Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials

Tobias Mayer, Santiago Marro, Serena Villata, Elena Cabrio

Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials

Tobias Mayer[†], Santiago Marro[†], Serena Villata[†], and Elena Cabrio[†]

[†]Université Côte d'Azur, CNRS, Inria, I3S, France

October 2020

Abstract

In the latest years, the healthcare domain has seen an increasing interest in the definition of *intelligent* systems to support clinicians in their everyday tasks and activities. Among others, also the field of Evidence-Based Medicine is impacted by this twist, with the aim to combine the reasoning frameworks proposed thus far in the field with mining algorithms to extract structured information from clinical trials, clinical guidelines, and Electronic Health Records. In this paper, we go beyond the state of the art by proposing a new end-to-end pipeline to address argumentative outcome analysis on clinical trials. More precisely, our pipeline is composed of *(i)* an Argument Mining module to extract and classify argumentative components (i.e., evidence and claims of the trial) and their relations (i.e., support, attack), and *(ii)* an outcome analysis module to identify and classify the effects (i.e., improved, increased, decreased, no difference, no occurrence) of an intervention on the outcome of the trial, based on PICO elements. We annotated a dataset composed of more than 500 abstracts of Randomized Controlled Trials (RCT) from the MEDLINE database, leading to a labeled dataset with 4198 argument components, 2601 argument relations, and 3351 outcomes on five different diseases (i.e., *neoplasm*, *glaucoma*, *hepatitis*, *diabetes*, *hypertension*). We experiment with deep bidirectional transformers in combination with different neural architectures (i.e., LSTM, GRU and CRF) and obtain a macro F1-score of .87 for component detection and .68 for relation prediction, outperforming current state-of-the-art end-to-end Argument Mining systems, and a macro F1-score of .80 for outcome classification.

1 Introduction

In the healthcare domain, there is an increasing interest in the development of *intelligent* systems able to support and ease clinicians' everyday activities. These systems deal with heterogeneous kinds of data, spanning from textual documents to medical images to biometrics. Concerning textual documents (e.g., clinical trials, clinical guidelines, and Electronic Health Records), such solutions range from the automated detection of PICO elements [1] in health records to evidence-based reasoning for decision making [2, 3, 4, 5]. These applications aim at assisting clinicians in their everyday tasks by extracting, from unstructured textual documents, the exact information they necessitate and to present this information in a structured and machine-readable format, easy to be (possibly semi-automatically) analyzed. The ultimate goal is to aid the clinician's deliberation process [6].

Analyzing argumentation from the computational linguistics point of view has recently led to a new field called Argument(ation) Mining (AM) [7, 8, 9, 10] which deals with detecting, classifying and assessing the quality of argumentative structures in text. Standard tasks in AM are the detection of argument components (i.e., *evidence* and *claims*), and the prediction of the relations (i.e., *attack* and *support*) holding among them. Our motivation to rely on argumentation mining stems from its aptness in providing us with methods to automatically detect in text the argumentative structures that are at the basis of Evidence-Based Medicine (EBM), which is the *"conscientious, explicit, and judicious use of current best evidence"* [11] to guide clinical decision-making with scientific information from systematic reviews. These information needs cannot be directly tackled by current methods (e.g., clinical document classification [12], clinical question answering [13], or extractive summarization [14]), and require the development of novel approaches within the argumentation mining field.

Alas, despite its natural employment in healthcare applications, only few approaches have applied AM methods to this kind of text [15, 16, 17, 18], and their contribution is limited to (part of) the Argument Mining pipeline, disregarding the combination of this argumentative information with other kind of clinical data that clinicians usually look at when searching for relevant evidence. Indeed, when clinicians search for relevant evidence, they use a specialised framework called PICO, which stands for ***P****atient Problem* or *Population*, ***I****ntervention*, ***C****omparison* or *Control*, and ***O****utcome*. The idea is to ask well-built clinical questions [19], which should be answered by clinical trials. Searching for relevant trials and finding meaningful answers is a time consuming and laborious task for clinicians. Automating this process of evidence collection and argumentative analysis from documents could unburden the clinicians substantially.

In our work, we take up these challenges and formulate them into the following research questions:

- How to adapt models from argumentation theory on large corpora of clinical text for modeling argumentation and outcome-based evidence in Randomized Controlled Trials?

- What computational approaches can be used to analyze arguments and evidence on outcomes in Randomized Controlled Trials?
- What is the impact of argumentative structures and PICO elements on evidence-based deliberation?

In this article, we answer to these research questions, with the goal of addressing the previously discussed challenges and issues. First, we apply a structured argumentation model [20] combined with the effects of an intervention on an outcome from PICO evidence to manually annotate a new huge resource of 660 abstracts of Randomized Controlled Trials. Second, we propose a deep bidirectional transformer approach combined with different neural networks to address in a pipeline both the AM tasks of component detection and relation prediction, and the outcome classification in Randomized Controlled Trials, evaluating it on the corpus we annotated. Third, we discuss the impact of argumentative information and PICO evidence on the clinician's deliberation process both with respect to the data contained in the annotated dataset and to the results of the end-to-end pipeline we propose.[1]

To summarize, the contributions of this paper are as follows:

- We create a new dataset which is, to the best of our knowledge, the largest dataset that has been annotated within the argumentation mining field on clinical data. The dataset is built from the MEDLINE database, consisting of 4198 argument components and 2601 argument relations on five different diseases (*neoplasm*, *glaucoma*, *hepatitis*, *diabetes*, *hypertension*). A novel aspect of the corpus is the annotation of the effects (i.e., improved, increased, decreased, no difference, no occurrence) of an intervention on 3351 outcomes.
- We experiment on the annotated data using various Machine Learning methods relying on deep bidirectional transformers combined with different neural networks, i.e., Long Short-Term Memory (LSTM) networks, Gated Recurrent Unit (GRU) networks, and Conditional Random Fields (CRFs) in order to extract argument structure and classify outcomes from RCTs. We propose several novel feature sets and identify configurations that run best in in-domain and cross-domain scenarios depending on the diseases present in the dataset. To foster research in the community, we provide the annotation guidelines, the annotated data as well as all the experimental software.[2];

[1] We used the dataset and core methods from this article in our publication Mayer et al., 2020 [18]. The main difference is that this article focuses mainly on the argument-based annotation and analysis of the dataset, and it introduces the outcome annotation and classification together with the discussion of the argumentation and outcome analysis for evidence-based deliberation, whereas in Mayer et al., 2020 [18] we study what are the best methods for identifying argument components and predicting argument relations in Randomized Controlled Trials.

[2] The source code is available here: `https://gitlab.com/tomaye/ecai2020-transformer_based_am` and the dataset together with the annotation guidelines is available here: `https://gitlab.com/tomaye/abstrct`

- Our extensive evaluation allows to characterize argumentative components using the effects on the outcomes we classified, such that we can now identify for instance when a claim reports about an outcome as being *safe* or *efficient* but the associated side effects are classified as *increased*. This combined analysis reveals more fine-grained categorization of the statements in RCTs.

The paper is organised as follows. In Section 2, we discuss the related literature pointing out the main advantages of our approach. Section 3 defines the main guidelines of our annotation studies, and describes the creation process of our annotated dataset of clinical trials. Section 4 introduces our argumentative outcome analysis pipeline and discusses the methodological choices we address in conceiving it. In Section 5, we detail the experimental setting and we report on the obtained results together with an in-depth error analysis. Conclusions end the paper.

2 Related Work

In this section, we first introduce the main achievements in the area of Argument Mining, and then we discuss the main results presented in the literature to apply the Argument Mining pipeline to different application scenarios, highlighting the main advantages of our approach and the peculiarity of the clinical trial scenario. Finally, we present the other approaches to evidence-based medicine, stressing the importance of combining both argumentative components and PICO elements to achieve more insightful analyses of clinical trials.

Argument Mining One of the latest advances in the field of artificial argumentation [21] deals with the automatic processing of text to extract argumentative structures. This new research area is called *Argument(ation) Mining* (AM) [7, 8, 9, 10], and it mainly consists of two standard tasks: *(i)* the identification of arguments within the text, that may be further split in the detection of argument components (e.g., claims, evidence) and the identification of their textual boundaries; *(ii)* the prediction of the relations holding between the arguments identified in the first stage. These relations are then used to build the argument graphs, where the retrieved argumentative components represent the nodes of the graph and the predicted relations correspond to the edges. Different methods have been employed to address these tasks, from standard Support Vector Machines (SVMs) to Neural Networks (NNs). AM methods have been applied to heterogeneous types of textual documents, e.g., persuasive essays [22], scientific articles [23], Wikipedia articles [24], political speeches and debates [25, 26], and peer reviews [27]. However, only few approaches [28, 15, 16, 17, 18] focused on automatically detecting argumentative structures from textual documents in the medical domain, such as clinical trials, clinical guidelines, and Electronic Health Records.

Argument Mining pipeline The whole AM pipeline (i.e., mining both argumentative components and the relations connecting them) has been implemented in few application scenarios. In particular, Stab and Gurevych [22] propose a feature-based Integer Linear Programming approach to jointly model argument component types and argumentative relations in persuasive essays. Differently from our data, essays have exactly one major claim each. The authors impose the constraint such that each claim has no more than one parent, while no constraint holds in our case. In contrast with this approach, Eger et al. [29] present neural end-to-end learning methods in AM, which do not require the hand-crafting of features or constraints, using the persuasive essays dataset. They employ TreeLSTM on dependency trees [30] to identify both components and relations between them. They decouple component classification and relation classification, but they are jointly learned, using a dependency parser to calculate the features. In this paper, we also decouple the two classification tasks, in line with the claim of [29] that decoupling component and relation classification improves the performance. Furthermore, the same work addresses component detection as a multi-class sequence tagging problem [31]. Differently from their approach, which does not scale with long texts as it relies on dependency tree distance, our approach is distance independent. In addition, whilst persuasive essay components are usually linked to components close by in the text, in our dataset links may span across the whole RCT abstract.

Ajjour et al. [32] proposed a deep learning approach for segmentation of text into argument units. Here, the task is, again, formulated as a sequence tagging problem, where a label is assigned to each token following the BIO-tagging scheme. The authors only tackle the argument unit segmentation (argumentative vs non-argumentative) without the further classification of the components. Contrary to the performed five class argument component detection, this translates to a three class classification problem, i.e., *Arg-B, Arg-I* and *Arg-O*. The best performing model consists of two BiLSTM, where one is using word embeddings and the other syntactic, structural and pragmatic input features (one-hot vectors). Both BiLSTM outputs are concatenated and put through a dense layer before it is passed to another (upper) BiLSTM. The output of the last (upper) BiLSTM is used in the final classification layer. The authors noted an decreased number of invalid BI sequences with the addition of the second (upper) BiLSTM. In later work, the authors in [33] further investigated this architecture with minor changes: they used solely one BiLSTM with word embeddings as input features and tested the efficacy of the second (upper) BiLSTM. Moreover, they investigated the effects of adding various attention layers. The results did not show any major changes in performance with respect to adding the second (upper) BiLSTM. Also, the addition of attention layers did not improve the results. In line with these observations, no stacked RNNs or attention layers are added for the sequence tagging architectures evaluated for argument component detection in this work. The idea is to reduce the number of invalid BI sequences not with a second (upper) RNN layer, but with a CRF. Recent approaches for link prediction rely on pointer networks [34] where a sequence-to-sequence model with attention takes as input argument components and returns

the links between them. In these approaches, neither the boundary detection task nor the relation classification one are tackled. Another approach to link prediction relies on structured learning [35]. The authors propose a general approach employing structured multi-objective learning with residual networks, similar to approaches on structured learning on factor graphs [36]. Recently, the argument classification task was addressed with contextualized word embeddings [37]. However, differently from our approach, they assume components are given, and boundary detection is not considered. In line with their work, we experimented with the BERT [38] base model to address parts of the AM pipeline [17]. Contrary to this preliminary work, we employed and evaluated various contextualized language models and architectures on each task to span the full AM pipeline as well as the outcome analysis.

