LitTx: A New Treatment Relation Extraction Dataset

Yuhang Jiang*, Md Sultan Al Nahian†, Li Hao Richie Xu*
Rani Chikkanna*, and Ramakanth Kavuluru*
*University of Kentucky, Lexington, Kentucky, USA
†Penn State Harrisburg, Middletown, Pennsylvania, USA

Abstract

The interest in biomedical relation extraction (RE) continues to persist even in the LLM era owing to RE being a prominent way to build knowledge graphs, which further ground LLM applications, especially in preventing hallucinations. Therapy-disease treatment relations from scientific literature are an important type in RE as they indicate emerging therapeutic hypotheses and off-label usages being explored in the community. An automatically extracted evolving knowledge-base of such relations will be of great utility to researchers because doing it manually is not viable with the exponential growth of biomedical articles. In this paper, toward this end, we introduce a new expert-annotated dataset **LitTx** for identifying treatment relationships discussed in literature given the lack of such datasets in the recent past. Besides confirmed or implied positive relations, we also introduce a new "conditional treatment" relation type where hedging or a potential relationship is indicated. Our baseline RE models with this new dataset demonstrate promising results, while also revealing clear areas for improvement. To foster innovation and ensure replicability in the biomedical RE community, we release our dataset, code, and annotation guidelines publicly: https://github.com/bionlproc/LitTx_dataset.

1. Introduction

Biomedical *entities* and the *relations* that connect them are at the core of knowledge discovery. Diseases, medications, procedures, genes, variants are entity types that are central to modern medicine. However, what is more important is how these entities are linked with each other driving etiology, pathology, diagnosis, prognosis, and recovery. Some popular types of such links or relations (with applications in parentheses) are: protein-protein interactions (to understand disease etiology and progression), gene-disease associations (to identify potential drug targets), drug-disease treatment relations (to spot off-label usage or assess potential for repositioning), and drug-drug interactions (to surveil adverse events). Biomedical knowledge graphs (KGs) formed with entities as nodes and relations as edges are valuable resources used as background knowledge to perform question answering (QA) and facilitate free text search for a set of relevant documents either from literature or clinical notes (Goodwin and Harabagiu, 2017). Direct reasoning over knowledge graphs can also lead to discovery giving rise to the so called problem of knowledge base question answering (KBQA) (Raghavan et al., 2021; Sima et al., 2021). Recently, biomedical KGs have also been shown to be useful in reducing hallucinations and grounding knowledge discovery and reasoning tasks with large language models (LLMs) (Xiong et al., 2025). Overall, we can see that relational information will continue to be a crucial driver of biomedical data science to either enable or enhance capabilities of automated methods in biomedical research and healthcare operations. Discovering certain types of new relations (e.g., gene-disease associations) is nontrivial and is done based on extensive experimental research. Newly discovered relations are typically reported in scientific literature after undergoing sufficient peer review. Certain types of clinical relations (e.g., drug–side-effect) may also be reported in clinical notes or also on social media as patients report them. Overall, the task of relation extraction (RE) from natural language sources (literature, clinical notes, social media posts) is an indispensable precursor step to build evolving KGs and support downstream discovery tasks.

An important type of biomedical relation is the therapeutic agent-disease treatment relation. Consider the sentence: "We conclude that tamoxifen therapy is more effective for early stage breast cancer patients." From this sentence an RE model is expected to extract the relation (tamoxifen, TREATS, breast cancer) where tamoxifen is the *subject* entity, breast cancer is the *object* entity, and TREATS is the relation type or more formally the *predicate*. Note that subject and object entities are strings (spans) from the input sentence. Let's consider a different sentence: "Condition of postmenopausal breast cancer patients has been observed to improve significantly after they have been on tamoxifen for five years post surgery." Although it is worded differently compared to first sentence, an RE model is expected extract the same relation (tamoxifen, TREATS, breast cancer) from this one too. Often there is an extra entity normalization step where synonyms are collapsed to a concept in a standardized vocabulary (e.g., RxNorm for drugs). For instance, instead of tamoxifen if we see "Nolvadex" in either sentence, it may also be extracted as (tamoxifen, TREATS, breast cancer) since Nolvadex is a brand name for a particular

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1352–1360
11–16 May 2026.

tamoxifen formulation. Since normalization is often treated as an auxiliary step (although important), we do not further discuss this in this paper.

