ClinScope Corpus - Clinical Notes Annotated for Hedge and Negation

Anonymous ACL submission

Abstract

There has been great interest in knowledge extraction from biomedical texts. Part of this research involves hedge and negation assertion detection as doctors often use these assertions during the diagnostic process to specify likelihood or ruling out other possible diseases and conditions. Although natural language processing has been growing rapidly in the biomedical field, available corpora for clinical free-texts are still limited with research relying on limited available corpora where many are not annotated. In addressing this issue, we propose this ClinScope Corpus, a new clinical text corpus focused negation and hedge annotations. Our sampling allows for higher concentrations of assertion cues along with their scope and medical foci to aid in detecting when cues directly negate or mark medical entities uncertain.

1 Introduction

Knowledge from clinical texts is invaluable for improving patient care, epidemic detection and management, and identifying patients eligible for research (Frankovich et al., 2011; Chapman et al., 2001a,b). However, medical reports often contain doctors' notes in narrative form (Chapman et al., 2001b), increasing the difficulty of manual data analysis. Through automated data analysis, medical professionals can quickly reference clinical notes and other texts to expedite patient care.

However, information retrieval techniques commonly do not index or take negation and hedge assertion cues into consideration (Chapman et al., 2001b). One study showed that approximately half the conditions analyzed in clinical reports were negated (Chapman et al., 2001a). For hedges, another study found that most clinical document categories have at least one hedge phrase in at least half of the associated documents (Hanauer et al., 2012). Since these cues are prevalent in clinical texts, it is vital that automation algorithms accurately detect when medical statements are negated or speculations (Lakoff, 1973).

Negation cues can be simply defined as words performing predicate denial or negating the meaning of the modified expression (Horn, 2001). They can come in multiple forms such as: 1) an affix such as *un-* in *unable*, 2) a single word such as *not*, 3) multiple words such as *rule out*, or 4) contractions, such as *don't*. Hedge cues can be one word or multiple words and are used to express uncertainty if the modified expression leans true (positive) or false (negative). Figure 1 shows both type of cues analyzed in this paper and annotations for *scope* and *medical foci*, which are medical expressions within the cues' scopes that the cues directly negate or mark as uncertain.

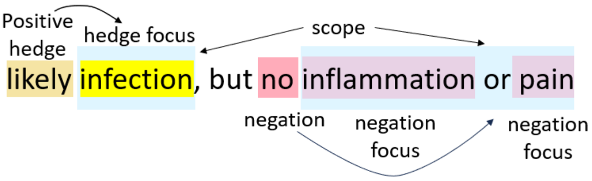

Figure 1: This example[1] demonstrates how we annotate for negation and hedge cues, scopes, and medical foci.

The objective of this paper is to introduce our new annotated clinical text corpus focused on negation and hedges as shown in Figure 1[1]. We algorithmically populated this corpus through sentence parsing and extraction from MIMIC-III's notes (Johnson et al., 2016a). This corpus also incorporates algorithmic sampling to increase the concentration of cues in this corpus. Our annotations for this dataset include labelling medical foci within scopes to align with the end goal improving identification of whether clinical observations are absent or uncertain.

[1] All provided example sentences are not directly from MIMIC-III but derived for demonstration purposes. The dataset itself will have real sentences from MIMIC-III and requires PhysioNet (Goldberger et al., 2000) access.

2 Related Work

2.1 Negation and Hedge Detection

Earlier negation research began with rule-based systems such as NegEx (Chapman et al., 2001b) and NegFinder (Mutalik et al., 2001) where both used their own dataset and a predefined set of negation terms. Morante's group (Morante, 2010) explored negation cues cited in previous works and analyzed how negation cues are used in Bioscope (Vincze et al., 2008). There has also been other work over the years for detecting negation involving dependency graphs (Slater et al., 2021), machine learning (Morante and Daelemans, 2009b; Fancellu et al., 2016; Sergeeva et al., 2019) and large language models (LLMs) (van Aken et al., 2021), where the last work also focused on hedge detection. Some other approaches for detecting hedge cues includes work using machine learning algorithms (Medlock and Briscoe, 2007; Morante and Daelemans, 2009a; Agarwal and Yu, 2010) to detect hedge cues in full-text papers from genomics and the Bioscope Corpus (Vincze et al., 2008). Hanauer's group (Hanauer et al., 2012) also analyzed the use of hedges in clinical documents from their institution's electronic health record (EHR) system.