Evidence-based medicine Only few approaches have applied AM methods to the kind of text relevant for systematic reviews [15, 16], and their contribution is limited to the detection of argument components. In addition, no huge annotated dataset for AM is available for the healthcare domain. There exists a corpus of contradicting claims [39], which was created using research abstracts of studies considered in systematic reviews related to cardiovascular diseases, but this corpus does not contain the corresponding evidence backing those claims. Some systems assisting in automatic evidence extraction have been proposed in the literature. For instance, ExaCT [40] extracts information containing PICO elements based on a SVM. It was designed to search full text articles, but was limited by the scarce training data available. Whereas nowadays, there is the EBM-NLP corpus [41], which is a collection of considerable size of sentences annotated with PICO elements. Similarly, Trenta et al. [42] proposed a maximum entropy classifier to mine characteristics of randomized clinical trials in form of PICO elements. Their dataset comprises 99 manually annotated abstracts. Recently, Jin and Szolovits [43] proposed deep learning models to address PICO elements detection, such as BiLSTM CRF combinations, and methods to improve the generalization of these models. Another system facilitating the evidence gathering process is RobotReviewer [44], which summarizes the key information of a clinical trial. These key information comprise the interventions, trial participants and risk of bias, where the latter is related to finding potential design flaws of the studies. Recently, Lehman et al. [45] proposed an approach to infer if a study provides evidence with respect to a given intervention, comparison and outcome. Additionally to the classification, the model returns a sentence from the document supporting the classification result. These *rationals* [46] are important evidence which support the classification result in a human readable way. Our approach is similar, but focuses more on these rationals, which we call argument components. While they start with a prompt of PICO elements, we set up our pipeline in the opposite direction, where the first step is to find evidence in the form of an argumentation graph, which is human readable and then, in a subsequent step, enrich these graphs with information about the contained PICO elements. This way,

the comparison between the intervention and comparator, which is an essential part of medical evidence, is not explicitly modelled in a machine-readable format, as the above mentioned Evidence Inference task does it. This information about the direct comparison is only available in form of natural language text in the nodes of the argumentation graph and a future research direction could be to explicitly model and formalize this comparative relation. However, our approach has the advantage to provide a more articulated and richer kind of evidence through argument graphs, i.e., including outcome unspecific information in the argumentation graph (e.g., limitations of the study where the authors state that their findings need further confirmation), which is crucial in judging the results of a study.

3 Annotation studies and dataset creation

In this section, we present an extension of the dataset of Randomized Controlled Trials annotated with Argument Mining labels we firstly introduced in [16]. The reasons for the augmentation of this dataset are manifold. Firstly, the previous version of the dataset was relatively small and therefore not reliable enough to make robust predictions about the generalizabilty of a model. Secondly, the dataset was annotated with respect to the argument component identification layer only, and it was thus missing a fundamental part of the argument structure, i.e., the relations holding between argumentative components. In addition, the possibility of adding among the main topics a more body-part-unspecific disease, as described in the following section, further offered the opportunity to reuse the previous smaller disease specific subsets as separated test sets and examine the model potential generalizabilty. Finally, after collecting feedback on the dataset from medical domain experts, we decided to incorporate information about the observed outcome in the argument structure. We expect that this additional information makes the argumentative approach to clinical trials more approachable for clinicians, which usually do not have any background in argumentation, but are very familiar with the meaning and use of PICO elements.

In the following, Section 3.1 describes the data collection phase. Section 3.2 describes the three types of annotations carried out on the collected dataset, namely :

- **Argument Components**: Comprising major claims, claims and evidence, where a *major claim* is a general statement about properties of treatments or diseases, a *claim* is a concluding statement, and an *evidence/premise* is an observation or measurement in the study (see subsection 3.2.1).
- **Argumentative Relations**: The relations are connecting argumentative components to form the graph structure of an argument. Components can be either *supporting, attacking* or *partially-attacking* other components (see subsection 3.2.2).

- **Effect-on-Outcome**: It describes the effect an intervention has on each outcome (evaluated parameter) of a study. Effects were annotated when they *improved, increased*, or *decreased*, or when there was no observable difference or an outcome did not occur (see subsection 3.2.3).

Section 3.3 reports on the annotation process and the Inter Annotation Agreement, and discusses cases of disagreement.

3.1 Data collection

As stated in the previous section, we annotate Randomized Controlled Trials (RCTs) to be in line with Evidence-Based Medicine (EBM) guidelines. EBM builds the decision-making on analysing scientific information from systematic reviews of clinical trials. While clinical trials also comprise observational studies, in EBM one opts for randomized controlled trials, which provide more compelling evidence [47] than the observational studies making RCTs the most valuable sources of evidence for the practice of medicine [48]. Albeit there are more factors for this decision, one crucial aspect is the random process of assigning trial participants to at least two comparison groups, which eliminates selection bias. One group receives the intervention under assessment, while the other group, the control group/arm, receives either an established treatment, a placebo or no intervention at all. The intervention efficacy is determined as a comparison with respect to the control group(s). Due to the randomized allocation of participants allowing the use of probability theory, the likelihood that any difference between the groups was by chance can be estimated [49]. The documentation of the study is defined by the CONSORT[3] policies. Due to this comparative nature of the underlying data, for AM, this means that the argumentation is also built mostly on relative statements[4].

We decided to restrict our work to the abstracts of the trials following the argumentation of Trenta et al. [42], such that *"abstracts are the first section readers look at when evaluating a trial"*. Moreover, they are freely accessed, while full text articles may require a paid subscription to unlock.

We rely on and extend our previous dataset AbstRCT [16], the only available dataset of randomized controlled trial abstracts annotated with the different argument components (i.e., evidence, claims and major claims). Such dataset contains the same abstracts used in the dataset of RCT abstracts presented by Trenta et al. [42], that were retrieved directly from PubMed[5] by searching for the disease name and specifying that it has to be a RCT, adopting Strategy 1 in [42]. The first version of the dataset with coarse labels contained 919 argument components (615 evidence and 304 claims) from 159 abstracts comprising 4 different diseases (i.e., *glaucoma*, *hypertension*, *hepatitis b*, *diabetes*). To obtain

[3] `http://www.consort-statement.org/`

[4] In our dataset, about 70% of the annotated argumentative components contain either an explicitly stated comparison or an implicit comparison reported as measured values.

[5] PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`) is a free search engine accessing primarily the MEDLINE database on life sciences and biomedical topics.

more training data, we have extracted from PubMed 500 additional abstracts following the aforementioned strategy. We selected *neoplasm*[6] as a topic, assuming that the abstracts would cover experiments over dysfunctions related to different parts of the human body (providing therefore a good generalization as for training instances).

3.2 Data annotation

The annotation of the dataset was started after a training phase based on the annotation guidelines we defined[7], where amongst others the component and outcome boundaries were topic of discussion. Gold labels were set after a reconciliation phase, during which the annotators tried to reach an agreement. While the number of annotators vary for the three annotation phases (i.e., argumentative component, argumentative relation and effect-on-outcome annotation), the inter-annotator agreement (IAA) was always calculated with three annotators based on a shared subset of the data. The third annotator was participating in each training and reconciliation phase as well.

In the following, we describe the data annotation process for the argument components layer in the neoplasm dataset (i.e., the newly added topic with respect to the preliminary version of AbstRCT [16]), the argumentative relations layer in the whole dataset, and for the effect-on-outcome layer also in the whole dataset.

3.2.1 Argument Components

Following the guidelines for the annotation of argument components in RCT abstracts provided in [16], two annotators with background in computational linguistics[8] carried out the annotation of the 500 abstracts on neoplasm. In the following, example annotations of the abstract or parts of it are shown, where claims are written in bold, major claims are highlighted with a dashed underline, and evidence are written in italics. An illustration of an annotated abstract is shown in Example 3.1.

Example 3.1 *Extracellular adenosine 5'-triphosphate (ATP) is involved in the regulation of a variety of biologic processes, including neurotransmission, muscle contraction, and liver glucose metabolism, via purinergic receptors. [In nonrandomized studies involving patients with different tumor types including non-small-cell lung cancer (NSCLC), ATP infusion appeared to inhibit loss of weight and deterioration of quality of life (QOL) and performance status]. We conducted a randomized clinical trial to evaluate the effects of ATP in patients*

[6] While neoplasms can either be benign or malignant, the vast majority of articles is about malignant neoplasm (i.e., cancer). We stick with *neoplasm* as a term, since this was the MeSH term used for the PubMed query.

[7] Guidelines can be found here: `https://gitlab.com/tomaye/abstrct`

[8] In [50], researchers with different backgrounds (biology, computer science, argumentation pedagogy, and BioNLP) have annotated medical data for an AM task, showing to perform equally well despite their backgrounds.

with advanced NSCLC (stage IIIB or IV). [...] Fifty-eight patients were randomly assigned to receive either 10 intravenous 30-hour ATP infusions, with the infusions given at 2- to 4-week intervals, or no ATP. Outcome parameters were assessed every 4 weeks until 28 weeks. Between-group differences were tested for statistical significance by use of repeated-measures analysis, and reported P values are two-sided. Twenty-eight patients were allocated to receive ATP treatment and 30 received no ATP. [Mean weight changes per 4-week period were -1.0 kg (95% confidence interval [CI]= 1.5 to -0.5) in the control group and 0.2 kg (95% CI =-0.2 to +0.6) in the ATP group (P=.002)]$_1$. [Serum albumin concentration declined by -1.2 g/L (95% CI=-2.0 to -0.4) per 4 weeks in the control group but remained stable (0.0g/L; 95% CI=-0.3 to +0.3) in the ATP group (P =.006)]$_2$. [Elbow flexor muscle strength declined by -5.5% (95% CI=-9.6% to -1.4%) per 4 weeks in the control group but remained stable (0.0%; 95% CI=-1.4% to +1.4%) in the ATP group (P=.01)]$_3$. A similar pattern was observed for knee extensor muscles (P =.02). [The effects of ATP on body weight, muscle strength, and albumin concentration were especially marked in cachectic patients (P=.0002, P=.0001, and P=. 0001, respectively, for ATP versus no ATP)]$_4$. [...] This randomized trial demonstrates that ***[ATP has beneficial effects on weight, muscle strength, and QOL in patients with advanced NSCLC]$_1$.***

Claims In the context of RCT abstracts, a *claim* is a concluding statement made by the author about the outcome of the study. It generally describes the relation of a new treatment (intervention arm) with respect to existing treatments (control arm) and is derived from the described results. An example of comparative conclusions can be seen in the Examples 3.2 and 3.3, where the latter is negated.

