Predicting Intervention Approval in Clinical Trials through Multi-Document Summarization

Georgios Katsimpras and **Georgios Paliouras**
NCSR Demokritos, Athens, Greece
{gkatsibras,paliourg}@iit.demokritos.gr

Abstract

Clinical trials offer a fundamental opportunity to discover new treatments and advance the medical knowledge. However, the uncertainty of the outcome of a trial can lead to unforeseen costs and setbacks. In this study, we propose a new method to predict the effectiveness of an intervention in a clinical trial. Our method relies on generating an informative summary from multiple documents available in the literature about the intervention under study. Specifically, our method first gathers all the abstracts of PubMed articles related to the intervention. Then, an evidence sentence, which conveys information about the effectiveness of the intervention, is extracted automatically from each abstract. Based on the set of evidence sentences extracted from the abstracts, a short summary about the intervention is constructed. Finally, the produced summaries are used to train a BERT-based classifier, in order to infer the effectiveness of an intervention. To evaluate our proposed method, we introduce a new dataset which is a collection of clinical trials together with their associated PubMed articles. Our experiments demonstrate the effectiveness of producing short informative summaries and using them to predict the effectiveness of an intervention.

1 Introduction

Clinical Trials (CT) present the basic evidence-based clinical research tool for assessing the effectiveness of health interventions. Nevertheless, only a small number of interventions make it successfully through the process of clinical testing. Approximately, 39%-64% of interventions actually advance to the next step of each phase of clinical trials (DiMasi et al., 2010). The uncertainty of a CT outcome could lead to increased costs, prolonged drug development and ineffective treatment for the participants. At the same time, the volume of published scientific literature is rapidly growing and offers the opportunity to explore a valuable knowledge. Therefore, there is a need to develop new tools which can i) integrate such information and ii) enhance the process of intervention approval in CT.

Predicting the approval of an intervention, a task that describes the ability of a system to predict whether an intervention will reach the final stage of clinical testing, is a topic that has been studied before (Gayvert et al., 2016; Lo et al., 2018). The majority of these studies use various traditional machine learning methods and rely on structured data from various sources, including biomedical, chemical or drug databases (Munos et al., 2020; Heinemann et al., 2016). However, only a few studies take into account the textual information that is available online, and mostly in a supplementary manner (Follett et al., 2019; Geletta et al., 2019). In fact, employing natural language processing (NLP) techniques to address the outcome prediction task has been hardly explored.

Recognising this lack of related studies, the work presented here addresses the task of predicting intervention approval with the use of NLP. Particularly, we relied on generating concise and informative summaries from multiple texts that are relevant to the intervention under evaluation. In a sense, we built an intervention-specific narrative which combines key information from multiple inter-connected documents. The benefit of using multiple articles to generate summaries is that they can cover the inherently multi-faceted nature of an intervention's clinical background.

More precisely, given an intervention, our system retrieves all PubMed abstracts that are relevant to the intervention and refer to a clinical study. It then extracts the evidence sentences from each abstract using a BERT-based evidence sentence classifier, in a similar fashion to (DeYoung et al., 2020). This set of evidence sentences, which captures the consolidated narrative about the intervention, can

grow gradually, as new articles become available. Thus, further analysis is necessary in order to select the most important information. Using the set of evidence sentences for each intervention, we generate short summaries by leveraging the power of language models (BERT or BART). The resulted summaries are then fed to a BERT-based binary sequence classifier which makes a prediction about the likely approval or not of the intervention.

Overall, the main contributions of the paper are the following:

- We propose a new approach for predicting the approval of an intervention which is based on a three-step NLP pipeline.

- We provide a new dataset for the task of intervention approval prediction that consists of 704 interventions and 15,800 PubMed articles in total.

- We confirm through experimentation the effectiveness of the proposed approach.

2 Related Work

Intervention Success Prediction The prediction of intervention approval belongs to a broader category of medical prediction tasks. Relevant work includes clinical trial outcome prediction (Munos et al., 2020; Tong et al., 2019; Hong et al., 2020), drug approval (Gayvert et al., 2016; Lo et al., 2018; Siah et al., 2021; Heinemann et al., 2016), clinical trial termination (Follett et al., 2019; Geletta et al., 2019; Elkin and Zhu, 2021), predicting phase transition (Hegge et al., 2020; Qi and Tang, 2019). All these studies rely either on specific types of structured data or on combining structured data with limited unstructured data.

