Clinical Contradiction Detection

Anonymous ACL submission

Abstract

Detecting contradictions in text is essential in determining the validity of the literature and sources that we consume. Medical corpora are riddled with conflicting statements. This is due to the large throughput of new studies and the difficulty in replicating experiments, such as clinical trials. Detecting contradictions in this domain is hard since it requires clinical expertise. In this work, we present a distant supervision approach that leverages a medical ontology to build a seed of potential clinical contradictions over 22 million medical abstracts. As a result, we automatically build a labeled training dataset consisting of paired clinical sentences that are grounded in an ontology and represent potential medical contradiction. The dataset is used to weakly-supervise state-of-the-art deep learning models showing significant empirical improvements across multiple medical contradiction datasets.

1 Introduction

Determining whether a pair of statements is contradictory is foundational to fields including science, politics, and economics. Detecting that statements contradict can shed light on fundamental issues. For instance, mammography is an integral routine in modern cancer risk detection, but there is conflicting material about its efficacy (Boyd et al., 1984). Recognizing that a certain topic has opposing points of view, signifies that this issue may deserve further investigation. Medicine is a particularly interesting domain for contradiction detection, as it is rapidly developing, of high impact, and requires an in-depth understanding of the text. According to the National Library of Medicine, the PubMed (Canese and Weis, 2013) database averaged 900k citations for the years 2018-2021, with a quickly growing trajectory (med, 2022). The publication of contradictory papers is not uncommon in scientific research, as it is part of the process of validating or refuting hypotheses and advancing knowledge in a field. A study on high impact clinical research found that 16% of established interventions had their outcome refuted (Ioannidis, 2005). Extrapolating this to PubMed, over 5 million articles would disagree with a previous finding.

The problem of contradiction detection in text has been studied in the task of natural language inference (NLI). This task was developed to tackle the problem of recognizing whether a pair of sentences are contradictory, entailing, or neutral. Deep learning approaches have reached impressive results for this task. Specifically, large models with hundreds of millions of parameters such as DeBERTa (He et al., 2020) and BioELECTRA (raj Kanakarajan et al., 2021), are considered the state-of-the-art (SOTA) for this task. However, in medical research, defining and detecting a contradiction is more difficult. Sometimes more context is needed to detect contradiction due to the difficulty of the material. Consider the example below:

1. "However, in the valsartan group, significant **improvements in left ventricular hypertrophy and microalbuminuria** were observed."
2. "Although a bedtime dose of doxazosin can significantly **lower the blood pressure**, it can also **increase left ventricular diameter**, thus **increasing the risk of congestive heart failure.**"

Detecting that this pair contradicts requires knowing that *improvements in left ventricular hypertrophy* is a positive outcome, whereas an *increase [in] left ventricular diameter* is negative outcome with regards to heart failure.

To tackle natural language understanding tasks using deep learning methods, large datasets are required (Conneau et al., 2017). However, few datasets exist to train such algorithms in the clinical contradiction domain. Time and cost of labeling complex medical corpora, could be a potential reason for this. The MedNLI dataset (Romanov and

Shivade, 2018) for instance, required the expert labeling of 4 clinicians over the course of 6 weeks [1]. Yet, MedNLI is fabricated since each of the clinicians was given a clinical description of a patient and came up with a contradicting, entailing, and neutral sentence to pair up with that description. However, in this work we are interested in naturally occurring sentences in clinical literature as opposed to manually curated texts. Specifically, we focus on sentences representing clinical outcomes and attempt to identify whether they contradict.

One of the approaches to overcome the lack of large enough data is distant supervision (Mintz et al., 2009). Distant supervision is used for training machine learning models on a large corpus of data without manual annotation. It works by using existing knowledge sources (such as a database) to automatically label a large amount of data. The quality of the labels can be noisy, so the goal is to train models that are robust and can still learn meaningful patterns. We propose a novel methodology leveraging distant supervision and a clinical ontology - the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT or SNOMED for short) (Stearns et al., 2001). SNOMED is developed by a large and diverse group of medical experts (Donnelly et al., 2006) and it contains extensive information about clinical terms and their relationships. Our methodology uses knowledge extracted from SNOMED to classify pairs of "naturally occurring", potentially contradictory sentences. PubMed's database of medical abstracts is our source for naturally occurring sentences.

We perform empirical evaluation over multiple manually labeled clinical contradiction datasets. We fine tune SOTA deep learning models on the aforementioned ontology-driven created dataset. The results demonstrate that the distant-supervision-based methodology we propose yields statistically significant improvements of the models for contradiction detection. The average results of 9 different models see an improvement on our main evaluation set (Section 4.1.1) over previous SOTA. Specifically, we find that the improvement is consistent across both small models and those that are considered to be SOTA on NLI tasks, which is the closest task to that of contradiction detection.