Example 3.2 [Trabeculectomy was more effective than viscocanalostomy in lowering IOP in glaucomatous eyes of white patients.]

Example 3.3 [Latanoprost 0.005% is not inferior (i.e., is either more or similarly effective) to timolol and produces clinically relevant IOP reductions across pediatric patients with and without PCG]

Example 3.4 [Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response]

Additionally to the comparative statements, *claims* can also assert general properties, e.g., that an intervention was well tolerated or had beneficial effects with respect to an outcome, like in Examples 3.1 and 3.4. These statements can be in a coordinate structure, which poses the question how to split them. Ideally, the goal is to make an argument component as small and self-contained as possible. For coordinate structures, this means to split them into separated components. For instance, in Example 3.4, this translates to one claim talking about the long-term ocular hypotensive effect and another one about the low rate of allergic response. Dividing the conclusions in these smaller claims makes

the argumentative structure more transparent, because it is clear which assertion an evidence supports. While for a coordination it cannot necessarily be seen at first glance, especially for general outcomes with multiple aspects like *quality of life*. In practice, most of these fine-grained discriminations are prohibited by the syntactic structure of a sentence. Usually conjunctive and disjunctive coordinations are written in an elliptical manner, as it is shown in Example 3.4. The problem with elliptical coordinate structures is that if we divide them into their single conjuncts, these conjuncts are not self-contained: the necessary contextual information, usually the omitted subject, is missing, preventing them to be a stand-alone argument component. This forces the annotators to treat them as one component increasing the complexity of the subsequent relation annotation and classification task.

Major claims *Major claims* are usually defined as a stance of the author in the AM literature. Here, they are defined more as a general/introductory *claim* about properties of treatments or diseases, which is supported by more specific claims. They do not necessarily occur at the end of an abstract as a final conclusion, but are mostly introduced before as a general hypothesis to be tested or as an observation of a previous study to be confirmed. A major claim with the goal of representing an introductory *claim* is shown in Example 3.1. Given the negligible occurrences of major claims in our dataset (only 3% of the components are major claims) and the structural similarity to normal claims, we merge them with the claims for the classification task.

Evidence An *evidence* in RCT abstracts is an observation or measurement in the study, which supports or attacks another argument component, usually a *claim*. Those observations comprise side effects and the measured outcome of the intervention and control arm. They are observed facts, and therefore credible without further justifications, as this is the ground truth the argumentation is based on. *Evidence* can either state exact measurements, see for instance evidence 1-3 in Example 3.1, or explicitly expressed comparisons, as shown in Examples 3.5, 3.6 and 3.8. A common part in medical argumentation are outcomes which were not observed. For clinical decision making not only the observed change in outcomes play an important role, but also the absence of, for example, a side-effect. Section 3.2.3 elaborates more on this matter. Since these observations of absence are important, we consider them as *evidence* in the argumentation, as illustrated in Example 3.7.

Example 3.5 [*Headache, fatigue, and drowsiness were similar in the 2 groups.*]

Example 3.6 [*Pulse rate was significantly reduced with timolol, but not with latanoprost.*]

Example 3.7 [*No evidence of tachyphylaxis was seen in either group.*]

Example 3.8 [*Dry mouth was more common in the brimonidine-treated group than in the timolol-treated group (33.0% vs 19.4%)*]$_1$, [*but complaints of burning*

and stinging were more common in the timolol-treated group (41.9%) than in the brimonidine-treated patients (28.1%)]$_2$.

Example 3.9 [*Mean (+/-SD) preoperative and 1-year postoperative intraocular pressures in the 5-fluorouracil group were 26.9 (+/-9.5) and 15.3 (+/-5.8)mm Hg, respectively. In the control group these were 25.9 (+/-8.1)mm Hg, and 15.8 (+/-5.1) mm Hg, respectively*]

Similarly to the aforementioned *claims*, *evidence* are often stated as conjunctive coordinations and it is important that multiple observed measures are annotated as multiple pieces of the same *evidence*. Again, the problem of how to divide them into separated self-contained units arises. In Example 3.5, the syntax does not allow splitting the conjunction and therefore the sentence as a whole is annotated as one single *evidence*. Exceptions can be adversative coordinations (e.g., *but, except for*). While they are usually also elliptical (see for instance Example 3.6), in some cases they are not and can be seen as a separated *evidence*, as illustrated in Example 3.8. Here, evidence 2 is self-contained and can be processed without evidence 1. In rare cases, *evidence* can span multiple sentences, like in Example 3.9. As stated before, the efficacy of an intervention in a RCT is measured as a comparison to the control group. In Example 3.9, each sentence on its own misses the relevant information to make the comparison from the other group. In terms of argumentation, this is a linked argument structure, where multiple premises require each other to support a conclusion. Given the interdependence of the premises in such a structure, we decided to annotate it as one component.

3.2.2 Argumentative Relations

In order to identify complex argumentative structures in the data, it is crucial to annotate the relations, i.e., directed links connecting the components. Those relations are connecting argument components to form the argumentation graphs representing the structure of an argument. Existing approaches in AM try to form a tree structure with one root node [22]. Our approach is more data driven, and we assume that a trial abstract contains at least one argument in form of a tree, where an argument consists of at least one *claim* which is supported by at least one *evidence*. In practice, the average clinical trial in our dataset has between one and two trees, depending on the number and topic of the claims and major claims. In general, the annotated arguments are convergent[9] or a combination of convergent and sequential[10] arguments [51]. Removing one *evidence* does not weaken the other. Given that *claims* often have a coordinate structure or make general statements, i.e., that an intervention was well tolerated, there are various independent pieces of *evidence* linked to a single *claim* making most of the arguments in our data convergent. In our data, sequential arguments

[9] A *convergent* argument consists of a *claim*, which is supported by independent *premises/evidence* [51].

[10] *Sequential* arguments consists of at least two *premises/evidence*, where one supports the other, which is supporting the final *claim*.

can be seen mostly in combination with two supporting claims or major claims. There, one *claim* supported by *evidence* supports or attacks another (major) *claim*. In 19% of the cases, *claims* are linked to other (major) *claims*.

Generally speaking, an argumentative relation is a directed link from an outgoing node (i.e., the *source*) to a target node. The nature of the relation can be supporting or attacking, meaning that the source argumentative component is justifying or undermining the target argumentative component. Links can occur only between certain components: evidence can be connected to either a claim (in 92% of the cases) or another evidence (in 8% of the cases), whereas claims can only point to other claims (including major claims). The polarity of the relation (supporting or attacking) does not limit the possibility to what type of component a component can be connected. Theoretically, all types of relations are possible between the allowed combination pairs. Practically, some relations occur rather seldom compared to the frequency of others. For example, in 78% of the cases when an *evidence* is linked to another *evidence* it is an attack or a partial-attack. As stated previously, in rare cases, components can be unconnected. Additionally to the aforementioned occurrence, this can happen for *major claims* in the beginning of an abstract, whose function is to point out a general problem, unconnected to the outcome of the study itself.

As shown in Example 3.1, argument components can contain negations. For many text mining tasks negation detection and scope resolution are important subtasks, because negations entirely change the meaning of a sentence. Especially in the biomedical domain, the use of negative assertions (in particular, negating negative phrases, like *not inferior*) is abundant [52]. This poses further challenges for the automatic processing of this kind of text. In the case of AM, negations do also play an important role. Here, the impact is related rather to the correct classification of the relation than the correct linking of the components. Failing to correctly detect a negation can culminate in assigning the wrong polarity label, i.e., *attack* instead of *support*. Again, posing a great challenge for the relation classification part of the AM pipeline on clinical trials.

Attack A component is attacking another one, if it is *i)* contradicting the proposition of the target component, or *ii)* undercutting its implicit assumption of significance, e.g., stating that the observed effects are not statistically significant. The latter case is shown in Example 3.10. Here, Evidence 1 is attacked by Evidence 2, challenging the generality of the prior observation.

Example 3.10 *[True acupuncture was associated with 0.8 fewer hot flashes per day than sham at 6 weeks,]*$_1$ $\xleftarrow[Attack]{}$ *[but the difference did not reach statistical significance (95% CI, -0.7 to 2.4; P = .3).]*$_2$

We further make the assumption that when the trial reports allergic reactions or other adverse effects, the author as a domain expert knows if these observations are disproportional or acceptable. So, when an intervention is claimed to be well tolerated, the *evidence* reporting these effects is considered

as supporting unless the opposite is clearly stated, e.g., in form of *severe* or other modifiers.
The *partial-attack* is used when the source component is not in full contradiction, but weakening the target component by constraining its proposition. Those can be implicit statements about the significance of the study outcome, which usually occur between two claims, as in Example 3.11. Attacks and partial-attacks are identified with a unique class for the relation classification task, because these relations are underrepresented in the dataset. In the training set only 2,5% are attack and 12% are partial-attack relations.

Example 3.11 [**SLN biopsy is an effective and well-tolerated procedure.**]$_1$ $\cdot\overleftarrow{_{Partial-attack}}$ [**However, its safety should be confirmed by the results of larger randomized trials and meta-analyses.**]$_2$

Support Contrary to the attack relations, the support relation is not further subdivided. While an *evidence* usually provides support for a certain aspect of the more general *claim*, it would have been often ambiguous to distinguish between partially and fully support relations, especially with respect to the impact of observed adverse effects. Thus, all statements or observations justifying the proposition of the target component are considered as supporting the target (even if they justify only parts of the target component). In Example 3.1, all the evidence support Claim 1.

We carried out the annotation of argumentative relations over the whole dataset of RCT abstracts, including both the first version of the dataset [16] and the newly collected abstracts on neoplasm.

3.2.3 Effect-on-Outcome

Argumentative structure annotations alone are for most domain specific AM use cases sufficient. In the case of EBM, where one wants to facilitate the analysis process of trials by clinicians, further medical annotations can be beneficial. For this reason, we decided to annotate the effect an intervention has on an Outcome (one of the PICO elements), e.g., if the outcome was *increased*, *decreased* or was not affected. Contrary to Lehman et al. [45], which also use these three labels[11], we added two extra labels, which we consider essential to fully cover the reports about an outcome. These labels are (i) the *NoOccurrence* label, when an outcome, e.g., a side effect, did not occur, and (ii) the *Improved* label for cases in which it is not clear from the text if the beneficial effect is due to a decrease or increase in the measured value of the outcome. We consider the addition of the *NoOccurrence* label important for medical argumentation, even though these reports are less frequent. For decision-making, it is not only relevant which effects were observed, but also which (side-)effects did not occur.

[11] We dropped the *significantly* from the labels, because even though we made an implicit assumption of significance earlier, we do not know beforehand how many of the outcomes are significant, since the model cannot take components undercutting this assumption into account.