2.2 Corpora

Currently, there are not many available clinical corpora with clinical notes as many had been pulled from public access. The most prevalent corpora is the MIMIC dataset which provides the largest amount of medical records, albeit not annotated. We list some of the clinical corpora below:

- BioScope Corpus (Vincze et al., 2008) which originally included annotated clinical free-texts (Pestian et al., 2007) (data now retracted) and also contains the Genia Corpus (Ohta et al., 2002) annotated for negation and hedges.
- i2b2 Clinical Records (Uzuner et al., 2011) - currently available through n2c2
- TREC Medical Records (Voorhees, 2013)- retracted from public use (other later datasets may be available)
- MIMIC-III (Johnson et al., 2016b) and MIMIC-IV (Johnson et al., 2023) - largest quantity of public un-annotated clinical reports

3 Information Extraction Tasks

Recent work has vastly moved past negation and uncertainty detection and focused other aspects of clinical texts and tasks (Lee et al., 2020; Shah and Mohammed, 2020; Lin et al., 2021; Lehman and Johnson, 2023; Agrawal et al., 2022; Yang et al., 2022; Eysenbach, 2023). However, the issue remains if LLMs, the current state-of-the-art, actually perform well on information extraction, especially with texts containing negated or uncertain probabilities of medical concepts. Prior research shows there is still a high need for annotated in-domain training data for negation as tested approaches have mixed results in negation detection with limited generalizability for arbitrary clinical text (Wu et al., 2014). Specialized clinical LLMs often perform better than general LLMs even when trained on limited annotated data (Lehman et al., 2023; Wornow et al., 2023). Another group found that medical pre-training improves models, but clinical language models still suffer from errors (van Aken et al., 2021). When considering our research directions, we performed preliminary experiments on existing algorithms (described in Section 4).

4 Experimental Observations on NLP Algorithms

We performed analysis of NegEx, van Aken's group's best performing clinical language model (van Aken et al., 2021), and GPT-3.5 through Microsoft Azure[2] (Boyd, 2023). All algorithms were tested using MIMIC-III data and using individual sentences and full clinical reports. The details of the experiments were omitted to conserve space, but our findings showed that although there were improvements from the initial NegEx algorithms, there is still detection sensitivity issues when it comes to denoting if a medical concept (i.e., disease) is present, absent, or uncertain for both the clinical language model and GPT-3.5. For the performance on one report, the clinical model had 63% accuracy while GPT-3 had 61% when analyzing the 51 medical entities in the report (details in Appendix A). We note that the errors can be severe - if the information about the patients' records is reported incorrectly, this can potentially lead to incorrect treatments and misdiagnoses.

[2] Microsoft Azure was chosen and used with content logging turned off to remain compliant to MIMIC-III's data use agreement.

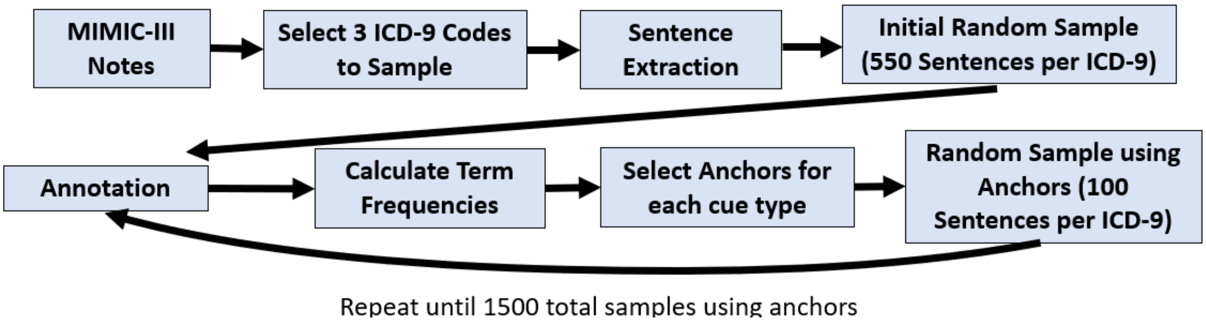
MIMIC-III Notes
Select 3 ICD-9 Codes to Sample
Sentence Extraction
Initial Random Sample (550 Sentences per ICD-9)
Annotation
Calculate Term Frequencies
Select Anchors for each cue type
Random Sample using Anchors (100 Sentences per ICD-9)
Repeat until 1500 total samples using anchors

Figure 2: Flowchart summarizing the steps used to generate the ClinScope corpus.

5 ClinScope Corpus Generation

The corpus contains 3,150 sentences extracted from MIMIC-III's clinical notes (Johnson et al., 2016a). This section details the methods used to sample and create this corpus (flowchart available in Figure 2). All sentences are also linked back to the original reports and ICD-9 codes for traceability.