Therapy-disease treatment relations are often reported in literature based on clinical trial results or other prospective studies. They are also reported in case reports especially in the context of off-label usage. As such, scientific literature is a goldmine to continuously extract treatment relations in an automated fashion using RE methods. The biomedical natural language processing (BioNLP) community has a rich history of building *supervised* RE methods to extract various types of relations from text including treatment relations (Luo et al., 2017; Huang et al., 2024). An important indispensable resource needed to build RE methods is human expert annotated data that can be used for training and testing supervised models. In the zero shot setting, they can be used purely as test benchmarks. Our current paper introduces one such physician annotated new dataset for treatment relations.

As our main contribution, we introduce a new **Lit**erature derived **T**reatment (**LitTx**) dataset annotated by two practicing physicians who are hospitalists (Internal Medicine), treating inpatients in a large academic hospital in the USA. Each instance is annotated by both of them with disagreement resolution meetings after each round. We use a fine-grained label scheme where "conditional treatment" is used to indicate hedging or inconclusive statements where experiments are conducted but results are not stated. We also report on baseline experiments with encoder-only, encoder-decoder, and decoder-only models for the created dataset along with an examination of errors. The full dataset, annotation guidelines, and code are available here: `https://github.com/bionlproc/LitTx_dataset`. Please note that due to label leakage concerns in the context of modern frontier LLMs, following recent recommendations (Jacovi et al., 2023), we have encrypted our data. We do provide instructions to decrypt using the `decrypt.py` script in the GitHub repository.

The structure of this paper is organized as follows. Section 2 discusses related work on prior treatment relation extraction datasets and methods typically used for RE. Section 3 describes the dataset creation process, including task definition, data collection, annotation procedure, the resulting dataset statistics, and evaluation protocol. Section 4 presents the baseline models used in our experiments. Section 5 reports experimental results, including multi-class and binary-label settings as well as preliminary error analysis. Finally, Section 6 concludes the paper, followed by ethical considerations and references.

2. Related Work

Gold standard datasets have played a central role in advancing biomedical relation extraction. An important effort in gold standard datasets for biomedical RE was by Kilicoglu et al. (2011), which has 126 treatment relations in the final dataset although the entire dataset has over a dozen predicates (and their negated versions) besides the TREATS predicate. Although treatment relations represent only one subset of the overall schema, this work demonstrated the feasibility of fine-grained semantic annotation for biomedical RE. Another dataset is based on the crowd sourcing effort by Dumitrache et al. (2018) where there were a total of 606 sentences annotated for potential treatment relations with 295 being positive. The annotations were conducted by third year medical students and each sentence was only annotated by exactly one student. While this approach enabled scalable annotation, the fact that each sentence was annotated by a single annotator limits formal assessment of inter-annotator agreement and adjudication.

Recently, Tiktinsky et al. (2022) introduced a gold standard dataset for annotating combination drug therapies where seven biomedical engineering graduate students who took a course on the topic were used in the annotation with domain expert supervision. The test set has 272 instances each annotated by at least two coders and reviewed by the domain expert. In this dataset, although combinations that were deemed helpful were identified, the corresponding disease was *not* annotated, limiting direct applicability to drug–disease TREATS extraction. All the three datasets mentioned above are from scientific literature (PubMed abstracts). In contrast, the 2010 i2b2/VA challenge dataset (Uzuner et al., 2011) was generated from clinical notes. This dataset has two relation types that are pertinent for our effort: (a) treatment improved a medical condition (203 relations) and (b) treatment administered for medical condition (2613 relations) (D'Souza and Ng, 2014). Each instance was annotated by at least two annotators, which included nurses, nursing students, a respiratory therapist and some non-clinicians (South et al., 2011). This dataset is qualitatively different from literature derived treatment relations as it is focused on administering care in a hospital, with each instance relating to what happened to a single patient. Despite these contributions, several gaps remain. Many literature-derived datasets contain relatively limited numbers of explicitly adjudicated TREATS relations. Annotation expertise also varies, with relatively few datasets annotated directly by practicing physicians. In addition, some resources focus on specialized treatment scenarios such as combination therapies

without explicitly annotating the corresponding disease entity, thereby limiting their applicability for drug–disease treatment relation extraction.