Class	#outcomes	%
Improved	831	25
Increased	765	23
Decreased	782	23
NoDifference	897	27
NoOccurrence	76	2
Total	3351	100

Table 1: Statistics of the outcome dataset. Showing the numbers of Improved, Increased, Decreased, NoDifference and NoOccurrence classes independent of the disease-based subsets.

Note that we decided to not annotate our data with the other PICO elements. Firstly, because argumentative components contain information about the trial population only in roughly 1-2% of the cases. And secondly, there exists already a larger corpus specialised on PICO annotations [41]. Before we started annotating the Effect-on-Outcome, we assessed whether the argumentative components contain enough description of those effects to have a comprehensive coverage in our dataset. Theoretically, following the CONSORT statement [53] authors should report all PICO elements in the abstract. We found that claims contain approximately in 72% of the cases at least one PICO element (P: 2%, I/C: 51%, O: 47%) and evidence contain it approximately in 87% (P: 1%, I/C: 27%, O: 72%) of the cases. For our annotation, we consider explicit mentions of effects on an outcome. From our 4198 argument components, 2195 fulfilled this criteria. The others report either only the measured numerical values of outcomes (704) making the effect implicit, or general statements without an indication of a trend, e.g., that some side effect was *mild* or *common*. Moreover, many components, especially claims, give conclusive statements, e.g., that a treatment is *safe* or *efficient*, without listing the specific outcomes. Note that the annotation (and later the classification) is even more complex as about 50% of the effect-on-outcome containing argument components report either the outcome or the intervention in an abbreviated form. This trend is similar to the distribution of abbreviations in all argument components, where about 45% contain an abbreviation of either the intervention, or outcome or both. The detailed annotation statistics are reported in Table 1.

Increased/Decreased These labels are used when it is stated that the outcome was higher, like in Example 3.12, or lower after an intervention, like in Examples 3.12 and 3.14. Generally, it should not contain a sentiment, like *better score*. In rare cases, where an outcome was reported as *worse*, annotation guidelines were set to infer the value, e.g., a worsened side-effect usually means an increased/more intense and not a decrease occurrence. There were only a handful of cases were this was not achievable without fundamental medical expertise. These examples have been discarded.

NoDifference An effect on an outcome is labeled as *NoDifference*, when there was no change in the outcome or when the two treatments resulted in similar values, i.e., there was no difference in the outcome between the two treatment arms. The latter case is shown in Example 3.12, where the *response rates* of both interventions are similar.

Example 3.12 *Raltitrexed showed similar [response rates]$_{NoDifference}$ to the de Gramont regimen, but resulted in greater [toxicity]$_{Increased}$ and inferior [quality of life]$_{Decreased}$.*

NoOccurrence This label is used when an outcome, usually an adverse effect, was not observed, as shown in Example 3.13. Moreover, this example illustrates the division of coordinate structures in a single component. Contrary to argument components, the problem with ellipses preventing the division is lower, because the annotation units are smaller.

Example 3.13 *No cases of drug-related [neutropenic fever]$_{NoOccurrence}$, [sepsis]$_{NoOccurrence}$, or [death]$_{NoOccurrence}$ occurred.*

Improved This label is used when the described outcome explicitly had a beneficial effect and no information if the measured value increased or decreased is provided, like in Example 3.14. There, two problems come together. First, *bleb morphology*, like *quality of life*, is a general term comprising various subscales, for instance, *bleb wall reflectivity*, *visibility of drainage route* or *presence of hyper-reflectivity area*. Second, the effect description *better* does not allow any conclusions about the measured values without concrete expert knowledge about which subscale should be increased or decreased to result in a better bleb morphology. Thus, the only certain information, which can be drawn from this statement, is that the bleb morphology improved.

Example 3.14 *Ologen resulted in a lower long-term [postoperative IOP]$_{Decreased}$, a better [bleb morphology]$_{Improved}$, and fewer [complications]$_{Decreased}$.*

3.3 Inter-Annotator Agreement

In total for all tasks, three annotators were participating in the annotation process. During the training phase the guidelines were refined in multiple rounds of discussion between all annotators. After the training phase, where the annotators made themselves familiar with the tasks and the data, in order to validate the annotations, the inter-rater reliability or inter-annotator agreement (IAA) was calculated on a reserved and previously unseen subset of the data. The subset was sampled randomly from the collected data and each rater annotated the data independently. While the subsequent full annotation of each sub-task was not always conducted with all three annotators, the corresponding IAA subset was always annotated by all three annotators and the agreement was calculated respectively.

As the statistical measure for assessing the reliability of the annotations, we used Fleiss' kappa [54], a generalization of Scott's pi. It is suitable for a finite nominal-scale and contrary to the latter, it can be used for more than two raters. Another plausible measure would have been Krippendorff's alpha. While it is more flexible and allows other scales and missing data, our data is purely nominal and complete. Furthermore, having a highly imbalanced dataset could lead to instances being correctly classified by chance. Both measures control this providing a more reliable agreement score. While Krippendorff's alpha is based on the observed disagreement corrected for disagreement expected by chance, Fleiss' kappa considers the observed agreement corrected for the agreement expected by chance [55]. In the case of complete nominal data[12], both measures are similar in representing the reliability [55, 56].

Argument Components For this task, the IAA was calculated for token-level annotation. This way not only the label mismatch between *claim* and *evidence* is considered, but also the disagreement in boundary annotation. IAA among the annotators has been calculated on 30 abstracts, resulting in a Fleiss' kappa of 0.72 for argumentative components and 0.68 for the more fine-grained distinction between claims and evidence. Both values are higher than 0.61 meaning substantial agreement for both tasks [57].

Argumentative Relations Contrary to the other tasks reported in this paper, here, the IAA was calculated not on token-level but considering each argument component as a unit. Annotation was considered as agreed, when both, the relation label and the assigned target component, were the same. IAA has been calculated on the same 30 abstracts annotated in parallel by three annotators (the same two annotators that carried out the argument component annotation, plus one additional annotator). The resulting Fleiss' kappa was 0.62, meaning substantial agreement.

Effect-on-Outcome Similarly to the argument component annotation, the agreement was calculated on token-level. Since the Effect-on-Outcome descriptions occur only on a subset of the argument components, we increased the number of abstracts included in the IAA calculation to 47. This resulted in a Fleiss' kappa of 0.81, which means almost perfect agreement [57].

3.3.1 Disagreement

In the following, we discuss the observed disagreement between the annotators and the associated difficulties, which were examined in the reconciliation phase.

For the argument component annotation, raters disagreed on the exact determination of the boundaries. For example, conjunctive adverbs like *however*

[12]In our dataset, all N observations are assessed by all n raters, which makes our IAA subset complete per definitionem.

or *in general* can play an important role. In Example 3.15, *in general* is an important modifier which should be included in the component. Also, for phrases like *this suggests*, it can be argued that they are an important part of the argument component, because they underline the conclusive function of a claim and therefore serve as potential discriminators, in particular for cases where it is not directly clear if the statement is an observed outcome or a drawn conclusion. This is mostly the case when no exact measurement or p-value is stated, as in Example 3.16.

Example 3.15 In general, the tolerance to medication was acceptable.

Example 3.16 Latanoprost provided greater mean IOP reduction than did Brimonidine.

Further common disagreement was observed between claims and major claims, which can be very similar in their function as a (general) summary or conclusion. This strengthened us in the decision to merge these two labels later in the classification. Another common conflict was the annotation of too general or co-referring components, which would not be self-contained after removing the context.

Concerning the relation annotation, most of the disagreement was not in annotating the relation label, but in assigning the target component, with an exception for the attack and partial-attack labels. As for the claims and major claims, this further endorsed the label merge for classification. Linking components lead to conflict in cases where multiple claims were very similar. One could either see a sequential structure if one considers one of the claims less specific, or two separated claims, which share parts of their evidence. In the reconciliation phase, we decided against the latter option to avoid this kind of divergent argument structures.

For the effect-on-outcome annotation, one of the main disagreements between the annotators was regarding how to annotate enumerations separated by a backslash (e.g., anthralogia/myalgia), whether to annotate both as one outcome or annotate them as separated entities. It was decided to label them separately. Similar to this, the coordination of outcomes (e.g., mood, QOL or healthcare utilization) were also labeled like that, unless the separation implicates losing information related to the outcomes (e.g., liver and cardiac toxicities).

Another topic of discussion was about the inclusion of extra information relevant to the outcome or not, i.e., setting the exact boundaries. This led to further discussion on what is considered relevant information. In the end, it was decided to only include the tokens that directly affect the semantics of the outcome (e.g., *overall* QoL, *global* QoL *scores*, emphirreversible toxicity). The tokens left apart were those that do not change the semantic of such (e.g., *severity of other* toxicities, *rating of* cosmetic results, *quality adjusted* survival time). A full sentence is provided in Example 3.17.

Example 3.17 Ratings of [cosmetic results]$_{\text{Decreased}}$ decreased with time, in line with clinical observations of long-term side-effects of radiotherapy.

Dataset	#Evi	#Claim	#MajCl	#Sup	#Att
Neoplasm	2193	993	93	1763	298
Glaucoma	404	183	7	334	33
Hepatitis	80	27	5	65	1
Diabetes	72	36	11	44	8
Hypertension	59	26	9	53	2
Total	2808	1265	125	2259	342

Table 2: Statistics of the extended dataset. Showing the numbers of evidence, claims, major claims, supporting and attacking relations for each disease-based subset, respectively.

As previously discussed, in the dataset we have a few sentences that present two different polarities at the same time, for instance Example 3.18. Most of them are a comparison between the intervention and the control group where the outcome has different results for each. This was the main disagreement between the annotators, whether to annotate the outcome twice with each different result or to follow one of the group results. Ultimately, it was decided to always follow the intervention group results.

Example 3.18 Men in the control group had significant increases in [fatigue scores]$_{\text{NoDifference}}$ from baseline to the end of radiotherapy (P=0.013), with no significant increases observed in the exercise group (P=0.203).

Accordingly, with respect to Example 3.18, *control group* qualifies as the baseline and *exercise group* as intervention, meaning that the outcome *fatigue scores* is annotated as *NoDifference*. These cases pose additional challenges to the effect classifier.

3.4 Dataset Statistics

To summarize, Table 2 reports on the statistics of the argumentative component and relation annotation, and Table 1 on the Effect-on-Outcome annotations of the final AbstRCT dataset.

Concerning the argumentative annotations, there are about as half as many claims as evidence for every data split. While the average rate of evidence to claim is 2.2, the average claim has 1.87 components pointing at it. The difference is due to unconnected pieces of evidence and pieces of evidence pointing at other pieces of evidence, which are in total 22% of all snippets annotated as evidence. Major claims and attack relations are not as balanced in their distribution over the various data splits, mostly because of their rare occurrence in general. As previously stated, the average trial contains one to two argument graphs in form of trees, with the highest average of 1.98 arguments on the neoplasm subset and the lowest with 1.3 on the hypertension subset.