Differently from this line of work, the authors of (Lehman et al., 2019) proposed an approach that employs NLP to infer the relation between an intervention and the outcome of a specific clinical trial. Their method is based on extracting evidence sentences from unstructured text. An extension of this work suggests the use of BERT-based language models for the same task (DeYoung et al., 2020). Another closely related study (Jin et al., 2020), performs a large-scale pre-training on unstructured text data to infer the outcome of a clinical trial. Our approach builds upon this related work, aiming to incorporate information from multiple articles. This extension is motivated by the

assumption that the inter-connected clinical knowledge, coming from multiple sources can provide a more holistic picture of the intervention, facilitating more precise analysis and accurate prediction.

Although all these prior efforts tackle, more or less, the problem of intervention approval, none of them attempted to predict the effectiveness of an intervention using summarization methods.

Summarization The goal of summarization is to produce a concise and informative summary of a given text. There are two main categories of approaches: i) *extractive*, which tackles summarization by selecting the most salient sentences from the text without changing them, and ii) *abstractive*, which attempts to generate out-of-text words or phrases instead of extracting existing sentences. Early systems were primarily extractive and relied on sentence scoring, selection and ranking (Allahyari et al., 2017). However, both extractive and abstractive approaches have advanced significantly due to the novel neural network architectures, such as Transformers (Vaswani et al., 2017). The Transformers architecture is utilized by the BERT (Devlin et al., 2018) and BART (Lewis et al., 2019) language models which are used by the state-of-the art solutions for multiple NLP tasks, including summarization. Although most of the summarization literature focuses on single-document approaches, there is also a line of work that applies summarization on a set of documents, i.e. multi-document summarization (Ma et al., 2020). Such approaches are of particular relevance to our work, as we aim to summarize a set of sentences about a particular intervention.

Summarization in the Medical Domain Summarization has been used to address various problems in the field of medicine. These include electronic health record summarization (Liang et al., 2019), medical report generation (Zhang et al., 2019; Liu et al., 2021), medical facts generation (Wallace et al., 2021; Wadden et al., 2020) and medical question answering (Demner-Fushman and Lin, 2006; Nentidis et al., 2021).

Our work is inspired by recent work on multi-document summarization of medical studies (DeYoung et al., 2021). Apart from introducing a new summarization dataset of medical articles, that work also proposed a method to generate abstractive summaries from multiple documents. Their model is based on the BART language model, appropriately modified to handle multiple texts. Our

model differs in the way it handles the input texts. Instead of concatenating all texts into a single representative document, we order them chronologically and split them into equal-size chunks. Doing so, we expect the clinical studies that were conducted during a similar time period, to reside in the same chunk.

3 Task Overview

According to the U.S. Food and Drug Administration (FDA), a CT addresses one of five phases of clinical assessment: Early Phase 1 (former Phase 0), Phase 1, Phase 2, Phase 3 and Phase 4. Each phase is defined by the study's objective, the interventions under evaluation, the number of participants, and other characteristics[1]. Notably, Phase 4 clinical trials take place after FDA has *approved* a drug for marketing. Therefore, we can assume that a CT in Phase 4 assesses effective intervention. On this basis, our task is to predict whether an intervention will advance to the final stage of clinical testing (Phase 4), as shown in Figure 1.

We model the task of predicting the success or failure of an intervention as a binary classification task. All data relevant to Phase 4 are omitted from the training stage.
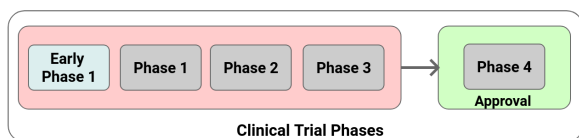

Figure 1: The phases of a clinical trial.

4 Data

In this work, we introduce a new dataset[2] for the task of predicting intervention approval. The dataset is a collection of structured and unstructured data in English derived from clinicaltrials.gov and PubMed during May-June 2021.

As a first step in the construction of the dataset, we retrieve all available CT studies from clinicaltrials.gov that satisfy some criteria. Then, we associate each CT with PubMed articles based on the CT study identifier. Following some cleaning process (i.e. deduplication and entity resolution) we generate the final dataset.

Clinical Trials Studies At the time of writing, more than 350,000 studies were available online at *clinicaltrials.gov*. We focused on *cancer* related clinical testing and we retrieved approximately 85,000 studies related to this topic using a list of associated keywords[3].