The contribution of our work is threefold: (1) We present the novel problem of contradiction analysis of naturally occurring sentences in clinical data. (2) We create a clinical contradiction dataset by using distant supervision over a clinical ontology, yielding improvements of SOTA deep learning models when fine-tuning on it. (3) We empirically evaluate numerous manually labeled clinical contradiction datasets showing improvements of SOTA models when fine-tuned on the ontology-driven dataset.

[1]To access MedNLI, users must be MIMIC-III certified.

2 Related Work

The field of NLI primarily focuses on textual entailment, starting with the RTE challenges proposed by Dagan et al. (2013) and Dagan et al. (2005). The task involves determining if the meaning of one sentence can be inferred from another. Over time, new data and classification criteria have been introduced, including the labeling of contradictions in the third challenge (Giampiccolo et al., 2007). However, the medical domain brings additional challenges requiring clinical expertise.

Despite the complexity of medical literature and the reality of contradictions in publications, there is surprisingly little work in this area. Large NLI corpora contain relatively easy contradiction pairs, partly due to the cost of annotating complex contradictions. The contradiction is often a negation through words like 'not'. An example from a large NLI corpus, MultiNLI (Williams et al., 2017) is:

1. "**Met my** first girlfriend that way."
2. "**I didn't meet my** first girlfriend until later."

Scientific fact-checking is a related task, where a claim is verified against evidence (Wadden et al., 2020; Sarrouti et al., 2021). The work in this field does not deal with direct contradiction detection between two pairs of naturally occurring sentences in medical literature, but rather a popular claim which is justified by evidence coming from a medical source. In the case of (Kotonya and Toni, 2020) the health data comes from popular sources of media such as the Associated Press and Reuters News, as opposed to medical literature.

Alamri and Stevenson (2016) developed a dataset labeled for contradictory research claims in abstracts related to cardiovascular medicine. This corpus has complex sentence-pairings and is annotated by experts in the field. There are works which address contradiction of a clinical query and a claim. Given a sentence and a question, Tawfik and Spruit (2018) use a combination of handcrafted features to build a classifier, whereas (Yazi et al., 2021) use pure deep neural networks (DNN). Unlike these approaches, we focus on classifying

any given pair of medical sentences representing a clinical outcome. To our knowledge, no work addresses contradiction detection between naturally occurring sentences in clinical literature.

Following the distant supervision work of Mintz et al., Nguyen and Moschitti extended it to larger knowledge bases like YAGO. Since then, distantly-supervised relation extraction shifted into improving performance through neural networks (Zeng et al., 2015; Zhang et al., 2019). We propose to leverage distant supervision for the task of identifying contradictions between clinical sentence-pairs representing clinical outcomes. This is done by weakly-supervising SOTA deep learning models during fine-tuning and using the relational knowledge of a clinical ontology. Unlike common distant supervision approaches (Smirnova and Cudré-Mauroux, 2018; Purver and Battersby, 2012), we have unknown relation labels. Instead, we use the structure and attributes of a clinical ontology to infer whether terms contradict. In addition, our setup provides positive and negative term-relationships, unlike the classic distant supervision models (Smirnova and Cudré-Mauroux, 2018). To our knowledge, we are first to use distant supervision for contradiction detection in the clinical realm.

3 Methods

We aim to create a model for classifying whether clinical outcomes contradict. We focus on nontrivial examples requiring deep subject understanding. This model brings awareness to conflicting findings in medicine. Locating disagreement can elicit further investigations or general consciousness.

3.1 SNOMED CT Ontology

SNOMED is an ontology containing over 350,000 clinical terms (Stearns et al., 2001). The terminology has information about a plethora of health concepts, containing useful attributes such as relationships to other terms and interpretations. The structure of SNOMED allows us to group terms based on their relationships. We hypothesize that using this structure coupled with synonyms and antonyms, will enable us to create a corpora of contradicting and non-contradicting clinical terms. We use the 2022 SNOMED version in this work.

3.1.1 SNOMED Node Attributes

Each term in the SNOMED ontology is a node in a tree-like structure. A subset of these nodes have useful attributes which we use to determine their relationships. Each of these nodes belongs to a group parented by the group root. In addition, each node has a simple interpretation which is a defined attribute within the ontology. In Figure 1, the group consists of nodes describing the group root *cardiac output*. The green (right) node, *increased cardiac output*, has the interpretation - *increased*.

We claim that groupings of terms with these attributes have a logical connection. Pairing up child nodes yields a combination of contradicting and non-contradicting pairs of phrases. Determining the relationship between a pair of SNOMED terms is done partially through comparing their interpretations. In Figure 1 the left node has the interpretation *decreased*, whereas the right node has the interpretation *increased*. We assign the pair an *attribute* label ($A_{i,j}$) of contradiction. In Algorithm 1, $A_{i,j}$ is assigned on Line 12.

The size of the groupings can get large. For instance, the group root *Cardiac function* has 275 children. Since *cardiac function* is very general, its child terms may not be related - for example the terms *aortic valve regurgitation due to dissection* and *dynamic subaortic stenosis*. Both terms are impairments of *cardiac function*, but it would not be fair to claim that the two are related outcomes. Though these large groupings can yield many pairings of phrases, we see why they may also be less accurate. Some of this testing is in Section 5.2, where we investigate the effects of group sizes.