5.1 Sampling

To guide the sampling, we chose three ICD-9 codes (disease codes associated with medical reports) where we considered the frequency of known cues from other works, severity of the diseases, and less similarity between the chosen diseases. The end result was the selection of these three codes: ICD-9 codes 4280 (Congestive Heart Failure not otherwise specified (NOS)), 51881 (Acute Respiratory Failure), and 5849 (Acute Kidney Failure NOS).

From there, we used an algorithm to parse the reports into sentences prior to sampling. We tested SciSpacy (Neumann et al., 2019) and our own algorithm (which uses regular expressions)[3] and found that our algorithm was comparable or better at handling section headers, numerical bulleting, medical acronyms, and other unique issues found in medical notes while being over 45 times faster (4-6 minutes for each ICD-9 code vs 4-5 hours for SciSpacy).

For the initial sample seed, we chose to randomly sample 550 sentences from each of the three ICD-9 codes with the intention of 500 samples and an additional 10% sampling to adjust for sentence extraction errors. We justified the size using statistics (Arya et al., 2012) with full details on the calculation in Appendix B.

After annotating the initial set of sentences (annotation described in Section 5.2), we use the concept of anchors (Halpern et al., 2014) to sample an additional 1500 sentences. This is performed through five rounds of sampling through choosing 3-5 new anchors for each cue type (negation, positive hedge, negative hedge) for each round. These anchor terms are chosen using frequency and likelihood of leading to a cue existing in the sentence based on the meaning of the chosen anchor term. For example, *but* was selected as an anchor as it is linguistically used for contrasting parts of a sentence and thus an increased likelihood of negation occurring in the contrast. These anchor terms are then use to select 100 sentences at random for a given ICD-9 code for each of the five rounds.

5.2 Annotation

We annotated sentences using brat (Stenetorp et al., 2012) and customized the annotation tool to notate for the following cues: negation, positive hedge, and negative hedge. We then instruct annotators to only annotate the cues if they affect the probability of the expression being true or false. We specify that positive hedges are denoted as leaning towards the probability of being true while negative hedges lean towards false. We have these distinctions as we find it is informative to medical professionals when the report denotes if the medical observation is likely, unlikely, or absent for making medical decisions.

Scope and medical foci are also annotated in this corpus. We follow scope annotations by Morante's group (Morante and Daelemans, 2009b) with a few changes. For example, we do not annotate the cues as part of their scopes. As brat allows for annotating relationships, we instruct the annotators to add annotations to designate which scope belongs to which cue. In the case that a cue is in the middle of its scope, separating the scope into two parts, the annotators are instructed to connect the fragments with the brat tool. For medical foci, the medical terms in the scopes are annotated as foci if their probability of being present/absent is af-

[3] Link to algorithm provided after anonymous submission.

fected by the cue. We provide Example A where the positive hedge cue, "suggestive" modifies the scope "of mild encephalopathy" where we mark "encephalopathy" as a medical focus with a higher probability of being true.

<Suggestive> (Positive Hedge) [of mild <encephalopathy> (Hedge Focus)]. (A)

Finally, we must also define what we consider as "medical terms" that can be medical foci. We considered what can affect a diagnosis and may be necessary for doctors to know when diagnosing and treating the patients. Thus, we define "medical terms" as follows:

1. Diagnoses like medical conditions/diseases

2. Signs/Symptoms and causes

3. Procedures/tests and associated observations

4. Medical treatments

Table 1: Preliminary statistics summarizing the results from the first annotator. This provides the percent of sentences containing cues (negation, positive hedge, negative hedge) for the two set of sentences; initial random sample and anchoring sampling.

Sentence Set	Cue Type	% of Sentences
First 1,650 Sentences	Negation	12.4%
	Pos Hedge	4.6%
	Neg Hedge	2.2%
1,500 Anchored Sentences	Negation	43.9%
	Pos Hedge	21.3%
	Neg Hedge	11.1%

Table 2: Cue Frequency Comparison between ClinScope and Bioscope. We only analyze the statistics for the clinical texts in BioScope for this comparison.