In terms of methods for biomedical RE, most efforts used some type of deep neural architecture over the past decade. Researchers used convolutional neural networks (Liu et al., 2016), recurrent architectures (Kavuluru et al., 2017), and ensembles of these two types of models with linear models (Peng et al., 2018). Since transformers were invented, most RE models employed encoder-models (Wei et al., 2020; Jiang and Kavuluru, 2025) or sequence to sequence methods (Giorgi et al., 2022; Jiang and Kavuluru, 2023). While joint models that predict both entities and relations together in a single pass (Miwa and Bansal, 2016; Tran and Kavuluru, 2019b) have been popular, there is also a rise in pipeline models (Zhong and Chen, 2021). The transformer based encoder-decoder T5 models (Phan et al., 2021) and the decoder-only LLMs such as Llama (Hu et al., 2026) are currently more well known for RE.

3. The Dataset Creation and Composition

3.1. Task definition

The main RE task here is relation classification: given an input sentence $\mathcal{S}$ and a pair of entities e_1 and e_2 that are marked up as spans in the sentence, the goal is to categorize the semantic relationship ρ between these entities from a set of predefined relationships $\mathcal{R}$. More concretely, we would like a classification model $\mathcal{M}(\mathcal{S}, e_1, e_2) \rightarrow \rho$, where $\rho \in \mathcal{R}$.

3.2. Data collection

We employ our prior distant supervision approach specific to treatment relations using the MeSH Subheading-based Distant Supervision (MSDS) method (Tran and Kavuluru, 2019a), which leverages medical subject headings (MeSH) assigned to PubMed articles along with the so called "MeSH subheadings" that capture fine-grained details covered in an article. MeSH is a standardized vocabulary used by the U.S. National Library of Medicine (NLM) to index biomedical articles facilitating better search on PubMed. It is organized as a hierarchy with nearly 30,000 unique terms as of 2025. Each article NLM indexes in PubMed is assigned nearly 13 MeSH headings (or terms) on average to capture the main themes covered in it. MeSH also has subheadings which further qualify the main heading. For example, *Breast Neoplasms* is a MesH heading and any article which is tagged with this heading is expected to be about breast tumors in some way. A subheading can further qualify and offer hints on what specifically is covered in the article about this main theme. For example, *Breast Neoplasms/therapy* is used to indicate that the article is discussing the treatment aspects of breast tumors via the *therapy* subheading. Likewise, *Breast Neoplasms/epidemiology* conveys that the article talks about distributional aspects of the condition in the population given the *epidemiology* subheading. To further illustrate the idea of MeSH headings and subheadings, we use the example from Tran and Kavuluru (2019a): "A 15-year-old female adolescent developed drug hypersensitivity syndrome 4 weeks after starting minocycline therapy for acne vulgaris." The article containing this sentence is indexed with two MeSH headings: *Minocycline/therapeutic use* and *Acne Vulgaris/drug therapy*. Clearly, a drug with the *therapeutic use* subheading and a disease with the *drug therapy* subheading indicate a "potential" treatment relation if both of them occur in the same sentence, as is demonstrated in the example sentence. Using this insight, Tran and Kavuluru mine for sentences that likely contain treatment relations and show that this MSDS approach helps with training supervised models without any hand labeling by domain experts.

Of course, there is no guarantee that sentences identified using this MeSH heading/subheading heuristic always express treatment relations unless human experts look at them. We take up this endeavor here and search for PubMed abstracts that contain both the *therapeutic use* and *therapy* MeSH subheadings with MSDS. Please note that the *therapy* subheading has *drug therapy* as a child in the subheading hierarchy. The broader *therapy* subheading allows us to cast a wide net for therapies beyond drugs including procedures, medical devices, and supplements. We follow the procedure as exactly described in the original paper (Tran and Kavuluru, 2019a) including considerations for certain exceptions and nuances they make. Specifically, we follow the semantic type constraints on the subject and object entities for the TREATS predicate used by the creators of the SemMedDB predication database at the NLM (Kilicoglu et al., 2012). In the end, the subheadings enable us to automatically identify and extract sentences that likely describe treatment relationships between therapies and diseases, streamlining the creation of a gold standard dataset focused on such relations. The MSDS method provides both a sentence and a pair of entities – a therapy entity and a disease entity – with the potential for expressing a treatment relation between them. This pair is identified by the MetaMap concept mapping tool made available by the NLM (Aronson and Lang, 2010). By using MSDS, we improve the effi-