Below are pairings of contradictions in various medical domains that our methodology yields:

- suppressed urine secretion ↔ polyuria
- elevation of SaO2 ↔ oxygen saturation within reference range
- joint stable ↔ chronic instability of joint

3.1.2 Synonyms

After exploiting ontological structure, we consider linguistic elements. Although synonyms and antonyms do not always indicate whether sequences of words are contradictory, they provide a strong signal in our structural construction. Since clinical terms are already grouped, we know that all the terms in a grouping share a context, thereby allowing the use of simpler indicators to determine their relationship. We word-tokenize each clinical phrase, removing the intersection of the sets of tokens, leaving each with its unique tokens. A detailed visualization is found in Appendix B.1.

Algorithm 1 SNOMED Traversal

```
1:  function TRAVERSE(root)
2:      for n ∈ root.children do
3:          if n.num_childs ≤ group_size
4:              pairs ← DET_RELATION(n)
5:          end if
6:      end for
7:      return pairs
8:  end function

9:  function DET_RELATION(n)
10:     pairs ← {}
11:     for c_i, c_j ∈ n.child_pairs do
12:         A_{i,j} ← GET_ATTR_LABEL(c_i, c_j)
13:         S_{i,j} ← GET_SYN_LABEL(c_i, c_j)
14:         label_{i,j} ← A_{i,j}
15:         if S_{i,j} = contra or A_{i,j} = contra
16:             label_{i,j} ← contra
17:         end if
18:         pairs ← pairs ∪ {(label_{i,j}, c_i, c_j)}
19:     end for
20:     return pairs
21: end function

22: SNOMED ← TRAVERSE(root)
23: FINETUNE(Model, SNOMED)
```

3.1.3 Combining Attributes and Synonyms

To optimally combine $A_{i,j}$ and $S_{i,j}$ to form a final $label_{i,j}$, we build a validation set of the publicly available SNOMED term-pairs. Two human annotators with domain knowledge in the field labeled 149 SNOMED phrase-pairs - 70 of which were contradictory and 79 as non-contradictory. More details can be found in Appendix A.1. We find that when $A_{i,j}$ indicates contradiction, then it's highly likely that $label_{i,j}$ is a contradiction. The same is true for $S_{i,j}$. We define the explicit logic in Lines 15 through 17. We reach 79% accuracy through using this heuristic on the human-labeled SNOMED term-pairs with a Cohen's kappa coefficient of 0.853 for inter-annotator agreement.

3.2 Ontology-Driven Distant Supervision

Using the relational knowledge extracted from SNOMED, we weakly-supervise naturally occurring sentences in PubMed to build our SNOMED dataset. We fine-tune on this dataset to achieve significant improvements over existing baselines. Algorithm 1 summarizes the procedure. We use the 2022 PubMed version in this work. We search PubMed for sentences containing the phrase-pairs discussed in Section 3.1, resulting in a corpus of pairs of sentences. The sentence-pairs are then labeled through distant supervision as explained below. For a given pair of SNOMED terms (p_1, p_2), we label sentences (s_1, s_2) as formalized in Eq.1, where $label \in \{\text{contradiction, non-contradiction}\}$.

$$(p_1 \in s_1) \wedge (p_2 \in s2) \wedge ((p_1, p_2) \in label) \quad (1)$$

The assumption in this methodology is that if outcomes contradict, then the sentence pair is likely also contradictory. However, this ignores the possibility that there may be differing interventions or participants. More concretely, p_i may be a subset of s_i, so there may be information loss (statistics on average sizes of s_i are reported in Table 2). Given that the ultimate purpose of the SNOMED dataset is to increase the amount of training or fine-tuning data of a model, we find that this introduced noise is acceptable and still yields positive results.

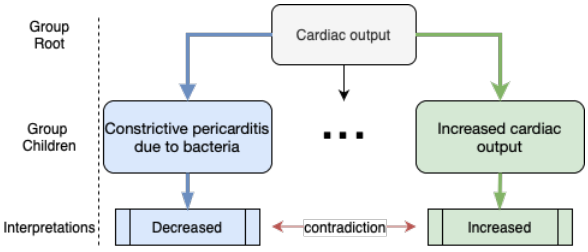

Figure 1: The group with *Cardiac output* as its root. The children depicted have contradicting interpretations.

3.3 Filtration

Naively, we pair-up any sentences satisfying Eq. 1, independent of whether they share context. Although two sentences contain their respective clinical SNOMED terms, they may be unrelated. The sentence-pair below exhibits this:

1. "The present results suggest that the upstream changes in blood flow are transmitted by the velocity **pulse faster** than by the pressure pulse in the microvasculature."
2. "His chest wall was tender and his **pulse slow** but the remainder of his physical examination was normal."

