure glaucoma.... CONCLUSIONS: Phacoemulsification and IOL
implantation is useful in IOP control for angle closure glaucoma after relief of pupillary block.
However, we should take care of operative complications because of a shallow anterior chamber,
poor mydriasis and zonular weakness."
        }
    ]
}
```

Figure 1: An example of the Document level structure

```
{
    "target_pmid":"16856103",
    "target_items":[
        {
            "target_text":"There is no evidence from good quality randomized trials or non-randomized
studies of the effectiveness of lens extraction for chronic primary angle-closure glaucoma.",
            "source_items":[
                {
                    "source_pmid":"9795850",
                    "source_text":"Glaucoma control was achieved in 15 eyes (68%) in Group 1 and in 17
eyes (68%) in Group 2."
                },
                {
                    "source_pmid":"12913321",
                    "source_text":"Phacoemulsification and IOL implantation is useful in IOP control for
angle closure glaucoma after relief of pupillary block."
                }
            ]
        }
    ]
}
```

Figure 2: An example of the Sentence level structure

Target (PMID 206650441)	Combination therapy of IVB and IVT did not result in any significant reduction in CSMT or gain in vision compared to treatment with IVB alone at any point in time.
BART XSUM	The IVB and IVT groups were compared in a double-blind, placebo-controlled trial to see if the treatment of a patient with a heart attack improved their heart function.
BART Cochrane	There is no evidence that IVB/IVT treatment is associated with a significant reduction in CMT and BCVA in the short term.
BioBART Cochrane	There is no evidence to support the use of IVB in the treatment of CMT.
Pegasus XSUM	The influence of the IVB and IVT groups on the change in CMT and BCVA between the IVB and IVT groups has been investigated.
Pegasus Pubmed	in this study, we report the results of a large - scale, multicenter, prospective, randomized, controlled trial (rct) to test the hypothesis that hiv-1 infection is a risk factor for type 2 diabetes mellitus (t2 dm). <n> the rct was conducted in a community with a high prevalence of t2 dm. <n> we collected peripheral blood samples, including whole blood (wbc), whole blood glucose (wbg), serum insulin, and insulin - like growth factor-1 (igf-1) levels, and urine samples. <n> urine samples were analyzed by gas chromatography / mass spectrometry (gc / ms). <n> we found that hiv-1 infection was a risk factor for t2 dm. <n> the rct was successful in achieving its primary objective, that is, to determine whether hiv-1 infection is a risk factor for t2 dm. <n> the primary objective of the rct was to determine whether hiv-1 infection is a risk factor for t2 dm. <n> the secondary
Bigbird Pubmed	key clinical messagea 35-year - old man was found to have a cerebral abscess secondary to a recurrent sigmoid diverticulum rupture.<n> both cultures grew streptococcus anginosus.<n> brain abscess is a rare but potential complication of sigmoid diverticulum rupture.
Pegasus Cochrane	There is no evidence to suggest that the use of IVB over IVT is more effective than IVT in the provision of high-speed Internet access.
Primera	No significant differences were detected in the changes of CMT and BCVA between the IVB and IVB/IVT groups (P = 0.99).
Primera Cochrane	There is insufficient evidence to support or refute the use of any intervention for the management of acute ischaemic stroke.

Table 10: Examples of **incorrect** generated summaries (excluding Primera).

```
{
  "target_pmid":"16856103",
  "target_items":[
    {
      "target_text":"There is no evidence from good quality randomized trials or non-randomized studies of the
effectiveness of lens extraction for chronic primary angle-closure glaucoma.",
      "p":"chronic primary angle-closure glaucoma",
      "i":"lens extraction",
      "c":"",
      "o":"effectiveness",
      "target_modality":"no evidence",
      "target_modality_span":"no evidence",
      "target_polarity":"n/a",
      "target_polarity_span":"",
      "discourse_relation":"contradiction",
      "source_items":[
        {
          "source_pmid":"9795850",
          "source_text":"Glaucoma control was achieved in 15 eyes (68%) in Group 1 and in 17 eyes (68%) in Group
2.",
          "polarity":"no effect",
          "polarity_span": "15 eyes (68%) in Group 1 and in 17 eyes (68%) in Group 2",
          "modality":"moderate",
          "modality_span":""
        },
        {
          "source_pmid":"12913321",
          "source_text":"Phacoemulsification and IOL implantation is useful in IOP control for angle closure
glaucoma after relief of pupillary block.",
          "polarity":"positive",
          "polarity_span":"useful",
          "modality":"moderate",
          "modality_span":""
        }
      ]
    }
```

Figure 3: An example of the Propositions level structure

PMID	Text	P	I	C	O	Modality	Polarity
27077733	anti-VEGF injections and RNFL thickness changes when all studies were examined together.		anti-VEGF injections		RNFL thickness changes	moderate	no effect
30325017	Anti-VEGF drugs are effective at improving vision in people with DMO with three to four in every 10 people likely to experience an improvement of 3 or more lines VA at one year.	people with DMO	anti-VEGF drugs		vision	moderate	positive
24641069	hyperemia in more patients than timolol.		latanoprost	timolol	conjunctival hyperemia	moderate	negative
16151432	The literature review confirmed a strong association between current smoking and AMD, which fulfilled established causality criteria.		current smoking		AMD	strong	negative
17712074	Tight glycemic and blood pressure control reduces the incidence and progression of DR.		tight glycemic and blood pressure control		incidence and progression of DR	moderate	positive

Figure 4: Examples of annotation for the polarity and modality classification task

In this task you are asked to match the sentence on the left (*target*) with one out of four sentences on the right (*input*). The sentence on the right should be the one most semantically close to the sentence on the left, i.e. most relevant to making the conclusion expressed in *target*. For example, for the following target:

Target: Our meta-analysis results suggest that current smokers are at significantly increased risk of developing POAG.