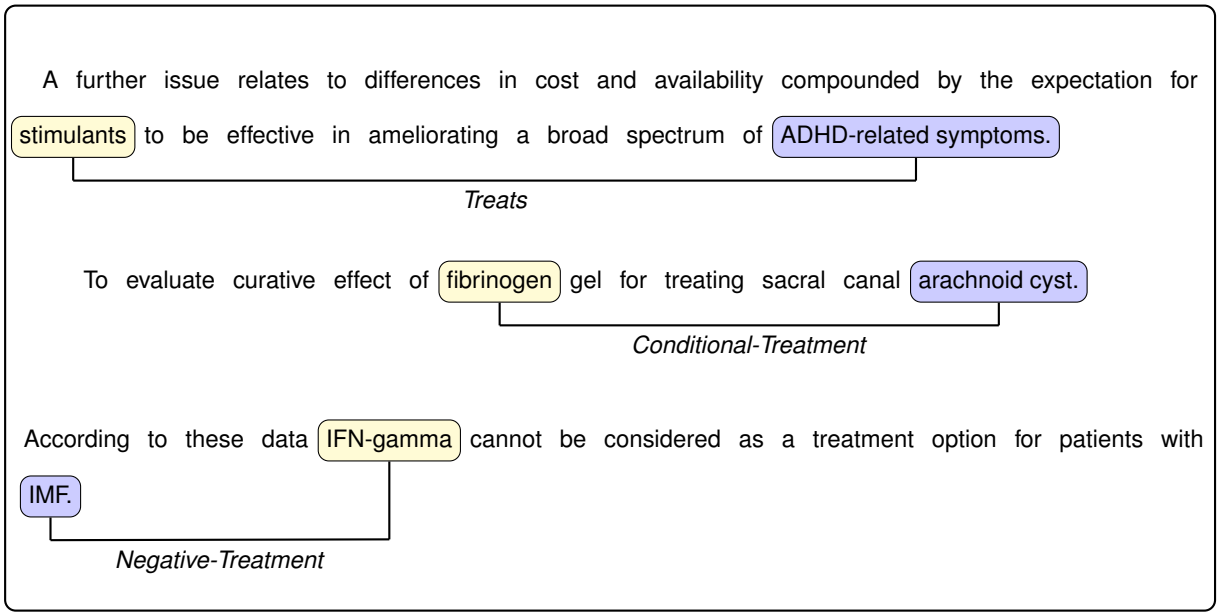

Figure 1: Examples of *Treats*, *Conditional Treatment* and *Negative Treatment* relations. Entities are shown in colored boxes, where yellow denotes the therapy entity and blue represents disease entity.

ciency and scalability of the data collection phase compared to traditional methods. It also helps to create a dataset that is more balanced across various types of meaningful interactions ready to be examined by our annotators.

3.3. Annotation process

LitTx has mainly two types of entities: therapeutic entities (drugs, procedures, supplements . . .) and disease entities. Given pre-spotted spans for a therapy subject and a disease object in the input sentence, the goal of expert annotation is to determine the most apt relationship from the following four types:

- **Treats**: the sentence indicates a therapeutic action or intervention such that the application or administration of the *therapy* results in a positive therapeutic outcome or alleviation of symptoms associated with the *disease*. This might also be an implication, rather than a direct assertion, depending on the verbiage used in the sentence.

- **Conditional Treatment**: the treatment relation is conditional or dependent on an experiment, trial, or evaluation that the authors (of the PubMed abstract) plan to conduct, as indicated in the sentence.

- **Negative Treatment**: the absence of a therapeutic effect is being indicated, and it could refer to negative results where a therapy is ineffective for a condition or cases where there is insufficient or minimal evidence supporting its effectiveness.