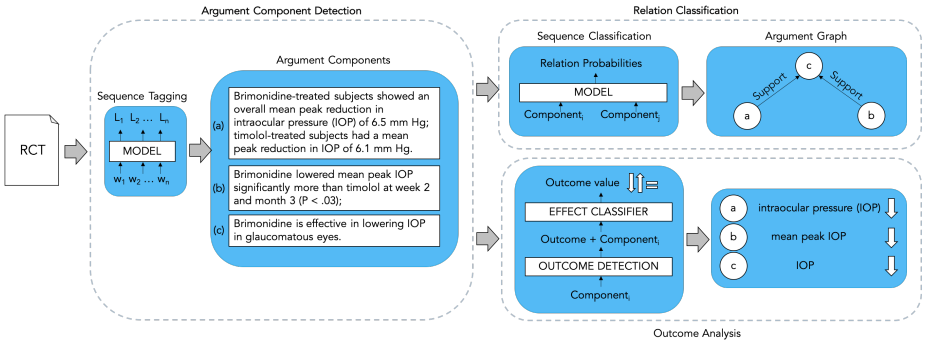


Figure 1: The full argument mining and outcome analysis pipeline on clinical trials.

4 AM and Outcome Analysis for Evidence-based Medicine

In this section, we describe the full argument mining and outcome analysis pipeline we defined, as visualized in Figure 1. More precisely, we first present our approach for the tasks of argument component detection (Section 4.1) and argument relation prediction (Section 4.2), and second, we define how to tackle the task of outcome detection and classification (Section 4.3).

4.1 Argument Component Detection

The first step of the AM pipeline (visualized in Figure 1) is the detection of argumentative components and their boundaries. As described above, most of the AM approaches classify the type of component assuming the boundaries of argument components as given. To merge the component classification and boundary detection into one problem, we cast the component detection as a sequence tagging task. Following the BIO-tagging scheme, each token should be labeled as either being at the **B**eginning, **I**nside or **O**utside of a component. As we have two component types (i.e., evidence and claims), this translates into a sequence tagging problem with five labels, i.e., *B-Claim*, *I-Claim*, *B-Evidence*, *I-Evidence* and *Outside*. To model the temporal dynamics of sequence tagging problems, usually Recurrent Neural Networks (RNN) are used. In our experiments, we evaluate different combinations of RNNs with various types of pre-trained word representations. Each embedding method is combined with uni- or bidirectional Long Short Term Memory (LSTMs) or Gated Recurrent Units (GRUs) networks with and without a Conditional Random Fields (CRF) as a last layer. The tested word embeddings range from **GloVe** [58] over **fastText** [59] to contextualized embeddings, such as **ELMo** [60], Contextualized String Embeddings (**Flair**) [61] or **BERT**) [38]. For a detailed overview of all tested embeddings, we refer the reader to [18]. Additionally to these embed-

dings, we are the first to do token level classification on AM by fine-tuning different transformer models. To this end, a shallow layer for sequence tagging is put on top of the transformer architecture. The shallow layer can be either a simple dense layer or one of the RNN CRF combinations for sequence modelling described above. Besides the original BERT, which is pre-trained on the BooksCorpus and English Wikipedia, we experiment with **BioBERT** [62], which is pre-trained on large-scale biomedical corpora outperforming the general BERT model in representative biomedical text mining tasks. The authors initialize the weights with the original BERT model and train on PubMed abstracts and full articles. Therefore, the vocabulary is the same as for the original BERT. Contrary to that, **SciBERT** [63] is trained from scratch with an own vocabulary. While SciBERT is trained on full papers from Semantic Scholar it also contains biomedical data, but to a smaller degree than BioBERT. We chose to use the uncased SciBERT model, meaning that we ignore the capitalization of words. As it was the case for the original BERT, the uncased model of SciBERT performs slightly better for sentence classification tasks than the cased model.

4.2 Relation Classification

After the argument component detection, the next step is to determine which relation holds between the different components (Figure 1). We extract valid **BI** tag sequences from the previous step, which are then considered to be the argumentative components of one RCT. Those sequences are phrases and do not necessarily correspond to full sentences. The list of components then serves as input for the relation classification.

As explained in Section 2, the relation classification task can be tackled with different approaches. We treat it as a sequence classification problem, where the sequence consists of a pair of two components, and the task is to learn the relation between them. For this purpose, we use self-attending transformers, since these models are dominating the benchmarks for tasks which involve classifying the status between two sentences [38]. Treating it as a sequence classification problem gives us two options to model it: *(i)* jointly modelling the relations by classifying all possible argumentative component combinations or *(ii)* predicting possible link candidates for each entity and then classifying the relation only for plausible entity pairs. In the literature, both methods are represented. Therefore, we decided to evaluate both ways of solving the problem. We experiment with various transformer architectures and compare them with state-of-the-art AM models, i.e., the Tree-LSTM based end-to-end system from Miwa and Bansal [30] as employed by Eger et al. [29], and the multi-objective residual network of Galassi et al. [35]. For option *(i)*, we use bi-directional transformers [38], which consist of an encoder and decoder which themselves consist of a multi-head self-attention layer each followed by a fully-connected dense layer. Contrary to the sequence tagging transformer, where each token of the sequence has a representation which is fed into the RNN, for sequence classification a pooled representation of the whole sequence is needed. This rep-

resentation is passed into a linear layer with a softmax which decodes it into a distribution over the target classes. We treat it as a three class classification problem (*Support*, *Attack* and *NoRelation*), where all possible component combinations of one trial are classified to determine the relations among them. To counter the class imbalance caused by this problem formulation, we also evaluate the effect of exchanging the cross entropy loss with weighted cross entropy loss to give more importance to the underrepresented classes during training. We refer to this type of transformer as **SentClf**. Using this architecture, one component can have relations with multiple other components, since each component combination is classified independently. This is not the case in a multiple choice setting (**MultiChoice**), where possible links are predicted taking the other combinations into account and which we employ for *(ii)*. Here, each component (source) is given the list of all the other components as possible target relation candidates and the goal is to determine the most probable candidate as a target component from this list. This problem definition corresponds to the grounded common sense inference problem [64]. To model components which have no outgoing link to other components, we add the *noLink* option to the choice selection. As an encoder for phrase pairs, we evaluate various BERT models as detailed above, just as we do for the SentClf task. With respect to the neural transformer architecture, a multiple choice setting means that each choice is represented by a vector $C_i \in \mathbb{R}^H$, where H is the hidden size of the output of an encoder. The trainable weight is a vector $V \in \mathbb{R}^H$ whose dot product with the choice vector C_i is the score of the choice. The probability distribution over all possible choices is given by the softmax, where n is the number of choices:

$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^{n} e^{V \cdot C_j}} \tag{1}$$

The component combination with the highest score of having a link between them is then passed into a linear layer to determine which kind of relation is holding between the two components, i.e., *Attack* or *Support*. The MultiChoice model is trained jointly with two losses, i.e., one for the multiple choice task and one for for the relation classification task. Similar to the experiments on sequence tagging, **BERT**, **SciBERT** and **BioBERT** are used. Furthermore, **RoBERTa** [65] is employed, another new model, which outperforms BERT on the General Language Understanding Evaluation (GLUE) benchmark. There, the BERT pre-training procedure is modified by exchanging static with dynamic masking, using larger byte-pair encoding and batches size, and increasing the size of the dataset.

Complementary to the results of the isolated relation classification on gold labels, we report the performance of the whole pipeline, which includes the component detection as a prior step. Here, we follow existing work on end-to-end Argument Mining systems [66, 29] to count true/false positives and false negatives for relation/component combinations to calculate the overall performance. In particular, the overlap percentage of tokens is used to determine the base if a predicted component matches the annotated component in the gold standard.

Similar to the aforementioned work, we report the results for a threshold of 50% and 100% of matched tokens. However, for determining if a gold component was detected, we ignore the difference between the argumentative labels (evidence and claim) and consider them as one class, since the discrimination between them is not relevant for our relation classification approach.

4.3 Outcome Detection and Classification

In Evidence-Based Medicine, PICO elements play an important role. However, to the best of our knowledge, we are not aware of any approach in EBM combining argumentative and outcome analysis to support clinicians in their investigation over clinical trials. With the goal of taking the automatic analysis of clinical trials a step further in this direction, in this paper we combine an analysis of the Effect-on-Outcome with an AM model, enriching the arguments with valuable medical information and leverage this way the advantages of both domains.

The outcome analysis is a pipeline composed of two major parts. First, the outcome detection, which finds and extracts the outcomes of an argumentative component, and second, the effect classifier, which predicts which consequence was seen for each outcome after an intervention. Similar to the argument component detection, we treat the outcome detection as a sequence tagging task with the BIO-tagging scheme and we employ the same transformer architecture for sequence tagging. From the prediction results, valid **BI**-sequences are extracted, which are considered to be the outcomes reported in a component. Each outcome together with the component it occurred in is provided as input into the effect classifier. Given this bipartite input, the problem is similar to the aforementioned relation classification and thus we treat effect classification the same way, namely as a sequence classification task. Contrary to the three class relation classification, in this case it is a five class (*Improved, Increased, Decreased, NoDifference, NoOccurrence*) classification problem. Experiments are conducted with the same pre-trained transformer model types as for relation classification, i.e., **BERT, BioBERT** and **SciBERT** (cased and uncased), with the exception of RoBERTa. For both parts of the pipeline, i.e., the outcome detection and effect classifier, the same type of transformer is employed. Similar to the relation classification, the isolated performance on gold standard and the overall performance with the prior component detection are reported. As for the evaluation of the overall performance of the argument mining part of the pipeline, the best performing sequence tagging model on the gold standard was selected, i.e., SciBERT in a combination with BiGRU and CRF, and the results reported for the 50% and 100% threshold of the component detection.

5 Experiments

This section presents experiments conducted on the annotated dataset of clinical trials introduced in Section 3. We first present the experimental setup of the

two tasks composing our pipeline, namely argumentative components detection and relation prediction (Section 5.1) and outcome detection and classification (Section 5.2). Secondly, we report and discuss on the obtained results, providing an in-depth error analysis (Section 5.3).

5.1 Argumentative components detection and relation prediction

For sequence tagging, each of the embeddings were combined with either *(i)* a GRU, *(ii)* a GRU with a CRF, *(iii)* a LSTM, or *(iv)* a LSTM with a CRF. For BERT, we use the PyTorch implementation of huggingface[13] version 2.3. For fine-tuning the BERT model, we used the uncased base model with 12 transformer blocks, a hidden size of 768, 12 attention heads, a learning rate of 2e-5 with Adam optimizer for 3 epochs. The same configuration was used for fine-tuning Sci- and BioBERT. For SciBERT, we used the uncased model with the SciBERT vocabulary. For BioBERT, we used version 1.1. For RoBERTa, we increased the number of epochs for fine-tuning to 10, as it was done in the original paper. The best learning rate was 2e-5 on our task. The number of choices for the multiple choice model was 6. Batch size was 8 with a maximum sequence length of 256 subword tokens per input example. The weight factor for each class in the weighted cross entropy loss for the SentClf transformer is the normalized number of training samples of this class[14]. To calculate the overall performance of our pipeline, we used the best performing component detection model, i.e., the SciBERT uncased with a BiGRU and CRF. We split our neoplasm corpus such that 350 abstracts are assigned to the train, 50 to the development, and 100 to the test set. Additionally, we use the first version of the dataset [16] to create two extra test sets, both comprising 100 abstracts. The first one includes only glaucoma, whereas the second is a mixed set with 20 abstracts of each disease in the dataset (i.e., neoplasm, glaucoma, hypertension, hepatitis and diabetes), respectively.