From this set, we were interested in interventional clinical trials and specifically in two categories that indicate the status of the trial: i) *"Completed"*, meaning that the trial has ended normally, and ii) *"Terminated"*, meaning that the trial has stopped early and will not start again. The resulting set of studies contains 34,517 completed and 6,872 terminated trials.

Interventions Dataset Using the selected CTs, we associated each intervention with its corresponding trials. Therefore, a clinical trial record was formed for each intervention. Then, we selected all interventions that are assessed in at least one Phase 4 CT to form our positive target class (i.e. approval). Likewise, we built our negative target class (i.e. termination) using interventions that led to a trial termination. In total, our dataset contains 404 approved and 300 terminated interventions.

For each intervention, we collect all articles from PubMed that are explicitly related to one of the CTs of the intervention. To achieve this, we combine two approaches. First, we search for eligible articles (or links to articles) in the corresponding structured results of *clinicaltrials.gov*. Secondly, we use the CT unique identifiers to query the PubMed database. Then, the selected PubMed articles are associated with the intervention. This way an intervention is linked with multiple studies that are inter-connected, and thus an intervention-specific narrative is developed. In our dataset, an intervention is associated on average with 22.4 pubmed articles, though for terminated interventions this number is just 1.4. This is because terminated interventions are usually not assessed in many CTs. Overall, our dataset contains 15,800 pubmed articles. The details of the dataset are presented in Table 1.

In addition, we attempted to evaluate[4] our approach on a previously used dataset (Gayvert et al., 2016), which consists of 884 (784 approved, 100 terminated) drugs along with a set of 43 features, including molecular properties, target-based properties and drug-likeness scores.

---

[1] https://clinicaltrials.gov/

[2] https://github.com/nneinn/ct_intervention_approval

[3] The complete list of the keywords used is: cancer, neoplasm, tumor, oncology, malignancy, neoplasia, neoplastic syndrome, neoplastic disease, neoplastic growth and malignant growth

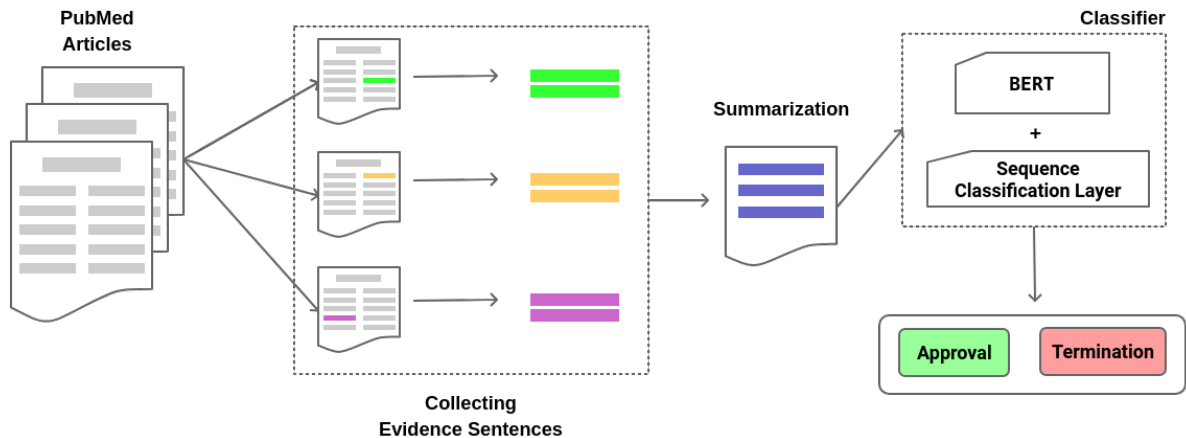[4] The results on this dataset are presented in appendix A

Figure 2: Overview of the proposed approach for classifying an intervention.

Type	|I|	|A|	avg
Approved	404	15,379	38.1
Terminated	300	421	1.4
Total	704	15,800	22.4

Table 1: The details of the interventions dataset. |I|, |A| and *avg* denote the number of interventions, the number of articles and the average number of articles per intervention respectively.

5 Methodology

In Figure 2, we illustrate the proposed approach, which consists of three main steps. Initially, we use the abstracts of the intervention's clinical trial record to extract evidence sentences. These sentences are then used to generate a short summary that contains information about the efficacy of the intervention. The summary is then processed by a BERT-based sequence classifier to make the final decision about the intervention. Each of the three steps is detailed in the following subsections.