- **Other**: when none of the above three categories are acceptable, indicating that no treatment relation is present between the therapy and the disease. That said, others types of relations may be present (e.g., therapies sometimes backfire and cause some conditions as side effects). But these others types of relations are not relevant to this project. Thus here "Other" indicates none from the above three types.

As indicated earlier, two practicing physicians annotated each instance independently based on multiple training sessions and subsequent practice sessions. The practice sessions led to refinement of the initial annotator guidelines. The annotation was carried out using the serverless browser based annotation tool MedTator (He et al., 2022). Although the annotators' medical knowledge was important to interpret the language in a sentence, they were directed to refrain from juxtaposing external knowledge into the sentential context. For example, they were directed to focus on what was expressed in the sentence even if the extracted treatment is not the current standard of practice, considering the task is to focus on RE from input text. Thus, it was not important to assess whether what is in the input aligns with how they would treat their patients.

Considering each instance is annotated by both annotators, cases where both agreed on the labels were directly included in the dataset as finalized annotations. For instances where the annotators disagreed on the labels, a structured discussion process was initiated: the annotators reviewed the instances together, referencing the annotation

Metric	Treats	Cond. Treatment	Neg. Treatment	Other	Overall
# Instances Agreed	402	141	9	495	1047
# Instances Disagreed	129	92	7	132	360
Agreement Percentage	75.7%	60.5%	56.3%	78.9%	74.4%

Table 1: Inter-annotator agreement measured against finalized annotation classes.

guidelines to clarify ambiguities in the presence of the project's NLP lead. Disagreements were resolved through consensus, with both annotators working to reach a mutual understanding of the correct label. After the discussion, a final label was assigned to the disputed instances, reflecting the consensus reached by the annotators.

3.4. Resulting dataset

The resulting dataset consists of 1,407 annotated sentences, each containing a therapeutic entity and a disease entity paired with an assigned label. The dataset is divided into training, validation, and test sets with approximate proportions of 50%, 20%, and 30%, respectively. The detailed information is demonstrated in Table 2, which clearly shows that in the entire dataset, around 45% have the "other" label as they do not participate in any relation. Around 16% belong to the "conditional treatment" label that we have not seen in other datasets. Nearly 38% belong to the main TREATS label and there are very few "negative treatment" labels. The latter may be expected since most papers do not want to discuss negative results. Overall, this is the largest literature derived treatment relation dataset (compared with the prior datasets reviewed in Section 1). Furthermore, this is the only treatment dataset annotated by practicing physicians who are likely to search literature for off-label use-cases and latest advances.

Split	# Treats	# Con.	# Neg.	# Other	# Total
Train	287	115	8	295	705
Valid	98	47	4	131	280
Test	146	71	4	201	422

Table 2: Overall statistics of the LitTx dataset.

Table 1 shows the agreement of the two annotators against the final consolidated dataset. Overall, their agreement is at nearly 75% and most of the confusion was between the main two classes of interest. Cohen's $\kappa = 0.61$ with the 4-way classification was also decent. It ranks as moderate to substantial as per the original rules of thumb by Landis and Koch (1977). However, as Sim and Wright (2005) discuss, for non-binary cases Cohen's κ can be misleading and trends low as the number of categories increases. We present three annotated examples from the training set in Figure 1 and point the readers to our code repository for the full dataset.

3.5. Evaluation

While our main contribution is the dataset created, we also provide some baseline experiments with the dataset splits provided in Table 2 to predict class labels for the annotated entities. We provide both micro-F_1 and macro-F_1 scores (sans the "other" class) under two evaluation scenarios:

- **Multi-class**: This is the default setup where all labels are treated separately and the problem is treated as 4-way classification.
- **Binary-label**: The *Treats* and *Conditional Treatment* labels are combined into a single unified positive class, representing any effective or conditionally effective treatment. Similarly, the *Negative Treatment* and *Other* labels are combined into a unified negative class, representing either ineffective treatment or no treatment relationship. In this simpler setting Cohen's $\kappa = 0.68$, jumps to a substantial range, as expected.

4. Models

We conducted experiments with three different kinds of models: an encoder-only model PURE, an encoder-decoder based sequence-to-sequence model T5, and a decoder-only LLM Llama-3.