5.2 Outcome detection and classification

For the sequence tagging architecture, we experimented with the GRU in combination with a CRF, because it provided slightly better results than the LSTM for the argument component detection. The outcome pipeline implementation was done with the same Python, PyTorch and transformer versions as the previous experiments. Both transformer models of the pipeline are of the same type and initialised with the same pre-trained weights. The effect-of-outcome annotations are converted into two datasets, one for each part of the pipeline. The first one in a CoNLL format for token-wise labels, and the second one in csv format, where each outcome-component pair is listed. This results in multiple entries, if a component contains more than one outcome. The fine-tuning of the

[13] `https://github.com/huggingface/transformers`

[14] The training set consists of 90% *NoRelation*, 8.5% *Support* and 1.5% *Attack* samples.

	Neoplasm			Glaucoma			Mixed		
Embedding	F1	C-F1	E-F1	F1	C-F1	E-F1	F1	C-F1	E-F1
GloVe	.58	.50	.66	.52	.36	.68	.50	.36	.64
fastText(fT)	.66	.61	.71	.65	.60	.71	.60	.52	.69
ELMo	.68	.59	.76	.72	.67	.77	.70	.67	.74
FlairPM	.68	.60	.75	.72	.69	.75	.68	.64	.72
fine-tuning BERT	.85	.78	.90	.86	.76	.89	.88	.81	.91
fine-tuning BioBERT	.84	.87	.90	**.91**	**.93**	**.91**	**.91**	**.91**	.92
fine-tuning SciBERT	**.87**	**.88**	**.92**	.89	**.93**	**.91**	.88	.90	**.93**

Table 3: Results of the multi-class sequence tagging task are given in macro F1. The binary F1 for claims are reported as C-F1 and for evidence as E-F1. Best scores in each column are marked in bold.

models is done separately, each task on its own dataset version, with the same configuration of hyper-parameters as for the argument component detection. Token-wise evaluation is done on the full pipeline output, which is reconverted to the CoNLL format to compare against the gold labels, taking the propagated error from the first pipeline part into account. We split our annotated dataset into a train and test set (80% and 20%, respectively) respecting the class distribution of the overall dataset in both subsets. This way we break with our previous methodological choice of having one in-domain test set and two out of domain sets, as it was the case for the evaluation of the AM pipeline. Given the size of our dataset and the fact that our annotations are imbalanced with respect to certain classes (see Section 3), it is not feasible to maintain these three test sets and ensure at the same time that they have the same size, as done for experiments on the AM pipeline (see Section 5.1), i.e., 100 abstracts each. Whilst it would be indeed interesting to see the effects the comparison of three different test sets offers, test sets with different sizes do not allow for a fair comparison.

5.3 Results

The following sections present and discuss the empirical results of our experiments on both modules of our pipeline. More precisely, the evaluation of our AM module for RCTs is reported in Section 5.3.1 and the outcome analysis module in Section 5.3.2. An in-depth error analysis completes each section.

5.3.1 Argument Mining Pipeline

Sequence Tagging We show the results for a selection of sequence tagging models in Table 3. For a more detailed report, we refer the reader to [18]. The differences of the various shallow layers, which are required to make BERT suitable for sequence tagging, are shown exemplary in Table 4 for the uncased BERT base model. Results calculated on token level are given on all three test sets in macro multi-class F1-score and for claim and evidence, respectively.

	Neoplasm			Glaucoma			Mixed		
	F1	C-F1	E-F1	F1	C-F1	E-F1	F1	C-F1	E-F1
dense layer	.60	.69	.83	.55	.63	.80	.57	.65	.83
CRF	.84	**.78**	**.90**	.85	.81	**.89**	.85	.79	.90
GRU+CRF	.84	**.78**	**.90**	.80	.81	.87	.81	.78	.90
LSTM+CRF	.65	.73	.89	.63	.78	.86	.64	.76	.88
BiGRU+CRF	**.85**	**.78**	**.90**	**.89**	.76	**.89**	**.88**	**.81**	**.91**
BiLSTM+CRF	.80	.77	.89	.81	**.82**	.88	.81	.79	.90

Table 4: Comparison of various architectures for the shallow layer extension of BERT for the sequence tagging task. Results are given in macro F1-score. The binary F1 for claims are reported as C-F1 and for evidence as E-F1.

Generally, evidence scores are higher than claim scores, leading to the conclusion that claims are more diverse than evidence. The explanation is that, since natural language reports of measurements in clinical trials vary mostly only in the measured parameter and its values, claims can be made about almost everything. Another observation is that the performance of the models trained on neoplasm data do not significantly decrease for test sets on other disease treatments. This fact supports our choice of a more general high level disease type like neoplasm for training the models. The performance for many model combinations even increases on the glaucoma test set. The glaucoma test set comprises only a handful of different glaucoma treatments and is therefore less diversified than the neoplasm or mixed test sets. This is ideal with respect to the application of such models, where clinicians will compare studies for a specific disease treatment. Looking at the main difference in the results, fine-tuning transformers shows a significant improvement to other models, where SciBERT with .87 F1-score is the best performing one.

Taking a look at the various options for the sequence modelling shallow layer on top of the transformer in Table 4, the most notable difference is achieved by adding a CRF. A CRF forces the model to consider all labels of a sequence instead of making an independent prediction for each token. Interestingly, adding a uni-directional GRU or LSTM between the transformer and the CRF does not increase the overall results. Replacing the uni-directional with a bi-directional RNN increases the performance only slightly with respect to having no RNN at all. Interpreting the results, this means that the transformer part actually captures the necessary information for the classification task, while the sequence modelling of the RNN becomes redundant. The only marginal increase of the bi-directional GRU is most likely more due to the increase in trainable network parameters than the actual recurrent architecture. In a direct comparison between GRU and LSTM, both RNN types deliver results in a comparable range, where the GRU does seem to show more reliable results. For example, the .65 macro F1-score on the neoplasm test set for the uni-directional LSTM is due to the complete failure of correctly detecting B-Claim tokens, which the GRU counterpart does not struggle with. Similar observations were found for the bi-directional variants. Here, the BiLSTM misclassifies B-tokens as I-tokens of

Method	Neoplasm	Glaucoma	Mixed
Tree-LSTM	.37	.44	.39
Residual network	.42	.38	.43
BERT MultiChoice	.58	.56	.55
BioBERT MultiChoice	.61	.58	.57
SciBERT MultiChoice	.63	.59	.60
BERT SentClf	.62	.53	.66
BioBERT SentClf	.64	.58	.61
SciBERT SentClf	**.68**	.62	.69
SciBERT SentClf (WeightedCrossEntropyLoss)	**.68**	**.70**	**.70**
RoBERTa	.67	.66	.67
RoBERTa (WeightedCrossEntropyLoss)	**.68**	.67	.67

Table 5: Results of the relation classification task, given in macro F1-score.

the correct component type, which translates into a lower macro F1-score.

Relation Classification The results for relation classification are shown in Table 5. Results are given on all three test sets in macro multi-class F1-score.

The Tree-LSTM based end-to-end system [29] performed the worst with a F1-score of .37. This can be explained by the positional encoding in the persuasive essay dataset being more relevant than in ours. There, components are likely to link to a neighboring component, whereas in our dataset the position of a component only partially plays a role, and therefore the distance in the dependency tree is not a meaningful feature. Furthermore, the authors specify that their system does not scale with increasing text length. Especially detailed reports of measurements can make RCT abstracts quite long, such that this system becomes not applicable for this type of data.

The residual network [35] performed better with a F1-score of .42. The main problem here is that it learns a multi-objective for link prediction, relation classification and type classification for source and target component, where the latter classification step is already covered by the sequence tagger and therefore unnecessary at this step.

Similar to sequence tagging, one can see a notable increase in performance when applying a BERT model. Comparing the specialized and general BERT model, the Bio- and SciBERT increase the performance by up to .06 F1-score. Interestingly, RoBERTa delivers comparable results even though it is a model trained on general data. We speculate that parts of the web crawl data which was used to train RoBERTa contain PubMed articles, since they are freely available on the web. Looking at the difference between the MultiChoice and SentClf architectures, the SentClf delivers slightly better results, but the drawback is that this technique tends to link components to multiple components. Since most of our components have only one outgoing edge, it creates a lot of false positives, i.e., links which do not exist. With respect to exchanging the loss function, the weighted cross entropy helps the model to learn a better representation of the underrepresented classes. This is notable in the slightly increased, but more stable performance of SciBERT on the glaucoma and mixed test sets.

Moreover, comparing the confusion matrices of the weighted and unweighted SciBERT model, shown below, indicates a reduced error rate for the *support* class. However, the improvement for RoBERTa is only marginal. Furthermore, the errors for the *attack* class increased, meaning that the model could learn a better representation when components are related, but not the actual polarity of the relation.

Full Pipeline Besides the performance of the single pipeline modules, the overall performance of the whole argument mining pipeline was assessed. As stated earlier, for this, the two best performing models were chosen, i.e., the SciBERT with BiGRU and CRF for the sequence tagging part and the SciBERT with the weighted cross entropy loss for the relation classification part. The F1-scores for the 50% threshold (at least 50% of the tokens of a gold component need to be classified as argumentative to be counted as a true positive) are .54, .51 and .49 for the neoplasm, glaucoma and mixed test set, respectively. After increasing the threshold to 100% (all tokens of a gold component must be classified as one of the argumentative classes), the F1-scores are .55, .54 and .51 on the test sets. It may surprise that the stricter constraint (100%) achieves better results. However, this is due to how the relation classification is addressed. The input for the relation classification is generated by combining each of the detected components with the others. The stricter constraint results in fewer detected components, since gold components which did not reach the threshold of correctly predicted tokens are getting discarded. A fewer number of components results in fewer components which are paired with false positives from the component detection, which leads to fewer false positives in the relation classification and thus results in a higher F1-score. In general, the significant difference between the relation classification on predicted components compared to gold components is mostly due to the above described way of how the SentClf approach propagates and multiplies false positive errors from the component detection module. This is a weakness, which becomes more obvious with increasing text lengths. While our dataset consists of article abstracts only for practical reasons, the pipeline can be applied on full text articles as well. Alas, we cannot provide a quantitative analysis on full articles due to missing annotated data. In preliminary experiments on full articles, we have observed a notable increase of false positives in the relation classification, which is the expected consequence of an increased number of components. Furthermore, with the number of components rising in the double-digit range, the multiple-choice architecture loses its predictive power. We leave further investigations to determine how to refine this architecture to be applied on full text articles as future work.