5.1 Evidence Sentences

Identifying evidence bearing sentences in an article for a given intervention is an essential step in our approach. Differently from other sentences in an article, evidence sentences contain information about the effectiveness of the intervention (Figure 3). Therefore, it is crucial that our model has the ability to discriminate between evidence and non-evidence sentences.

First, all abstracts related to the given intervention are broken into sentences. The sentences of each abstract are then processed one-by-one by a BERT-based classifier that estimates the probability of each sentences containing evidence about the

effectiveness of the intervention. For the classifier, we selected a version of the PubMedBERT (Gu et al., 2020) model, which is pre-trained only on abstracts from PubMed. We tested several models, including BioBERT (Lee et al., 2020), clinical-BERT (Alsentzer et al., 2019) and RoBERTa (Liu et al., 2019), but PubMedBERT performed the best in our task. On top of PubMedBERT, we trained a linear classification layer, followed by a Softmax, using the dataset from (DeYoung et al., 2020). This dataset is a corpus especially curated for the task of evidence extraction and consists of more than 10,000 annotations. The classifier is trained with annotated evidence sentences (i.e. positive samples) and a random sample of non-evidence sentences (i.e. negative samples). Regarding the ratio of positive to negative samples, cross-validation on the training set showed 1:4 to be a reasonable choice. The evaluation of the different BERT-based models was done based on the same data splits (train, test and validation) as in (DeYoung et al., 2020).
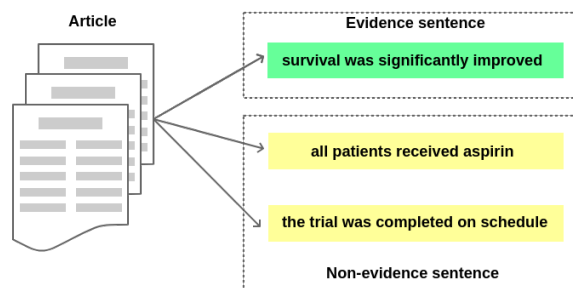

Figure 3: Evidence sentence identification. The evidence sentences constitute the positive instances whereas the non-evidence sentences the negative ones.

Once scored by the classifier, the highest scoring

sentence is selected from each abstract. Therefore, for each intervention we extract as many sentences as the number of abstracts in its clinical record.

5.2 Short Summaries

To generate short and informative summaries we explore both extractive and abstractive approaches.

Extractive Summaries were based on the evidence sentences extracted in the previous step. Specifically, we re-rank them and choose the top k ($k = 5$) to compose our final summary. The model we use here is the same BERT-based model as in Section 5.1.

Abstractive Considering that an intervention is linked to multiple abstracts and thus to multiple evidence sentences, we first order all evidence sentences chronologically and combine them into a single text. Then, we split them to equal chunks[5] and each chunk then is fed to a BART-based model to produce the final summary.

BART has been shown to lead to state-of-the-art performance on multiple datasets (Fabbri et al., 2021). Specifically, we used the pre-trained distilBART-cnn-12-6 model which is trained on the CNN summarization corpus (Lins et al., 2019). Since abstractive summarization produces out-of-text phrases, it needs to be fine-tuned with domain knowledge. In our case, we fine-tuned the BART model with the MS2 dataset (DeYoung et al., 2021), which contains more than 470K articles and 20K summaries of medical studies.

We limited the length of the output summary to 140 words. For the extractive setting, in case the top k sentences exceeded this limit, we removed the extra words. For the abstractive setting we iteratively summarized and concatenated the chunks for each intervention until the expected number of 140 words was accomplished.

5.3 Inferring Efficacy

We model the task of inferring the approval of an intervention as a binary classification task. In our approach, each intervention is represented by a short summary. For the classification of the summaries, we used again a PubMedBERT model. On top of it, we trained a linear classification layer, followed by a sigmoid, using the summaries generated in the previous step: Our positive training instances were the summaries of interventions that

---

[5] A chunk has length equal to the maximum input length of the BART model (1024).

have been approved, and correspondingly, the negative ones were the summaries of interventions that have been terminated. Hence, the model decides on the approval of the interventions.