The bolded clinical terms are central to the meaning of the sentences and are independently contradictory. However, when placed in context they may be less relevant to each other as in the example above. We experiment through imposing stricter criteria for filtering sentence matches - namely

MeSH (Medical Subject Headings) terms criteria (Lipscomb, 2000) and cosine similarity criteria.

MeSH terms categorize articles within PubMed and come from 2022 PubMed release. We hypothesize that sentences drawn from articles with related MeSH terms, likely discuss the same topic. Eq. 2 is our formulation for filtering via MeSH terms. $MeSH_i$ and $MeSH_j$ are the sets of MeSH terms for articles containing $sent_i$ and $sent_j$ respectively. Let t be a chosen threshold.

$$\mathbf{1}_A := \begin{cases} 1 & \text{if } \frac{|MeSH_i \cup MeSH_j|}{min(|MeSH_i|,|MeSH_j|)} \geq t \,, \\ 0 & otherwise \end{cases} \quad (2)$$

MeSH terms are powerful, but not perfect. The following sentence-pair achieves a score of 0.4 per the inequality in Eq. 2.

1. "In dogs challenged with endotoxin, the inhibition of nitric oxide production **decreased cardiac index** and did not improve survival."
2. "Intra-aortic balloon pumping **increased cardiac index** and aortic distensibility by 24% and 30%, respectively, and reduced myocardial oxygen demand by 31% ($P < .001$ for all alterations)."

Despite overlap in MeSH terms, they are very different - one discusses dogs and the other humans.

The second filtration method measures the cosine similarity between one-hot vectors. Topically related sentences should have a higher one-hot vector cosine similarity. Let $\mathbf{o}_i$ and $\mathbf{o}_j$ be the respective one-hot vectors of $sent_i$ and $sent_j$. Vector lengths are equal to the number of unique words spanning the sentence-pair. We compute the similarities between the vectors as shown in Eq. 3. The dog example above, yields a similarity score of 0.2.

$$\mathbf{1}_A := \begin{cases} 1 & \text{if } cosine(\mathbf{o}_i, \mathbf{o}_j) \geq t \,, \\ 0 & otherwise \end{cases} \quad (3)$$

We experiment with t, ultimately choosing $t = 0.35$ based on an external validation set.

4 Empirical Evaluation

In this section we discuss the medical corpora used in our evaluation of 9 different models.

4.1 Evaluation Datasets

4.1.1 Cardiology Dataset

Due to the difficulty of labeling medical data, there are few datasets labeled for medical contradictions. To evaluate the SNOMED dataset quality, we tweak ManConCorpus (Alamri and Stevenson, 2016), a corpus of potentially contradictory cardiovascular claims. The corpus consists of question-claim pairs. Each question has 'yes', 'no' claims. The claims naturally occur in PubMed and the questions are generated by experts. We convert ManConCorpus by pairing up claims, since we are strictly interested in naturally occurring sentences from PubMed. A pair is labeled as contradictory if each constituent claim answers the question differently. We coin this dataset as *Cardio* (see Table 1 for details).

Table 1: Cardiology Dataset Breakdown

Split	Total	Contra	Non-Contra
Train	1347	571	776
Dev	198	100	98
Test	227	55	172

4.1.2 Hard Cardiology Dataset

Through our analysis, we find that models tend to classify sentence-pairs as contradictory if negation words appear. For example:

1. "Our results indicate that atorvastatin therapy significantly improves BP control in hyperlipidemic hypertensive patients."
2. "Administration of a statin in hypertensive patients in whom blood pressure is effectively reduced by concomitant antihypertensive treatment **does not have** an additional blood pressure lowering effect."

We construct a version of Cardio through removing negation words. As expected, this version exposes some weaknesses of the models, since negation words are not deemed as important.

4.1.3 MedNLI Datasets

Inspired by SNLI (Bowman et al., 2015), MedNLI was created with a focus on the clinical domain (Romanov and Shivade, 2018). The dataset was curated over the course of six weeks, borrowing the time of four doctors. MedNLI consists of sentence-pairs which are grouped into triples - a contradictory, entailing, and neutral pair. The sentences are not naturally occurring in existing medical literature. The premise is shared across the three pairs, but each have a different hypothesis, yielding a different label. Since MedNLI deals with a 3-class problem, we relabel the dataset by making {*entailment, neutral*} map to *non-contradiction*.

Our focus is to show that the SNOMED dataset, which requires no expert intervention or expenses, is as powerful as the curated MedNLI dataset.

We find that the baseline on the relabeled version of MedNLI gives high results (Appendix E), so adding additional data makes little change. The largest labeled datasets containing naturally occurring sentences are at most hundreds of sentences. Therefore, we randomly sample 100 instances from MedNLI's train-split and report results on that.