4.1. PURE

The Princeton University Relation Extraction (PURE) model (Zhong and Chen, 2021) employs a BERT-based encoder, such as BiomedBERT in our case, to process input sequences and generate contextual embeddings for each token. Although PURE is a pipeline approach encompassing both a named entity recognition module and a relation module, we only employ the relation module in this setting. For each entity, it inserts two special typed markers before and after the entity span to denote the entity's start/end positions, and its type. Using the encoder model's output, the contextual

Relation Extraction Template for Llama-3

Model Input: You are a helpful medical expert, and your task is to extract the relation between the given entity pairs. Organize your output in a json formatted as Dict{{"answer_choice": str(A/B/C/...)}}. Your responses will be used for research purposes only, so please have a definite answer.
User: Here is the input text for relation extraction:
<Text>
What is the potential relationship between the therapy <Drug/Procedure/...> and the disease <Disease>?
Here are the potential choices:
A. Treats
B. Conditional Treatment
C. Negative Treatment
D. Other
Please generate your output in json.
Assistant:
Model Output: {"answer_choice": "X"}

Figure 2: An example prompt for RE with Llama-3 model.

representations of the two start markers are concatenated and passed through a softmax layer to predict the relation type.

4.2. T5

The T5 (Text-to-Text Transfer Transformer) models represent a versatile and powerful framework for NLP tasks (Raffel et al., 2020). Built on the encoder-decoder architecture, T5 adopts a unified "text-to-text" approach, where tasks like classification, translation, summarization, or question answering are cast into a single text-to-text format. We employ the T5-3B model and optimize the next-token prediction objective for relation classification. The relation bearing sentence is provided as the model's input, while a label word (e.g., "treats", "conditional") is generated as the prediction.

4.3. Llama-3

The LLaMA (Large Language Model Meta AI) decoder models have been widely used in recent years. With models available in different sizes, we select the latest Llama-3.1-8B-Instruct model (Dubey et al., 2024) for the need that balances efficiency with ease of fine-tuning and language instruction. We develop a template to guide the model in performing a multiple-choice RE task. To ensure the model generates more consistent outputs, especially to prevent it from producing unexpected responses, we enforce a JSON format for reliable output consistency. The whole template is shown in Figure 2.

5. Experimental Results

5.1. Multi-class results

We present the results for 4-way classification in Table 3. The largest model, Llama3.1-8B-Instruct, achieves superior performance compared to the other two models, both in terms of micro-averaged and macro-averaged F1 scores. The LLM demonstrates its strength in handling datasets with a small number of examples, showcasing its ability to generalize effectively from limited data. On the other hand, PURE with the BiomedBERT base model, which has only 110M parameters, also achieved a comparable performance in terms of micro-averaged F1 score. Despite its smaller size, PURE effectively leverages domain-specific pretraining on biomedical text, allowing it to excel in tasks where domain knowledge plays a critical role. Its lightweight architecture enables faster training and inference, making it a practical choice for environments with limited computational resources. Among the four labels, the main *treats* label has achieved highest F1 for all three models. The *conditional treatment label*, being more semantically complex, proved more challenging to determine. Meanwhile, the *negative treatment* type had very few examples in the test dataset, leading to highly variable performance across the models.

5.2. Binary-label results

The results for the binary setting are shown in Table 4. Recall that this evaluation computes the F1 score for the unified class of *treats* and *conditional treatment* labels, treating the two remaining labels together as the negative class. Note that we do

Model	*Treats*	*Con. Treatment*	*Neg. Treatment*	Micro-F_1	Macro-F_1
T5-3B	70.2	53.0	40.0	64.7	54.4
PURE-BiomedBERT	67.9	56.9	33.3	64.1	52.7
Llama-3.1-8B-Ins.	**72.1**	**60.1**	**75.0**	**67.8**	**69.1**

Table 3: Multi-class F_1 scores for each category for different models.