5.4 Technical set-up

All models were pre-trained and fine-tuned for the corresponding task. The maximum sequence size was 512 and 1024 for BERT-based and BART-based models respectively. The Adam optimizer (Kingma and Ba, 2015) was used to minimize the cross-entropy losses with learning rate 2e-5 and epsilon value 1e-8 for all models. We trained all models for 5 epochs, with batch sizes of 32, except the abstractive summarizer for which the batch size was decreased to 4 due to RAM memory limitations of our system. The implementation was done using the HuggingFace library (Wolf et al., 2020) and Pytorch(Paszke et al., 2019).

6 Results and Analysis

We followed different training approaches for the different trainable components of our pipeline. For the evidence sentence selection and the abstractive summarization models we split the data into development and test and then split the development set further into training (90%) and validation (10%). We kept the model that performed best on the validation set and evaluated it on the held-out test set of each task respectively, averaged over three random data splits. Considering the small size of the interventions dataset, we applied a 10-fold cross validation for the final classification task. For this task, we report macro averages of the evaluation metrics over the ten folds.

6.1 Ablation Study

Our experimentation started with a comparison of different variants and choices that were available for the various modules of our approach.

Evidence Classifier Coming early in the pipeline, the performance of the evidence classifier can play a significant role in downstream tasks. The chosen approach relied on domain-specific BERT models. As domain-specific training that can affect the performance of BERT-based models, we conducted a comparison between different variants of BERT. The results in Table 2 demonstrate that the performance of the models is comparable, with all models obtaining scores over 90% in terms of F1 and AUC. PubMedBERT model achieved the best

scores and was used in the rest of the experiments.

Model	P	R	F1	AUC
BioBERT	.928	.938	.933	.957
ClinicalBERT	.913	.925	.919	.945
RoBERTa	.905	.919	.912	.931
PubMedBERT	**.931**	**.956**	**.943**	**.969**

Table 2: The results of the domain-specific BERT variants that were used for the evidence classifier. All models were trained with negative sampling ratio 1:4. The results denote the averages over three random train-test splits.

Summarization Adequacy We assess the performance of the summarization methods on the MS2 dataset which is a collection of summaries extracted from medical studies. The task of the summarizers is to produce texts that approximate the target summaries. We measure the performance of the summarization methods using ROUGE and the results are presented in Table 3. As expected, the abstractive method achieves higher scores, as it has more flexibility in forming summaries. We also observed that domain-specific training improves performance. The $abstractive_{no}$ model is a generic BART model without fine-tuning in the domain. Comparing its performance to the $abstractive$ model, which was fine-tuned on a small sample of the MS2 dataset that was excluded from the evaluation process, we notice a statistically significant improvement.

Model	R-1	R-2	R-L
abstractive$_{no}$	24.85	4.34	15.48
abstractive	**39.38**	**11.98**	**20.13**
extractive	19.24	3.22	13.19

Table 3: Evaluation of summarization methods on the MS2 dataset. The abstractive$_{no}$ refers to the generic BART model without any fine-tuning in the domain.

Abstractive methods seem to provide better summaries, however, whether these are more useful than the extractive summaries for our donwstream task is still to be determined.

6.2 Predicting Intervention Efficiency

Having made the choices for the individual modules, we now turn to the ultimate task, which is the prediction of the efficiency of the intervention. We evaluate two variations of our proposed method; i) with abstractive summarization denoted as PIAS$_{abs}$ and ii) with extractive summarization denoted as

PIAS$_{ext}$. We compare their performance against two baselines:

- **BS**: This is a PubMedBERT model that is trained with a single evidence sentence per intervention (instead of a summary). The sentence is extracted from the most recent PubMed article relevant to the intervention.

- **BN**: This is similar to BS but instead of using a single sentence for each intervention it is trained with n evidence sentences extracted from n different articles ($n = 3$). The articles are selected randomly among the ones referring to the intervention.

The performance of all models is shown in Table 4. The proposed method outperforms the baselines independent of the summarization methods that is used. Interestingly, even selecting randomly selected evidence sentences seem to help, as BN achieved a higher performance than BS. Still, the use of summarization provides a significant boost over both baseline methods, validating the value of using short summaries to evaluate the efficiency of an intervention. Models that do not take advantage of the inter-connected documents suffer a significant drop in performance. Thus, this result justifies the design of the proposed method.