The most closely related input can be:

1) a (possibly distant) **paraphrase**: Current cigarette smoking was also associated with glaucoma (rate ratio = 2.9; 95% CI = 1.3 to 6.6).
2) **inference**: Associations were found with intraocular pressure greater than 21 mmHg (Odds ratio (OR) = 119.4, 95% confidence interval (CI), 17.62 to 4960.64), current cigarette smoking (OR = 2.77, 95% CI, 1.04 to 7.34) and Mongo ethnic subgroup (OR = 3.195, 95% CI, 0.93 to 9.57).
3) or a **different proposition with the same set of entities**: In these prevalence data, neither heavy drinking nor cigarette-smoking behavior was related to the prevalence of open-angle glaucoma.

Please chose and mark one of the four inputs (A to D) which is most closely related to the target.

1. pmid 12038941

The result of this meta-analysis suggested a statistically significant association between H. pylori infection and OAG.	A. Serum anti-H pylori-specific IgG was analyzed by enzyme-linked immunosorbent assay.
	B. Helicobacter pylori eradication may positively influence glaucoma parameters, suggesting a possible causal link between H pylori and glaucoma.
	C. Helicobacter pylori-positive patients received a triple eradication regimen (omeprazole, clarithromycin, and amoxicillin treatment), and all patients were observed for 2 years while remaining under the same antiglaucoma therapy.
	D. To evaluate the effect of H pylori eradication on the 2 most commonly used glaucoma parameters: intraocular pressure and visual field.

Figure 5: Annotation instructions and sample for the similarity task

Target (PMID 30320466)	Laser photocoagulation reduces the chances of visual loss and increases those of partial to complete resolution of DMO compared to no intervention at one to three years.
BART XSUM	Focal photocoagulation reduces the risk of blindness in diabetic eyes, and increases the chance of a successful treatment, according to a study in the Journal of the American Academy of Ophthalmology. (*has a hallucinated part*)
Target (PMID 25034616)	However, different stages of the treatment procedure produce varying levels of patient discomfort.
BART Cochrane	There is considerable variation in how patients tolerate different stages of the injection procedure. (*copied from the source*)
Target (PMID 28139151)	Late AMD was associated with elevated rates of all-cause (nine studies, hazard ratio (HR) 1.20, 95% confidence interval, CI, 1.02–1.41) and cardiovascular mortality (six studies, HR 1.46, 95% CI 1.13–1.98), but early AMD was not (all-cause mortality, 10 studies, HR 1.06, 95% CI 0.98–1.14; cardiovascular mortality, five studies, HR 1.12, 95% CI 0.96–1.31).
BioBART Cochrane	Early AMD is not associated with an increased risk of death. (*summarized for PICO: early AMD, mortality*
Target (PMID 28607930)	Findings from this research provide support that persons with AMD had a higher subsequent risk of mortality than persons without AMD.
Pegasus XSUM	Age-related macular degeneration (AMD) is a significant predictor of all-cause mortality.
Target (PMID 25034616)	However, different stages of the treatment procedure produce varying levels of patient discomfort.
Pegasus Pubmed	there is considerable variation in how patients tolerate different stages of injection procedure. <n> this variation is of considerable variation in how patients tolerate different stages of injection procedure. <n> it can affect the outcome of the procedure. <n> it can affect the cost of the procedure. <n> it can also affect the duration of hospital stay. <n> the variation in how patients tolerate different stages of injection procedure is of considerable variation in how patients tolerate different stages of injection procedure. <n> it can affect the outcome of the procedure. <n> it can affect the cost of the procedure. <n> the variation in how patients tolerate different stages of injection procedure is of considerable variation in how patients tolerate different stages of injection procedure. <n> it can affect the outcome of the procedure. <n> it can affect the cost of the procedure. <n> the variation in how patients tolerate different stages of injection procedure is of considerable variation in how patients tolerate different stages of injection procedure. <n> it can affect the outcome of the procedure. <n> it can affect the cost of the procedure. <n> the variation in how patients tolerate different stages of injection procedure is of considerable variation in how patients tolerate different stages of injection procedure. (*copied from the source and repeated with hallucinations*)
Target (PMID 26286384)	Both SLT and topical medication demonstrate similar success rates and effectiveness in lowering intraocular pressure in patients with open-angle glaucoma.
Bigbird Pubmed	reduction of primary openangle glaucoma (pog) by medical therapy alone has been shown to be similar to reduction of primary openangle glaucoma by laser trabeculoplasty. in this issue of the clinical glaucoma journal,<n> two authors argue that laser trabeculoplasty can be equally efficacious as latanoprost in reducing newly diagnosed pog . (*contains irrelevant hallucinations*)
Target (PMID 26258610)	Further analysis showed that this positive relation is observed only in POAG and NTG patients, but not in the PXFG patients.
Primera	We could not find a significant link between the anti-H. pylori IgG antibodies and the PXFG. (*copied from the source; summarised for PICO: anti-H. pylori, PXFG*)

Table 11: Examples of **correct** generated summaries.