Error Analysis Common mistakes for the sequence tagger are the invalid BIO sequences. Especially when there are multiple components in one sentence, the tagger tends to mislabel *B-* tokens as *I-* tokens. This is due to the natural imbalance between *B-* and *I-* tokens. Training the sequence tagging without the BIO

scheme using only *claim* and *evidence* as labels, poses problems when multiple components are following each other in the text. They would be extracted as one single component instead. This is a common case in concluding sentences at the end of a study, which strikingly often comprise multiple claims. Other notable mistakes arise for determining the exact component boundaries. Especially in the case of connectives, e.g., *however*, which have sometimes nothing but a conjunctive function, and in other cases signal a constraint of a previous statement. Another mistake is the misclassification of the description of the initial state of the participant groups as an observation of the study and therefore an evidence, e.g., *there were no significant differences in pregnancy-induced hypertension across supplement groups.* In the study abstract, these descriptions occur usually relatively close to the actual result description, which means that adding information of the position in the text will not avoid this error. While only some abstracts are structured, the full study report does usually have separated sections. This structure can be exploited when analysing full reports, and in the simplest case one would analyse only the sections of interest.

Concerning link prediction, general components like *the difference was not statistically significant* are problematic, since it could be linked to most of the components/outcomes of the trial. Here, a positional distance encoding could be beneficial, since those components are usually connected to the previous component. In general, most of the errors in the MultiChoice architecture were made in the multiple choice part by predicting a wrong link and not at the stage of classifying the relation type. Interestingly, comparing the two domain adapted models, Bio- and SciBERT, there were no regular errors, which allows any conclusion about the advantages or disadvantages of one model.

Looking at the confusion matrices, all tested SentClf models show a higher misclassification towards the *NoRelation* class. The confusion matrices for the SciBERT SentClf and its counterpart with the weighted loss function are shown exemplary in Figure 2. The *Support* relation was not as often misclassified as with the unweighted loss function. It can be further observed that the model could not learn a meaningful representation of the underrepresented *Attack* class, not even with the weighted loss function. There, the error rate even increased. Most of the attack relations were classified as *NoRelation*. These false negative errors indicate that in both cases the model is overly focusing on the *NoRelation* class. Concerning the learned representation of the relation classes, both transformer approaches have in common the problem of dealing with negations and limitations or associating the polarity of a measurement and therefore confusing support and attack, which might indicate that the model learns rather linguistic patterns than a deeper understanding of the components and their relations.

Example 5.1 [**more research about the exact components of a VR intervention and choice of outcomes to measure effectiveness is required**]$_{\textbf{source}}$ [**Conducting a pragmatic trial of effectiveness of a VR intervention among cancer survivors is both feasible and acceptable**]$_{\textbf{target}}$

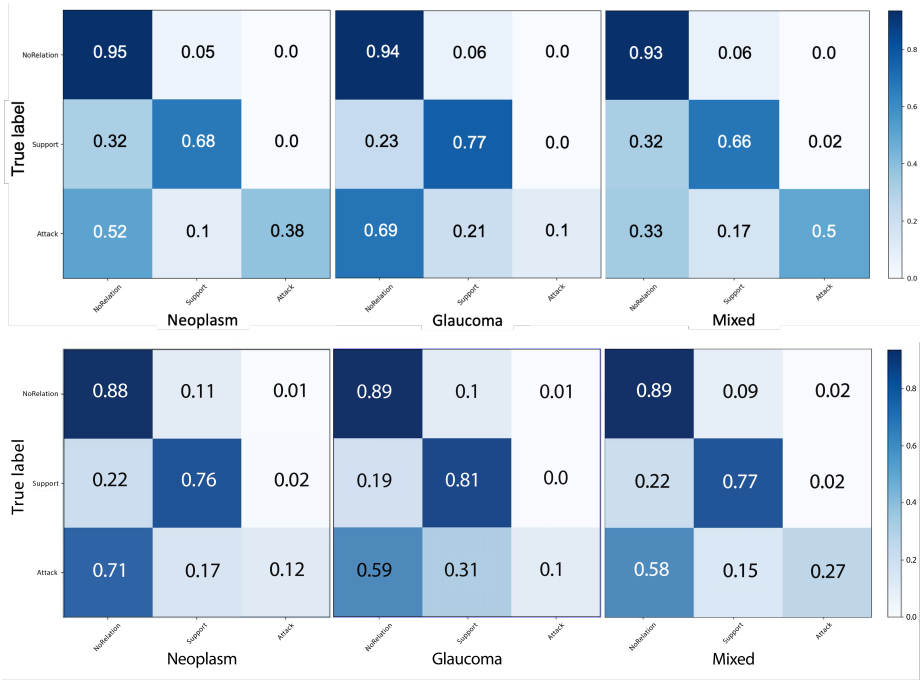


Figure 2: Confusion matrices of the predictions on the test set (neoplasm, glaucoma, mixed) of the relation classification task. SciBERT SentClf on top and SciBERT SentClf with weighted cross entropy loss on the bottom.

Example 5.2 [**this did not translate into improved progression-free survival (PFS) or overall survival**]$_{\textbf{source}}$ [**The addition of gemcitabine to carboplatin plus paclitaxel increased treatment burden, reduced PFS time, and did not improve OS in patients with advanced epithelial ovarian cancer**]$_{\textbf{target}}$

Example 5.1 shows two claims with a limiting (attacking) relation, which was wrongly classified as supporting. In Example 5.2, *not improving progression-free survival (PFS)* corresponds to a *reduced PFS time*, while for other factors reducing the value means it is beneficial, and therefore improving some study parameter. Here, the inclusion of external expert knowledge is crucial to learn these fine nuances. The polarity of a measurement cannot be learnt from textual features alone. Especially in the medical domain, there are complex interrelationships which are not often explicitly mentioned and therefore are impossible to capture with a model trained solely on character-based input. Phrases like *increased the blood pressure by X* or *showed no symptom of Y* can connote different messages depending on the context. Future work needs to consider this challenge of incorporating external expert knowledge. While we do not think this is a problem limited to a special domain, we consider it more relevant for

Model	macro	improved	increased	decreased	noDiff	noOcc
BERT (cased)	.62	.69	.65	.66	.75	.00
BERT (uncased)	.72	.72	.70	.72	.72	.50
BioBERT	.75	.74	.74	.77	.76	.54
SciBERT (cased)	.75	.71	.71	.73	.71	**.65**
SciBERT (uncased)	**.80**	**.81**	**.75**	**.81**	**.85**	.59

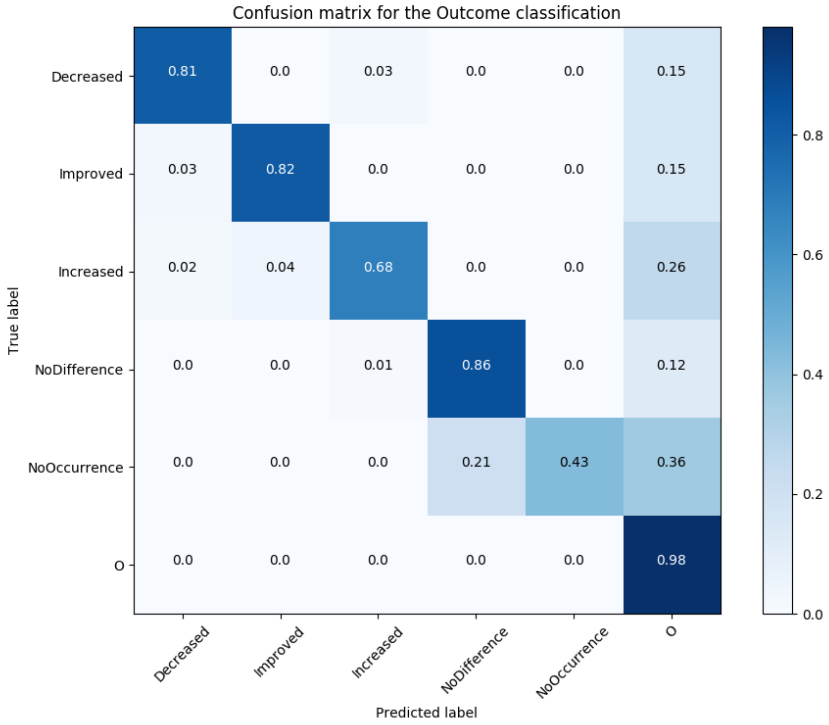
Table 6: Results for the outcome detection and classification tasks, given in F1-score.

understanding and representing medical text.

5.3.2 Outcome Analysis

The results for the outcome analysis tasks are shown in Table 6. Results are given on the test set in macro multi-class F1-score and as a binary F1-score for each of the five classes separately. Similarly to the relation classification results, we can observe an increase in performance on the specialized Bio- and SciBERT models compared to the general BERT model. In a direct comparison of the cased versions of these two specialised models, the overall F1-score is the same with .75. In the binary evaluation, BioBERT is slightly better with the exception of the *noOccurrence* class. Interestingly here, the SciBERT cased model performs the best with a F1-score of .65. Overall, SciBERT uncased is the best performing model with a macro F1-score of .80. It also outperforms the rest of the approaches in every F1-score measured except for the *noOccurrence* category, where the cased version has higher score. This category, in particular, suffers from sensitivity to class imbalance given that only 2% of the annotated data is labeled as such. For the other classes, the binary F1-scores are in a comparable range to each other, where the most prominent class in the annotated data, i.e., *noDifference* with 27%, has consistently the highest or second highest score. Besides the *noOccurrence* class, the *Increased* class has always the second lowest scores. Even for the best performing model, the difference compared to the worse performing models is not as massive as for the other classes. Notable in the confusion matrix, visualized in Figure 3, the classifier tends to wrongly predict it as *Improved*, which is a closely related class. The F1-score for the overall performance of the pipeline, i.e., with the argument component detection as a prior step, is .62 for the 50% and the 100% threshold. Both constraints produce a similar F1-score. Taking a look at the number of detected components for each of the constraints, there is only a total difference of 2 between them. Varying the threshold does not change the difference by much. We found that if the model detects a component most of the time at least 70% of the tokens are detected. Concerning the strong decrease from the gold label to the overall pipeline performance, we found that the *NoOccurrence* is the main reason, with not a single sample correctly predicted; either through not finding the component or, if detected, misclassifing the outcome with the wrong label. A similar situation was observed for the BERT cased model on

the gold standard, where the 0 F1-score of the *NoOccurrence* class lowered the macro F1-score significantly with respect to the other models. Ignoring the *NoOccurrence* class to estimate a performance value for the other classes, the macro F1-score would be at .74 for the whole pipeline.



:2:2

Figure 3: Confusion matrix of the predictions on the test set of the outcome classification task.