We can also observe that the best performance of the proposed method is achieved when using the extractive summarization method. Extractive summaries have demonstrated low ROUGE scores in Section 6.1. Still, they can properly capture the properties involved in the data for the classification task. On the other hand, although the abstractive summarizer achieved better ROUGE scores, it seems that the generated summaries cannot discriminate the target classes (approved or terminated) as well as the extractive ones. This indicates that the quality of the summary, in terms of the ROUGE score, is not decisive in the classification of the intervention.

Model	P	R	F1
BS	.717	.706	.702
BN	.732	.731	.731
PIAS$_{abs}$	.781	.774	.773
PIAS$_{ext}$	**.796**	**.793**	**.792**

Table 4: The classification results of all models. The reported precision, recall and f1 scores the macro averages over ten folds.

Analyzing further the performance of our best model, PIAS$_{ext}$, we report macro average scores for each target class in Table 5. We notice that the

class	P	R	F1
positive (approved)	.808	.819	.815
negative (terminated)	.778	.765	.772

Table 5: The performance of our best model, i.e. PIAS$_{ext}$, for each target class. The scores denote macro averages over ten folds.

model is slightly better at predicting the approval of an intervention rather than its termination. This can be explained by the fact that the approved interventions are associated with a considerably larger number of articles than the terminated ones. This leads to richer summaries for the approved interventions and thus to a more informed decision.

6.3 Predicting Phase Transition

Early prediction of approval To build our models, we considered all the available data from Phase 1, Phase 2 and Phase 3. However, predicting the success of an intervention at the earliest phase possible is compelling. Therefore, we examine the ability of our model in making early predictions. More precisely, we evaluate PIAS$_{ext}$ model on the following three transitions: Phase 1 to Approval, Phase 2 to Approval and Phase 3 to Approval.

To perform this experiment, we select the interventions that have CTs in various stages and there is least one article for each phase. In total, this subset contains 249 interventions (193 approved and 56 terminated). Then, we use 80% for training and 20% for testing. For each transition, we train our model only with training instances from the corresponding phase. In Table 6, we report the macro average scores over ten random splits of the data.

transition	P	R	F1
phase1 $\xrightarrow{to}$ approval	.39	.50	.44
phase2 $\xrightarrow{to}$ approval	.78	.70	.72
phase3 $\xrightarrow{to}$ approval	.81	.84	.82

Table 6: The performance of our best model, i.e. PIAS$_{ext}$, in predicting phase-to-approval transitions. The scores denote the averages over ten random runs.

The results indicate that prediction of approval, while at Phase 1 is very hard, but the transition

from Phase 2 and Phase 3 to approval can be predicted with considerable success. The large gap in performance between Phase 1 and Phase 2, 3 transitions is explained by the lack of clinical evidence in early phases.

Phase to Phase Another interesting and challenging task is to predict the transition of an intervention to the next phase of the clinical trial process. In this experiment, we want to predict Phase 1 to Phase 2 and Phase 2 to Phase 3 transitions. For each transition, we use data only from the former phase for training (e.g. for Phase 2 to Phase 3 transition we use data from Phase 2) for both target classes. Again, we use 80% for training and 20% for testing and present the average scores over ten random splits.

transition	P	R	F1
phase2 $\xrightarrow{to}$ phase3	.84	.82	.83
phase1 $\xrightarrow{to}$ phase2	.77	.76	.77

Table 7: The performance of our best model, i.e. PIAS$_{ext}$, in predicting phase-to-phase transitions. The scores denote the averages over ten random runs.

Table 7 shows the results for the two transitions, which are comparable to the overall predictive performance of the model. Considering the small size of the datasets used in both phase transition tasks, these results can serve only as an indication of how our model behaves. Further analysis and experiments should be conducted for a more thorough evaluation.

6.4 Explainability of Predictions

It is clinically very valuable to identify the factors that contribute most to a particular decision of the classifier. Interestingly, the summaries generated from our models can also serve that purpose very well.