To explore fields outside of cardiology, we create versions of MedNLI focused on gynecology (GN), endocrinology (Endo), obstetrics (OB), and surgery. To filter the data, we use the help of the same annotator introduced in Section 3.1.3. We sample from the train-split in the same fashion as explained above. Note that these datasets also have the same 2-class label structure as explained in Section 4.1.3 (see details in Appendix A.2).

4.2 SNOMED Dataset Analysis

Table 2 presents general statistics of the SNOMED dataset used for weak supervision. Specifically, the total number of articles in SNOMED analyzed and the number of sentences in PubMed containing a term from the SNOMED ontology. We now analyze difference noise sources of the dataset.

4.2.1 Phrase Matching Noise

The proposed phrase-matching introduces noise when p_i is not central to the meaning of s_i. To approximate this noise, we sample 100 sentences from the SNOMED dataset. A human annotator was asked to evaluate if p_i contributes to the central message of s_i. We observed a 91% accuracy.

4.2.2 SNOMED Labeling Noise

The automatic nature of the SNOMED dataset labeling may also introduce noise. Similar to (Mintz et al., 2009), we sample 100 instances from the dataset and manually label the sentence-pairs. The annotator is told to label each sentence pair as containing contradictory elements or non-contradictory. This gold label is compared to the distantly-supervised label. We extract both positive and negative relations from our ontology, thus we report accuracy to indicate the effectiveness of our methodology. We observe 82% accuracy. This is higher than the noise analysis of other weakly supervised datasets (Mintz et al., 2009), likely due to increased amount of information in ontologies.

4.3 Baseline Models

Yazi et al. (2021) achieve the SOTA on the ManConCorpus, which we turn into the Cardio corpus as explained in Section 4.1.1. They concatenate BERT embeddings for their question and claim, feeding this input into a multi-layer feed forward network. Our baselines do not use a siamese network, instead we feed in our sentence-pairs as input into the network. Our evaluation consists of 9 baseline models and comparing their performance when they are fine-tuned on the SNOMED dataset versus without. The task of classifying contradiction is most similar to NLI, so some of these baseline models are those that top leaderboards for the MNLI and MedNLI datasets - namely DeBERTaV3-Base (He et al., 2021), ALBERT (Lan et al., 2019), and BioELECTRA (raj Kanakarajan et al., 2021). ELECTRA (Clark et al., 2020) and BERT (Devlin et al., 2018) are also included as they are generally high-performing architectures. In addition, we are interested in seeing the performance of small models. They require less computing resources and may allow the SNOMED dataset to have a stronger influence during fine-tuning. Thus, we also include BERT-Small (Turc et al., 2019), ELECTRA-Small, and DeBERTaV3-Small (He et al., 2021). Finally, we include BioGPT (Luo et al., 2022) for completeness, as it has a decoder architecture and is also pre-trained on biological data. Table 3 contains a breakdown of the number of parameters per model.

All the baseline models are pre-trained on large corpora. The high-level architecture of the models is the same, so we use the functionalities of HuggingFace (Wolf et al., 2019) and the Sentence-Transformer library (Reimers and Gurevych, 2019). We add an uninitialized binary classification head to the model body. All hyperparameters come from the Sentence-Transformer library, except for training batch size - 8 for models above 30M parameters and 16 for models under 30M parameters.

Each baseline is tuned with the SNOMED dataset. The SNOMED dataset we create uses a group size of 25, sampling 10 sentence-pairs from

Table 2: SNOMED Dataset

Sentence length:	
NLTK token count	25.1
BioGPT token count	29.4
BioELECTRA token count	30.7
BERT-Base token count	36.8
Total Dataset Statistics:	
SNOMED term matches in PubMed	4.99M
Number of articles	2.87M
Number of qualifying pairs in SNOMED	0.63M

PubMed for every SNOMED term-pair. These hyperparameters are determined through ablation tests on the Cardio validation set.

5 Empirical Results

We explore the significance of the SNOMED dataset we create via our methodology (Section 3) and gather insights through ablation tests.

5.1 Main Result

Table 3 summarizes our main findings. We compare the performance of the baseline algorithms when fine-tuned over the original training split of each dataset (marked as *Base*) versus tuning using both the novel SNOMED dataset and *Base* (marked as *Ours*). We measure the area under the ROC curve of each baseline, and verify statistical significance through Delong's test (DeLong et al., 1988). Significant differences are marked with an asterisk (*). We observe that across all dataset the weak supervision over the SNOMED dataset reached superior results compared to fine tuning only on the original dataset and outperforms the SOTA model for contradiction detection (Yazi et al., 2021).

Cardio is a relatively difficult dataset of potentially contradicting, naturally occurring pairs in PubMed. The sentences are complex and require a deep medical understanding. We observe that fine tuning on the SNOMED dataset improves the baselines for 8 out of 9 models we evaluate.

The performance on Hard-Cardio drops relatively to Cardio as expected. This verifies our hypothesis that removing negations makes the problem more difficult. Further, 8 models fine-tuned on SNOMED outperform their baseline counterparts.