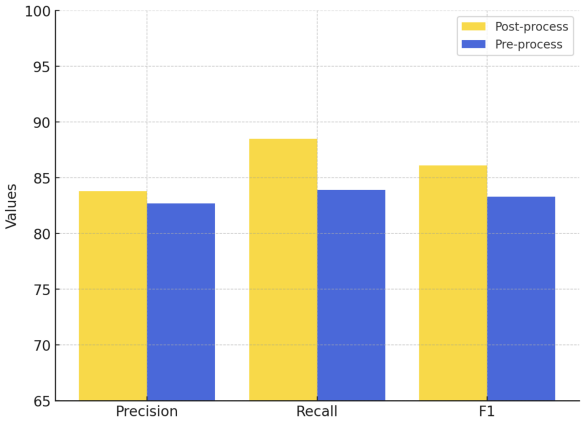

Figure 3: Comparison of binary-label performance with the pre- and post-processing strategies.

not retrain a model here but simply merge 4-way labels from the multi-class predictions to the new binary setting. The Llama-3.1-8B-Instruct model outperforms the other two models again, achieving at least a 4% improvement in F1 score. The fact that micro-F1 (from Table 3) jumps from mid sixties to low eighties here indicates that there is lot of confusion to distinguish between the main *treats* and *conditional treatment* labels with all models. However, we believe this distinction is important as sentences where the language is not clear should not be treated as conclusive.

Model	Precision	Recall	F_1
T5-3B	73.2	93.1	81.9
PURE-BiomedBERT	77.0	84.8	80.7
Llama-3.1-8B-Ins.	83.8	88.5	**86.1**

Table 4: Binary-label precision, recall and F_1 scores over different models. The results are directly computed from Table 3.

5.3. Training with binary labels

In Section 5.2, the binary evaluation was performed by merging labels from the 4-way predictions. So instead of post-processing 4-way labels, here we re-train the Llama-3.1-8B-Instruct model directly on pre-processed binary labels. As Figure 3 shows, training on the original 4-way labels and post-processing clearly outperforms training from scratch on binary-labels (pre-process) in all metrics: precision, recall and F1 score. Due to the richer feature learning enabled by the multi-class task, the model is likely exposed to more nuanced distinctions between categories. It may have developed a deeper understanding of the underlying patterns in the data, which subsequently improved its performance.

5.4. Preliminary error analysis

We discuss some interesting examples of errors caused by the models to throw light on where they are lacking, which can be used to guide future methods development on this task.

- **Over reliance on keywords**: At times, the model appeared to rely on keywords like “therapy" and “treatment", which strongly correlate with the main *treats* label. However, it overlooked contextual clues indicating a conditional relationship, such as mentions of “study” or “cohort”. For example, consider the sentence: “It was an ambispective cohort study carried out in 11 **HIV** units in Spain and involved 711 consecutive patients positive for HIV / HCV who started interferon plus **ribavirin** therapy.” Although there is no stated or implied treatment effect, it chooses *treats* instead of *conditional treats*, the gold label.
- **Inability to overcome strong lexical priors**: Considering the large sizes of existing models, lexical priors likely encoded strong signal based on pretraining data. Consider the sentence: “Using a nude rat human xenograft model of extremity **melanoma**, we analyzed tumors for glutathione (GSH), the main protein in the **melphalan** resistance pathway.” Although melphalan is a traditional treatment option for melanoma, this particular sentence is talking about analyzing tumors that maybe resistant to melphalan. The model was unable to overcome the embedded knowledge that melphalan is generally associated with treating melanomas and lacked focus on the “resistance pathway.” The gold label was *conditional treats* but the model predicted *treats* here.
- **Lack of focus on disease phrase modifiers**: Consider this sentence where the gold label

was *other*, where the model predicted *treats*: "Since conventional analgesics and sympatholytic drugs are of no benefit in the treatment of established postherpetic neuralgia, the sequelae of **herpes zoster** must, therefore, be recognized and treated with **amitriptyline** as soon as possible." Here the spotted condition was herpes zoster and the treatment effect of amitriptyline is being stated for "sequelae of" herpes zoster including postherpetic neuralgia. So the semantics of the modifier phrase "sequelae of" is not fully accounted for, leading to the incorrect prediction.