	ClinScope	BioScope
Total Sentences	3,150	6,383
# Negation Sentences	27.4%	13.6%
# Negation Cues	1,041	877
# Hedge Sentences	12.5%	13.4%
# Hedge Cues	722	1,189

6 Corpus Analysis

We conducted a preliminary analysis (Table 1) of the annotations from one annotator where we will confirm the findings when the other two annotators have completed their work. Although anchors did not guarantee that all the sentences had a cue, anchor sampling greatly increased the number of sentences with cues in the corpus. We compared the results to BioScope as it is one of the few corpora that provided cue frequency analysis (details in Table 2). Comparing only clinical texts, ClinScope corpus has approximately double the concentration of negation sentences than BioScope with 19% more negation cues in the corpus. Although our corpus has less hedge cues, the percent of sentences is less than 1% difference even though ClinScope is approximately half the total number of sentences as BioScope. Thus, our corpus with less sentences is able to provide more examples of negation cues while maintaining a similar level of hedge cues for use. In addition, anchor sampling led to 26 new negation cues and 32 new positive hedge cues, approximately doubling the number of unique cues for both categories. This also included finding cues that had not been described in previous work (i.e., "off", "c/w" (consistent with)). Finally, anchor sampling also increased the number of examples where cues do not lead to negation/uncertainty, such as cases where the cue is in conditional phrases (i.e., phrases using "if" and "unless"), which had not been discussed in previous works.

7 Conclusion

We provided a new annotated corpus for assertion detection of medical entities with a focus on negation and uncertainty. We employed targeted sampling to increase the concentration of sentences of cues in the corpus and cases where the cues do not lead to to negation or speculation of a medical entity. We had also found and included new cues that have not been discussed in previous works. However, we need more annotated corpora as the current state-of-the-art has room for improvement and more public corpora (especially of different sources for improved diversity) for training and improving algorithms will help. We plan to complete this work through the use of three annotators and calculating the inter-annotator agreement before releasing the dataset on PhysioNet (Goldberger et al., 2000), abiding to the data use agreement.

8 Limitations

We note that since we sampled only MIMIC-III for this corpus, our corpus suffers from not having a variety of reports from different institutions and from ICU patients only. Also, different ICD-9 codes may lead to different kinds of sentences in their reports as we only sampled three codes. We aim to increase the size of this corpora with three inter-annotator agreement using MIMIC-IV, i2b2, and other corpora to improve upon this limitation. We also do not annotate presence of medical entities, which may prove useful for general medical entity detection although there does exist other research that focuses on this realm.

Once completed, our dataset and any future associated work should and will be only provided in PhysioNet (Goldberger et al., 2000) to abide to the data use agreement for using MIMIC-III. PhysioNet grants public but restricted access to the MIMIC-III data to mitigate the risks of using classified patient data regardless if it has been de-identified to protect patient privacy. Users are prompted to complete the CITI Data or Specimens Only Research training and sign the data use agreement, including providing information for intended use. In addition, PhysioNet provides original MIMIC-III dataset de-identified prior to publishing.

Since the brat annotation file also includes texts from MIMIC-III (due to the way the brat tool annotates texts), we therefore ensure we meet data use agreement requirements by ensuring all data files are only provided on this same website. This will also restrict the use of the dataset to aligning with the original access conditions of MIMIC-III. We require that if the algorithm has potential of leaking the information from our annotated dataset (i.e., data leakage from LLMs), it must also be published on PhysioNet.

References

Shashank Agarwal and Hong Yu. 2010. Detecting Hedge Cues and their Scope in Biomedical Literature with Conditional Random Fields. *Journal of biomedical informatics*, 43(6):953–961.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ravindra Arya, Belavendra Antonisamy, and Sushil Kumar. 2012. Sample Size Estimation in Prevalence Studies. *The Indian Journal of Pediatrics*, 79(11):1482–1488.

Eric Boyd. 2023. General availability of Azure OpenAI Service expands access to large, advanced AI models with added enterprise benefits | Azure Blog | Microsoft Azure.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001a. Evaluation of negation phrases in narrative clinical reports. *Proceedings of the AMIA Symposium*, pages 105–109.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001b. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Gunther Eysenbach. 2023. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. *JMIR Medical Education*, 9(1):e46885. Company: JMIR Medical Education Distributor: JMIR Medical Education Institution: JMIR Medical Education Label: JMIR Medical Education Publisher: JMIR Publications Inc., Toronto, Canada.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural Networks For Negation Scope Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Jennifer Frankovich, Christopher A. Longhurst, and Scott M. Sutherland. 2011. Evidence-Based Medicine in the EMR Era. *New England Journal of Medicine*, 365(19):1758–1759. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp1108726.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220.

Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. 2014. Using Anchors to Estimate Clinical State without Labeled Data. *AMIA Annual Symposium Proceedings*, 2014:606–615.