Table 8 illustrates some examples of interventions along with their abstractive and extractive summaries as produced by our pipeline. For the first intervention, *pertuzumab*, it is notable that both summaries report a improved median progression-free survival which somewhat explains the prediction. For the second intervention, *taxane*, the summaries mention the greater incidence of serious adverse events and lower median overall survival, which counts against the approval of the intervention. We also notice that many numerical entities are randomly placed or changed in

Intervention		PIAS$_{abs}$	PIAS$_{ext}$
pertuzumab	✓	the primary endpoint of the study is progression-free survival. median progression-free survival was 12.4 months in the control group, as compared with 18.5months in the pertuzumab group. median survival was <dig> months, 12.3 months, and 12.5 months, respectively, in the p=0·0141 group and p =0·0% in the qtl group, respectively. the p <dig) group had a significantly improved pathological complete response rate compared with the group without complete response. p=dig> month and qtl were the most significantly different groups in both groups. p =dig>	Disease-free survival results were consistent with progression-free survival results and were 81% (95% CI 72-88) for group A, 84% (72-91) for group B, 80% (70-86) for group C, and 75% (64-83) for group D. Patients who achieved total pathological complete Three patients [1.5%; 95% confidence interval (CI) 0.31% to 4.34%] in cohort A experienced four New York No evidence of DDIs for pertuzumab on trastuzumab, trastuzumab on pertuzumab, or pertuzumab on chemotherapy PK was observed. The median progression-free survival (PFS) among patients who received NAT was 15.8 months compared with CNS ORR was 11% (95% CI, 3 to 25), with four partial responses (median duration of response, 4.6 months).
taxane	✗	the most common serious adverse events were anaemia, upper gastrointestinal haemorrhage, pneumonia, and pneumonia in the trastuzumab emtansine 24 mg/kg weekly group compared with pneumonia, febrile neutropenia, and anaemia in the taxane group. median overall survival was 11.8 months with trastzumab 2.4 mg/ kg weekly and 10.0 months with taxane.2) with taxanes.3) with t-dm1 was not associated with superior os or superior os versus taxane in any subgroup.5–10% of the patients with high body weight and low baseline trast	The most common serious adverse events were anaemia (eight 4), upper gastrointestinal haemorrhage (eight 4), pneumonia (seven 3), gastric haemorrhage (six 3), and gastrointestinal haemorrhage (five 2) in the trastuzumab emtansine 24 mg/kg weekly group compared with pneumonia (four 4), febrile neutropenia (four 4), anaemia (three 3), and neutropenia (three 3) in the taxane group. Median overall survival was 11.8 months (95 confidence interval ci, 9.3-16.3) with trastuzumab emtansine 2.4 mg/kg weekly and 10.0 months (95 ci, 7.1-18.2) with taxane (unstratified hazard ratio 0.94, 95 ci, 0.52-1.72).

Table 8: Examples of generated summaries from our models. These summaries can be used to explain the predictions of the classifier. The second column displays the prediction of the classifier for the specific intervention; ✓ denotes approval and ✗ denotes termination.

the abstractive summary. This contributes to the tendency of the abstractive methods to generate "hallucinated" evidence, as observed in the literature (Cao et al., 2018). However, the abstractive summaries look more readable. A more exhaustive analysis, including also a human evaluation, is needed to assess the ultimate explainability of these summaries.

7 Conclusion

Predicting intervention approval in clinical trials is a major challenge with significant impact in healthcare. In this paper, we have proposed a new pipeline to address this problem, based on state-of-the-art NLP techniques. The proposed method consists of three steps. First, it identifies evidence sentences from multiple abstracts related to an intervention. Then, these sentences are used to produce short summaries. Finally, a classifier is trained on the generated summaries in order to predict the approval or not of an intervention.

Moreover, we introduced a new dataset for this task which contains 704 interventions associated with 15,800 abstracts. This data was used to evaluate our pipeline against other baseline models. The experimental results verified the effectiveness of our approach in predicting the approval of an intervention and the contribution of each step of the proposed pipeline to the final result. Further evaluation on predicting phase transitions, showed that our model can assist in all stages of a clinical trial. Besides, the generated multi-document summaries can be naturally used to explain the predictions of the model.

There are multiple ways to extend this work. In terms of multi-document summarization, there is room to explore more advanced summarization models, quality and performance metrics as well as better explainability assessment. In the bigger picture, we shall also consider to expand the dataset by extending its size and incorporating different types of resources (e.g. drug interaction networks). Finally, we are interested in enhancing the proposed method to incorporate temporal information associated with the CTs to maintain the history of clinical changes.

Acknowledgements

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *International Journal of Advanced Computer Science and Applications*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 841–848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Joseph A DiMasi, Lanna Feldman, Abraham Seckler, and Andrew Wilson. 2010. Trends in risks associated with new drug development: success rates for investigational drugs. *Clinical Pharmacology & Therapeutics*, 87(3):272–277.