6. Conclusion

We introduced a new dataset **LitTx** for identifying treatment relations from scientific literature based on annotations by two practicing physicians. To our knowledge, this is the largest dataset of its kind and distinguishes between clear treatment signal and conditional or hedged statements. Our baseline models achieve promising results, while also highlighting clear opportunities for improvement especially in the 4-way setup. The dataset is ready for benchmarking of new RE methods with our error analysis offering some clues as to what issues need to be fixed.

7. Ethical Considerations and Limitations

Considering we use biomedical abstracts that are available for public use, there are no apparent ethical issues including copyright or privacy concerns. Our dataset and methods are limited to English only and that is a clear limitation of this project and cannot be easily overcome (owing to language skills of our annotators being somewhat limited based on our location.) Another caveat is that our dataset is limited to sentence-level relations and does not capture relations that manifest across sentences through co-references. This would be a future project as it is substantially more resource intensive to carry out.

Acknowledgments

This project has been funded through the U.S. National Library of Medicine via grant R01LM013240. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, pages arXiv–2407.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–20.

Jennifer D'Souza and Vincent Ng. 2014. Ensemble-based medical relation classification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1682–1693.

John Giorgi, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25.

Travis R Goodwin and Sanda M Harabagiu. 2017. Knowledge representations and inference techniques for medical question answering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(2):1–26.

Huan He, Sunyang Fu, Liwei Wang, et al. 2022. MedTator: a serverless annotation tool for corpus development. *Bioinformatics*, 38(6):1776–1778.

Yan Hu, Xu Zuo, Yujia Zhou, et al. 2026. Information extraction from clinical notes: are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, page ocaf213.

Ming-Siang Huang, Jen-Chieh Han, Pei-Yen Lin, et al. 2024. Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource. *Briefings in Bioinformatics*, 25(3):bbae132.

Alon Jacovi, Avi Caciularu, Omer Goldman, et al. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084.

Yuhang Jiang and Ramakanth Kavuluru. 2023. End-to-end n-ary relation extraction for combination drug therapies. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 72–80. IEEE.

Yuhang Jiang and Ramakanth Kavuluru. 2025. Relation extraction with instance-adapted predicate

descriptions. In *AMIA annual symposium proceedings*, pages 546–555.

Ramakanth Kavuluru, Anthony Rios, and Tung Tran. 2017. Extracting drug-drug interactions with word and character-level recurrent neural networks. In *2017 IEEE International Conference on Healthcare Informatics*, pages 5–12.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, et al. 2011. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, 12(1):486.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, et al. 2012. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Shengyu Liu, Kai Chen, Qingcai Chen, et al. 2016. Dependency-based convolutional neural network for drug-drug interaction extraction. In *2016 IEEE international conference on bioinformatics and biomedicine*, pages 1074–1080.

Yuan Luo, Özlem Uzuner, and Peter Szolovits. 2017. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, 18(1):160–178.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1105–1116.

Yifan Peng, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Extracting chemical–protein relations with ensembles of svm and deep learning models. *Database*, 2018:bay073.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, et al. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

Ana Claudia Sima, Tarcisio Mendes de Farias, Maria Anisimova, et al. 2021. Bio-SODA: enabling natural language question answering over knowledge graphs without training data. In *Proceedings of the 33rd International Conference on Scientific and Statistical Database Management*, pages 61–72.

Brett R South, Shuying Shen, Robyn Barrus, et al. 2011. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In *AMIA Annual Symposium Proceedings*, volume 2011.

Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, et al. 2022. A dataset for n-ary relation extraction of drug combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 3190–3203.

Tung Tran and Ramakanth Kavuluru. 2019a. Distant supervision for treatment relation extraction by leveraging MeSH subheadings. *Artificial Intelligence in Medicine*, 98:18–26.

Tung Tran and Ramakanth Kavuluru. 2019b. Neural metric learning for fast end-to-end relation extraction. *arXiv preprint arXiv:1905.07458*.

Özlem Uzuner, Brett R South, Shuying Shen, et al. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Qiang Wei, Zongcheng Ji, Yuqi Si, et al. 2020. Relation extraction from clinical narratives using pre-trained language models. In *AMIA annual symposium proceedings*, volume 2019.

Guangzhi Xiong, Eric Xie, Corey Williams, et al. 2025. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7849–7857.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 50–61.