We observe that even on synthetically created common datasets, such as MedNLI sentences, our methodology improves over *all* baselines for this corpus. We observe a similar trend when focusing on various sub-specialties. The improvements are consistent across almost *all* models when fine-tuning on SNOMED. This enables us to learn of the scalability of our methods for clinical contradiction detection through different fields within healthcare.

Analyzing our findings further, we see that there is a trend that smaller models are generally more affected by fine-tuning on SNOMED. All of the evaluation datasets improve over the baseline on *every* model under 30 million parameters.

5.2 Ablation Studies

Below we review ablation studies to find the impact of various system parameters on performance.

5.2.1 Group and Sentence Samples Size

We explain SNOMED term grouping in Section 3.1 and illustrate in Figure 1. Group size and pairing quality may be closely related. Larger groupings tend to have more terms which are less related to each other as explained in Section 3.1.1. Thus, we experiment with SNOMED datasets based on terms belonging to groups of at most 6, 12, 25, and 50.

During dataset creation, we choose how many sentence-pairs to sample per SNOMED pairing. In Figure 2, each line with a different color/marker represents a different number of samples averaged across all 8 models. The ablations we perform include 10, 25, and 50 samples per pairing.

Figure 2 shows 10 samples outperforms higher sampling numbers for almost all group numbers. Increased sampling results in over-saturation of certain term-pairs. This may result in overfitting. The best group size is 25 for small models and 12 for large models. These numbers strike the balance of creating a large amount of SNOMED phrase-pairings, while keeping their relationships accurate (as discussed in Section 3.1.1). Smaller models may benefit more from larger group sizes, because they have a more limited base knowledge than those of large models.

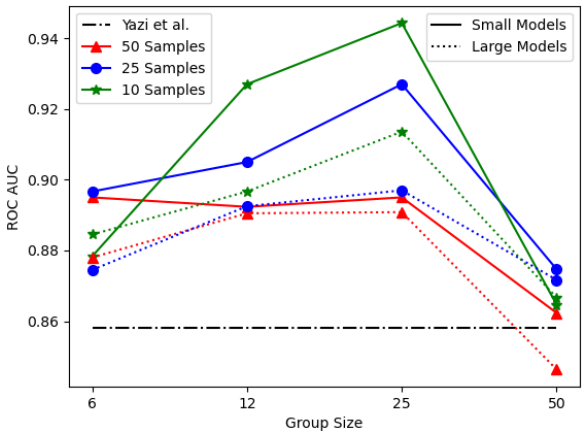

Figure 2: Small and large model performance across group sizes and sample numbers. Reported on Cardio.

5.2.2 Filtering Based on Similarity

To increase the chances that sentences are related, when sampling phrases from PubMed, we experiment with keeping pairs that exhibit high MeSH term or cosine similarity as explained in Section

Table 3: Performance of Models tuned with SNOMED vs. Without

		Algorithm (Number of Params)									
Dataset	Method	ALBERT Base (11.7M)	ELECTRA Small (13.5M)	BERT Small (28.8M)	ELECTRA Base (109.5M)	BERT Base (109.5M)	Bio-ELECTRA (109.5M)	DeBERTa Small (141.9M)	DeBERTa Base (184.4M)	Bio—GPT (346.8M)	(Yazi et al., 2021)
Cardio	Base	0.911	0.877	0.858	0.863	**0.914**	0.880	0.885	0.861	0.858	0.858
	Ours	**0.928**	**0.947***	**0.958***	**0.892**	0.878	**0.925**	**0.931***	**0.942***	**0.930***	-
Hard-Cardio	Base	0.876	0.785	0.717	0.847	**0.803**	0.850	0.842	0.845	0.762	0.687
	Ours	**0.925***	**0.853***	**0.794***	**0.873**	0.791	**0.925***	**0.917***	**0.936***	**0.871***	-
MedNLI-General	Base	0.609	0.559	0.587	0.602	0.752	0.585	0.581	0.704	0.816	0.518
	Ours	**0.817***	**0.721***	**0.718***	**0.791***	**0.816***	**0.820***	**0.735***	**0.878***	**0.850**	-
MedNLI-Cardio	Base	0.749	0.542	0.553	0.600	0.759	0.597	0.589	0.672	**0.840**	0.557
	Ours	**0.808**	**0.648***	**0.692***	**0.785***	**0.794**	**0.834***	**0.777***	**0.864***	0.833	-
MedNLI-GYN	Base	0.492	0.533	0.600	0.525	0.575	0.558	0.592	0.625	0.583	0.508
	Ours	**0.608**	**0.617**	**0.767**	**0.758**	**0.792**	**0.808**	**0.683**	**0.825**	**0.783**	-
MedNLI-Endo	Base	0.698	0.525	0.567	0.584	0.639	0.601	0.522	0.601	0.840	0.560
	Ours	**0.860***	**0.690**	**0.725**	**0.793***	**0.867***	**0.860***	**0.852***	**0.883***	**0.878**	-
MedNLI-OB	Base	0.532	0.502	0.513	0.505	0.557	0.549	0.502	0.579	0.507	0.505
	Ours	**0.616**	**0.542**	**0.581**	**0.667**	**0.625**	**0.702***	**0.618**	**0.740***	**0.630**	-
MedNLI-Surgery	Base	0.708	0.502	0.555	0.681	0.842	0.669	0.576	0.691	0.925	0.602
	Ours	**0.892***	**0.668**	**0.807***	**0.912***	**0.903**	**0.925***	**0.808***	**0.884***	**0.940**	-