David A. Hanauer, Yang Liu, Qiaozhu Mei, Frank J. Manion, Ulysses J. Balis, and Kai Zheng. 2012. Hedging their Mets: The Use of Uncertainty Terms in Clinical Documents and its Potential Implications when Sharing the Documents with Patients. *AMIA Annual Symposium Proceedings*, 2012:321–330.

Laurence R. Horn. 2001. *A Natural History of Negation*. The David Hume Series. Center for the Study of Language and Information.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. MIMIC-III Clinical Database.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016b. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

George Lakoff. 1973. Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, 2(4):458–508. Publisher: Springer.

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

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do We Still Need Clinical Language Models? In *Proceedings of the Conference on Health, Inference, and Learning*, pages 578–597. PMLR. ISSN: 2640-3498.

Eric Lehman and Alistair Johnson. 2023. Clinical-T5: Large Language Models Built Using MIMIC Clinical Text.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

Sarah BJ Macfarlane. 1997. Conducting a descriptive survey: 2. Choosing a sampling strategy. *Tropical Doctor*, 27(1):14–21.

Ben Medlock and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999, Prague, Czech Republic. Association for Computational Linguistics.

Roser Morante. 2010. Descriptive Analysis of Negation Cues in Biomedical Texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Roser Morante and Walter Daelemans. 2009a. Learning the Scope of Hedge Cues in Biomedical Texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28–36, Boulder, Colorado. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009b. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 21–29, Boulder, Colorado. Association for Computational Linguistics.

Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. 2001. Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *Journal of the American Medical Informatics Association*, 8(6):598–609.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the second international conference on Human Language Technology Research* -, page 82, San Diego, California. Association for Computational Linguistics.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic. Association for Computational Linguistics.

Elena Sergeeva, Henghui Zhu, Peter Prinsen, and Amir Tahmasebi. 2019. Negation Scope Detection in Clinical Notes and Scientific Abstracts: A Feature-enriched LSTM-based Approach. *AMIA Summits on Translational Science Proceedings*, 2019:212–221.

Jugal Shah and Sabah Mohammed. 2020. Clinical Narrative Summarization based on the MIMIC III Dataset. *International Journal of Multimedia and Ubiquitous Engineering*.

Luke T. Slater, William Bradlow, Dino FA. Motti, Robert Hoehndorf, Simon Ball, and Georgios V. Gkoutos. 2021. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. *Computers in Biology and Medicine*, 130:104216.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of* *the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

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

Betty van Aken, Ivana Trajanovska, Amy Siu, Manuel Mayrdorfer, Klemens Budde, and Alexander Loeser. 2021. Assertion Detection in Clinical Notes: Medical Language Models to the Rescue? In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 35–40, Online. Association for Computational Linguistics.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(11):S9.

Ellen M. Voorhees. 2013. The TREC Medical Records Track. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 239–246, Washington DC USA. ACM.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):1–10.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PLoS ONE*, 9(11):e112774.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):1–9.

A GPT-3.5 and Clinical Language Model: Detailed Errors Report

Table A1: Summary of assertion detection performance on an example report by GPT-3.5 a clinical language model (van Aken et al., 2021). The report used was one classified ICD-9 code 5849 (Acute Kidney Failure NOS). We checked 51 medical entities (including repeated entities) where the models were compared for presence, absence, and likelihood detection. We listed the number of missed entities, wrong assertion assignments, and "other" issues (incomplete assertion designations or errors outside of assertion detection).

Model	Missed	Wrong	Other
Clinical	12	4	3
GPT-3	8	9	3

B Sample Size Justification

In MIMIC III, there are over 2 million clinical notes and thus is at least 2 million sentences in size, but the actual number of sentences is unknown. Thus, the sample size is sufficiently large for us to use the following formula for calculating the minimum sample size (Arya et al., 2012):

$$n = \frac{(z^2)P(1-P)}{d^2} \tag{1}$$

The calculation of sample size (n) uses z-score (z), expected prevalence (P), and for allowable error (d). As we do not know the actual distribution for the number of sentences with negation and uncertainty cues (as we do not know if the reported known negation and uncertainty cues are all the cues in existence), we use $P = 0.5$, $d = 0.05$, and $z = 1.96$ used for 95% confidence level as recommended per convention (Macfarlane, 1997). Finite population correction is unnecessary as a sample size of 550 is $\leq 0.0275\%$ of the total MIMIC-III dataset, far less than the 5% minimum requirement. Thus, the result is:

$$n = \frac{(1.96^2)0.5(1-0.5)}{0.05^2} = 384.16 \tag{2}$$