Magdalyn E Elkin and Xingquan Zhu. 2021. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. *Scientific reports*, 11(1):1–12.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Lendie Follett, Simon Geletta, and Marcia Laugerman. 2019. Quantifying risk associated with clinical trial termination: a text mining approach. *Information Processing & Management*, 56(3):516–525.

Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301.

Simon Geletta, Lendie Follett, and Marcia Laugerman. 2019. Latent dirichlet allocation in predicting clinical trial terminations. *BMC medical informatics and decision making*, 19(1):1–12.

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 preprint arXiv:2007.15779*.

Stephan J Hegge, Markus Erik Thunecke, Matthias Krings, Leonard Ruedin, Jan Saputra Mueller, and Paul von Buenau. 2020. Predicting success of phase iii trials in oncology. *medRxiv*.

Fabian Heinemann, Torsten Huber, Christian Meisel, Markus Bundschus, and Ulf Leser. 2016. Reflection of successful anticancer drug development processes in the literature. *Drug discovery today*, 21(11):1740–1744.

Zhen Yu Hong, Jooyong Shim, Woo Chan Son, and Changha Hwang. 2020. Predicting successes and failures of clinical trials with an ensemble ls-svr. *medRxiv*.

Qiao Jin, Chuanqi Tan, Mosha Chen, Xiaozhong Liu, and Songfang Huang. 2020. Predicting clinical trial results by implicit evidence integration. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

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

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. *NAACL HLT 2019 - 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using ehr data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54.

Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral, Jamilson Batista, Bruno Tenorio, Rafael Ferreira, Rinaldo Lima, Gabriel de França Pereira e Silva, and Steven J. Simske. 2019. The cnn-corpus: A large textual corpus for single-document extractive summarization. In *Proceedings of the ACM Symposium on Document Engineering 2019*, DocEng '19, New York, NY, USA. Association for Computing Machinery.

Fenglin Liu, Shen Ge, and Xian Wu. 2021. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew W Lo, Kien Wei Siah, and Chi Heem Wong. 2018. Machine learning with statistical imputation for predicting drug approvals. *Available at SSRN 2973611*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *arXiv preprint arXiv:2011.04843*.

Bernard Munos, Jan Niederreiter, and Massimo Riccaboni. 2020. Improving the prediction of clinical success using machine learning.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2021. Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 239–263. Springer.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Youran Qi and Qi Tang. 2019. Predicting phase 3 clinical trial results by modeling phase 2 clinical trial subject level data using deep learning. In *Machine Learning for Healthcare Conference*, pages 288–303. PMLR.

Kien Wei Siah, Nicholas Kelley, Steffen Ballerstedt, Björn Holzhauer, Tianmeng Lyu, David Mettler, Sophie Sun, Simon Wandel, Yang Zhong, Bin Zhou, et al. 2021. Predicting drug approvals: The novartis data science and artificial intelligence challenge. *Available at SSRN 3796530*.

Ling Tong, Jake Luo, Ron Cisler, and Michael Cantor. 2019. Machine learning-based modeling of big clinical trials data for adverse outcome prediction: A case study of death events. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 269–274. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. 2019. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541.*

A Results on Proctor Dataset

To further evaluate our method, we attempted a comparison with the method presented in (Gayvert et al., 2016) using their data. The data contains a list of approved and terminated drugs together with various features. Using this dataset, we experienced two issues that made the comparison incomparable: i) For many drugs we could not find relevant articles in PubMed. The original dataset contains 828 drugs whereas we managed to collect information only for 537. Thus, the scores of our method are not directly comparable to the ones reported in (Gayvert et al., 2016) ii) Four important features that were used in (Gayvert et al., 2016) are missing in the dataset. Therefore, the reproduction of the exact model is not possible.

Despite these facts, we performed a comparison of the methods for the subset that we collected:

- **RF$_1$**: This model reports the scores from (Gayvert et al., 2016).

- **RF$_2$**: This is a Random Forest model similar to the original one, but it is trained only with the available features.

The overall performances of all models are depicted in Table 9.

Model	AUC
RF$_1$	.826
RF$_2$	.484
PIAS$_{ext}$	.586

Table 9: The classification results of all models on the Proctor dataset. The reported precision, recall and f1 scores are macro averages over ten folds.