3.2. Figure 3 shows the relationship between the filtration methods discussed above. As a continuation of the ablation visualized in Figure 2, we fix the number of samples to be 10 and the group size to be 25. The cosine methodology outperforms both the naive version (no filtering) and MeSH. Although MeSH terms are useful, it is possible that since they are tagged on an article-level, they cannot provide the same topic granularity as the one-hot vectors.

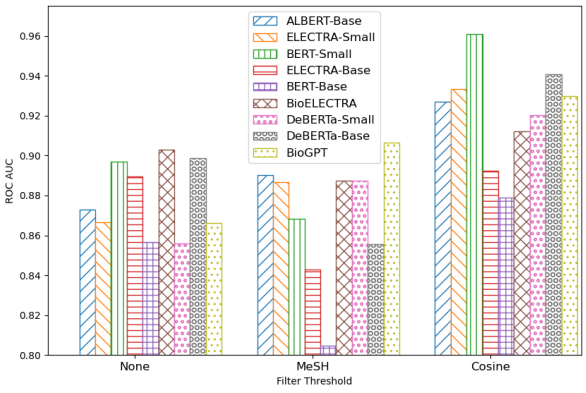

Figure 3: Performance across filtration methods. Number of samples is 10 and group size is 25. Reported on Cardio.

6 Conclusions

Contradiction detection is central to many fields, but is especially important in medicine due to human impact. With the rapid growth of the field, clinical research is exploding with findings as demonstrated by the growth of PubMed. Although contradictions are a subfield of NLI, there is less exploration in the clinical domain. Often, contradictions within medicine are more complex than other fields due to the need of additional context and domain knowledge. Labeling datasets which produce high results with deep learning models are costly.

We introduce a novel methodology of using a clinical ontology to weakly-supervise the creation of a contradiction dataset with naturally occurring sentences. We coin it the SNOMED dataset. The empirical results suggest that fine-tuning on the SNOMED dataset results in consistent improvement across SOTA models over diverse evaluation datasets spanning multiple medical specialties. We show that a balance exists between term group size and the number of sentences sampled from PubMed per pairing. In addition, we find that we can further improve results through filtering which PubMed sentences we include in our dataset.

For future exploration we suggest investigating more robust sentence filtration methods, such as topic modeling or sentence embedding similarity. Looking into how other clinical ontologies can be paired with SNOMED may also be fruitful.

This methodology is limited to SNOMED terms, many of which do not appear within PubMed. Due to the evolving nature of knowledge bases, terminology and information changes, potentially altering relations between terms. Finally, the structure we extract from the clinical ontology is not ground-truth, yielding noise during dataset creation.

Ethical Considerations

Whenever working within the clinical domain, ethical considerations are crucial. The data that we work with is all rooted in already publicly available corpora and PubMed. To the best of our knowledge the data we use does not contain any personal information of any humans involved in clinical trials. There is a potential risk of over representing common diseases and outcomes in our dataset, thereby not including enough data about other outcomes.

References

2022. Medline® citation counts by year of publication (as of january 2022)*.

Abdulaziz Alamri and Mark Stevenson. 2016. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics*, 7(1):1–9.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Norman F Boyd, Brian O'Sullivan, Eve Fishell, Imre Simor, and Gabriel Cooke. 1984. Mammographic patterns and breast cancer risk: methodologic standards and contradictory results. *Journal of the National Cancer Institute*, 72(6):1253–1259.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

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

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John PA Ioannidis. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Truc-Vien T Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 277–282.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Noha S Tawfik and Marco R Spruit. 2018. Automated contradiction detection in biomedical literature. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 138–148. Springer.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Fatin Syafiqah Yazi, Wan-Tze Vong, Valliappan Raman, Patrick Hang Hui Then, and Mukulraj J Lunia. 2021. Towards automated detection of contradictory research claims in medical literature using deep learning approach. In *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 116–121. IEEE.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. *arXiv preprint arXiv:1903.01306*.

A Annotation

As mentioned in Section 3.1.3, we work with annotators with domain knowledge. The annotators were used due to their expertise in the field.

A.1 SNOMED Term-Pairs

The annotators labeled 149 SNOMED term-pairs as either contradictory or non-contradictory. They were provided with a list of pairs, without any additional information about the ontological structure they came from. This was done in order to preserve fairness and integrity during the labeling process. The instructions were to come up with a binary label for each of the pairs.