Error Analysis With respect to the source of error in the pipeline, the two pipeline parts cause different observable errors in the overall output. Being a binary classifier, the outcome detection is the only part which predicts the negative class label (referred to as *O* in the confusion matrix). The second part, the effect classifier, assigns effect class labels (*Increased/Decreased*, etc.) to outcomes, which were found by the outcome detection module. Consequently, the impact of the propagated error from the first part of the pipeline can be observed in the confusion matrix in Figure 3. Effect classes are mostly not misclassified as other effect classes, but as the negative class *O*. This is reflect in a stronger coloration in the horizontal direction for the predicted *O* label in the confusion matrix. Since the only part in the pipeline which is responsible for the negative *O* label is the outcome detection, this means that the error

occurred in the first part of the pipeline. Accordingly, confusion of effect class labels are errors of the second part, the effect classifier, in the pipeline.

One of the most common mistakes of the models is the incomplete detection of outcomes. In many cases, the outcome to classify includes other words that complement it, for example in the sentence *The levels of VEGF were significantly lower*, the outcome to classify is *The levels of VEGF* while the model only catches *VEGF*. We also find that the model is effectively tagging outcomes in such a way that is different from the true labels, but correct nonetheless. For example, consider the sentence *Excess limb size (circumference and water displacement) and excess water composition were reduced significantly.* This sentence has as true labels the outcomes *Excess limb size* and *excess water composition*, both labeled as *Decreased*. The model detects and classifies those outcomes correctly, but also adding the words *circumferences* and *water displacement*, predicting the label *Decreased* which would be the correct label.

6 Conclusions and Future Work

In this article, we presented our original research Argumentation Mining in the healthcare domain. We conducted an annotation study on 660 Randomized Controlled Trials to filter argumentative structures and evidence-based elements like PICO. We then annotated the abstracts of those RCTs with a structured argumentation model where arguments are composed by evidence and claims linked by support and attack relations (components Fleiss' kappa: 0.68, relations: Fleiss' kappa: 0.62). Furthermore, we annotated the effects on the outcomes associated to the identified argumentative components (Fleiss' kappa: 0.81). We proposed a full pipeline considering both argumentation structures detection and the classification of the effects on the outcomes. We employed a sequence tagging approach combining a domain specific BERT model with a GRU and CRF to identify and classify argument components. We cast the relation classification task as a multiple choice problem and compare it with recent transformers for sequence classification. The same sequence tagging architecture with the LSTM in combination with a CRF was experimented for the outcome detection and classification. The proposed approach significantly outperformed standard baselines and state-of-the-art AM systems with an overall macro F1-score of .87 for component detection, .68 for relation prediction, and a macro F1-score of .80 for outcome classification. We examined in depth the errors made by the system and proposed future improvements.

As the field of Evidence-Based Medicine is still evolving, and to foster future research in the area of argument mining and outcome analysis on clinical trials, we make available to the research community our annotation guidelines, the annotated data, the source codes for the experiments, as well as the results of our system for error analysis. We believe this is a valuable contribution to motivate the community to build upon our work. Moreover, we have integrated the proposed pipeline into ACTA [17], the tool we have developed for automating the argumentative analysis of clinical trials (both the argument component and the

relation detection modules are fully integrated, while we are currently working at the integration of the Effect-on-Outcome module). Such tool has been designed to support doctors and clinicians in identifying the document(s) of interest about a certain disease, and in analyzing the main argumentative content and PICO elements.

Two main research lines are pursued for future work. First, whilst in this paper we concentrate on intra-argument relations only, we aim to focus also on inter-argument relations. To this aim, we will annotate relations across different RCTs to allow reasoning on the resulting argument graphs and clustering of arguments about the same disease with the aim to automatically identify, for instance, possible controversies among the conclusions of the RCTs about a certain disease. Second, as one of the main features of argumentation models is the capability to capture inconsistencies [21], we aim at mining argument components also in the full text of the RCTs. As it has been noticed in the literature [67], sometimes RCT abstracts contain a more positive reporting of the main findings of the article than what stated in the full text. Employing argumentation mining methods to automatically identify these instances of misrepresentation and distortion of the results in RCTs is a challenging and crucial research line for healthcare intelligent applications.

Acknowledgements

This work is partly funded by the French government labelled PIA program under its IDEX UCA JEDI project (ANR-15-IDEX-0001). This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002

References

[1] Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of BioNLP 2018 workshop*, pages 67–75, 2018.

[2] Anthony Hunter and Matthew Williams. Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190, 2012.

[3] Robert Craven, Francesca Toni, Cristian Cadar, Adrian Hadad, and Matthew Williams. Efficient argumentation for medical decision-making. In *Proceedings of KR 2012*, pages 598–602, 2012.

[4] Luca Longo and Lucy Hederman. Argumentation theory for decision support in health-care: A comparison with machine learning. In *Proceedings of BHI 2013*, pages 168–180, 2013.

[5] Malik Al Qassas, Daniela Fogli, Massimiliano Giacomin, and Giovanni Guida. Analysis of clinical discussions based on argumentation schemes. *Procedia Computer Science*, 64:282–289, 2015.

[6] Michael Chary, Saumil Parikh, Alex Manini, Edward Boyer, and Michael Radeous. A review of natural language processing in medical education. *Western Journal of Emergency Medicine*, 20:78–86, 12 2018.

[7] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, 2013.

[8] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10, 2016.

[9] Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of IJCAI*, pages 5427–5433, 2018.

[10] John Lawrence and Chris Reed. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818, 2019.

[11] D.L. Sackett, W.M.C. Rosenberg, J.A. Gray, bhaynes@mcmaster.ca Haynes, and W.S. Richardson. Evidence based medicine: What it is and what it isn't. *BMJ (Clinical research ed.)*, 312:71–2, 02 1996.

[12] Hamed Hassanzadeh, Mahnoosh Kholghi, Anthony N. Nguyen, and Kevin Chu. Clinical document classification using labeled and unlabeled data across hospitals. In *AMIA 2018*. AMIA, 2018.

[13] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases - International. Proceedings of Workshops of ECML PKDD 2019*, volume 1168 of *Communications in Computer and Information Science*, pages 727–740. Springer, 2019.

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

[15] Nancy Green. Argumentation for scientific claims in a biomedical research article. In *Proceedings of ArgNLP 2014 workshop*, 2014.

[16] Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. Argument mining on clinical trials. In *Proceedings of COMMA 2018*, pages 137–148, 2018.

[17] Tobias Mayer, Elena Cabrio, and Serena Villata. ACTA a tool for argumentative clinical trial analysis. In *Proceedings of IJCAI 2019*, pages 6551–6553, 2019.

[18] Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based argument mining for healthcare applications. In *Proceedings of ECAI 2020*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press, 2020.

[19] W. Richardson, M. Wilson, J. Nishikawa, and R. Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123 3:A12–3, 1995.

[20] Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014.

[21] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.

[22] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Comput. Linguist.*, 43(3):619–659, 2017.

[23] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP 2009*, pages 1493–1502, 2009.

[24] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of EACL 2017*, pages 251–261, 2017.

[25] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of AAAI 2018*, pages 4889–4896, 2018.

[26] Shohreh Haddadan, Elena Cabrio, and Serena Villata. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of ACL 2019, Volume 1: Long Papers*, pages 4684–4690. Association for Computational Linguistics, 2019.

[27] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In *Proceedings of NAACL-HLT 2019*, page 2131–2137, 2019.

[28] Jure Zabkar, Martin Mozina, Jerneja Videcnik, and Ivan Bratko. Argument based machine learning in a medical domain. In *Proceedings of COMMA 2006*, pages 59–70, 2006.

[29] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL 2017*, pages 11–22, 2017.

[30] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL 2016*, pages 1105–1116, 2016.

[31] Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of ACL 2016*, pages 231–235, 2016.

[32] Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128. Association for Computational Linguistics, September 2017.

[33] Maximilian Spliethöver, Jonas Klaff, and Hendrik Heuer. Is it worth the attention? a comparative evaluation of attention layers for argument unit segmentation. In *Proceedings of the 6th Workshop on Argument Mining 2019*, pages 74–82. Association for Computational Linguistics, August 2019.

[34] Peter Potash, Alexey Romanov, and Anna Rumshisky. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of EMNLP 2017*, pages 1364–1373, 2017.

[35] Andrea Galassi, Marco Lippi, and Paolo Torroni. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of ArgMining 2018 workshop*, pages 1–10, 2018.

[36] Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured SVMs and RNNs. In *Proceedings of ACL 2017*, pages 985–995, 2017.

[37] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of ACL 2019*, pages 567–578, 2019.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019.

[39] Abdulaziz Alamri and R.M. Stevenson. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of Biomedical Semantics*, 7, 05 2016.

[40] Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. Exact: Automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10:56, 09 2010.

[41] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of ACL 2018*, pages 197–207. Association for Computational Linguistics, July 2018.

[42] Antonio Trenta, Anthony Hunter, and Sebastian Riedel. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *CoRR*, abs/1509.05209, 2015.

[43] Di Jin and Peter Szolovits. Advancing PICO element detection in biomedical text via deep neural networks. *Bioinformatics*, 36(12):3856–3862, 2020.

[44] Iain Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12. Association for Computational Linguistics, July 2017.

[45] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the NACL 2019*, pages 3705–3717. Association for Computational Linguistics, June 2019.

[46] Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Proceedings of NACL 2007.*, pages 260–267. Association for Computational Linguistics, April 2007.

[47] Edward Hannan. Randomized clinical trials and observational studies guidelines for assessing respective strengths and limitations. *JACC. Cardiovascular interventions*, 1:211–7, 07 2008.

[48] Gordon H. Guyatt, R. Brian Haynes, Roman Z. Jaeschke, Deborah J. Cook, Lee Green, C. David Naylor, Mark C. Wilson, W. Scott Richardson, and for the Evidence-Based Medicine Working Group. Users' Guides to the Medical LiteratureXXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care. *JAMA*, 284(10):1290–1296, 09 2000.

[49] Kenneth F Schulz and David A Grimes. Generation of allocation sequences in randomised trials: chance, not choice. *The Lancet*, 359(9305):515 – 519, 2002.

[50] Nancy Green. Annotating evidence-based argumentation in biomedical text. *IEEE BIBM 2015*, pages 922–929, 2015.

[51] Leo A Groarke and Christopher W Tindale. Good reasoning matters: A constructive approach to critical thinking. 1997.

[52] György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45. Association for Computational Linguistics, June 2008.

[53] Sally Hopewell, Mike Clarke, David Moher, Elizabeth Wager, Philippa Middleton, Douglas G Altman, Kenneth F Schulz, , and the CONSORT Group. Consort for reporting randomized controlled trials in journal and conference abstracts: Explanation and elaboration. *PLOS Medicine*, 5(1):1–9, 01 2008.

[54] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[55] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16, 12 2016.

[56] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[57] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[58] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, 2014.

[59] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of LREC 2018*, pages 3483–3487, 2018.

[60] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237, 2018.

[61] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of COLING 2018*, pages 1638–1649, 2018.

[62] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2019.

[63] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3615–3620, 2019.

[64] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of EMNLP 2018*, pages 93–104, 2018.

[65] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[66] Isaac Persing and Vincent Ng. End-to-end argumentation mining in student essays. In *Proceedings of NAACL-HLT 2016*, pages 1384–1394, 2016.

[67] Isabelle Boutron and Philippe Ravaud. Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences*, 115(11):2613–2619, 2018.