A.2 Filtering MedNLI

A human annotator also helped with coming up with a list of sub-words which served as indicators for particular fields of medicine. For example, the sub-words *vulv* and *gyno*, are indicative of gynecology. These word lists were used to create the variations of MedNLI discussed in Section 4.1.3. You can find the lists of words in the code that is released with the paper.

B Additional Methodology Details

B.1 Synonym Extraction

Synonym extraction is a part of our methodology which is explained in Section 3.1.2. Figure 4 provides a depiction of this for the clinical terms *shortened p wave* and *prolonged p wave*. The respective unique tokens are *shortened* and *prolonged*. Since the unique tokens are antonyms, the *synonym* label for the pair is a contradiction. In Algorithm 1, the

Table 4: Cardio Dataset Additional Details

Sentence length:	
NLTK token count	26.7
BioGPT token count	30.6
BioELECTRA token count	31.8
BERT-Base token count	37.2

Table 5: MedNLI Additional Details

Sentence length:	
NLTK token count	13.2
BioGPT token count	16.3
BioELECTRA token count	17.1
BERT-Base token count	19.1

synonym label ($S_{i,j}$) is assigned on Line 13. Similarly, if the respective tokens are synonyms, then $S_{i,j}$ would be a non-contradiction.

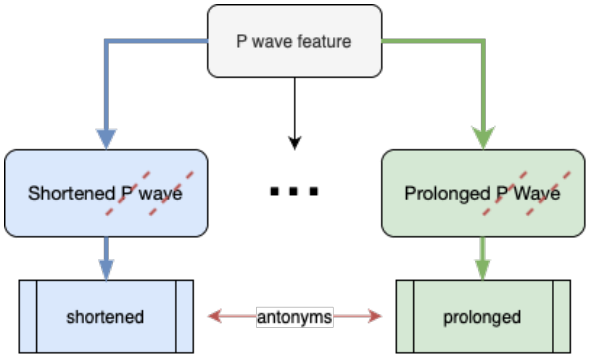

Figure 4: The terms *shortened p wave* and *prolonged p wave* are simplified to just *shortened* and *prolonged* after their common words are removed. The remaining words are antonyms.

C Additional Dataset Details

We include some additional details to the breakdown of the evaluation datasets. In particular, regarding the token lengths of the datasets. In Table 2 we include the sentence length breakdown according to various tokenizers of the SNOMED dataset. In the appendix we also include the break down of the Cardio dataset as well as the MedNLI dataset (Tables 4 and 5 respectively). Although our SNOMED dataset and Cardio dataset contain roughly the same number of tokens per sentence, the MedNLI dataset contains roughly half the number of tokens per sentence. This may serve as an indicator to the decreased difficulty of MedNLI as well as evidence that that sentences are not naturally occurring.

D SNOMED Dataset Examples

We include several randomly sampled examples from the SNOMED dataset. The data is also publicly available.

D.1 Contradiction Examples

Example:

Sentence 1: the plasma cck and luminal content of lcrf were measured by specific radioimmunoassays.;bile-pancreatic juice diversion significantly increased pancreatic secretion plasma cck and lcrf levels.

Sentence 2: blockade of the cck receptor results in decreased pancreatic secretion and atrophy.

Example:

Sentence 1: although the mutant does not swim still it is able to move and perform photobehavior.

Sentence 2: whereas the chev mutants still produced both types of flagella and were able to swim and swarm.

D.2 Non-Contradiction Examples

Example

Sentence 1: computed tomography (ct) scans showed bilateral contracted kidneys with a mass projecting from the lower pole of the right kidney.

Sentence 2: ultrasonography and computed tomography revealed a masslike expansion involving the upper pole of an otherwise small right kidney.

Example

Sentence 1: hearing loss tinnitus hyperacusis and difficulty hearing in noise remain persistent and in some cases progressive complaints for patients.

Sentence 2: chief complaints were longstanding localized pain and hearing difficulty.

E Full MedNLI Dataset

For completeness, we also report results on the full MedNLI dataset. Results can be see in Table 6. Notably, there are no statistically significant results. Although fine-tuning with the SNOMED dataset yields better results in majority of the models, we see that results are roughly the same. Therefore, we hypothesize that there is over-saturation which occurs at this stage.

Table 6: Performance of Models tuned with SNOMED vs. Without

		Algorithm (Number of Params)									
Dataset	Method	ALBERT Base (11.7M)	ELECTRA Small (13.5M)	BERT Small (28.8M)	ELECTRA Base (109.5M)	BERT Base (109.5M)	Bio-ELECTRA (109.5M)	DeBERTa Small (141.9M)	DeBERTa Base (184.4M)	Bio—GPT (346.8M)	(Yazi et al., 2021)
MedNLI	Base	0.946	0.934	**0.936**	0.962	0.951	0.973	**0.968**	**0.977**	**0.962**	0.934
	Ours	**0.951**	**0.934**	0.933	**0.962**	**0.952**	**0.973**	0.966	0.971	0.961	-