MedNLI Is Not Immune:
Natural Language Inference Artifacts in the Clinical Domain

Christine Herlihy
Department of Computer Science
University of Maryland
College Park, MD
cherlihy@cs.umd.edu

Rachel Rudinger
Department of Computer Science
University of Maryland
College Park, MD
rudinger@umd.edu

Abstract

Crowdworker-constructed natural language inference (NLI) datasets have been found to contain statistical artifacts associated with the annotation process that allow hypothesis-only classifiers to achieve better-than-random performance (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018). We investigate whether MedNLI, a physician-annotated dataset with premises extracted from clinical notes, contains such artifacts (Romanov and Shivade, 2018).

We find that entailed hypotheses contain generic versions of specific concepts in the premise, as well as modifiers related to responsiveness, duration, and probability. Neutral hypotheses feature conditions and behaviors that co-occur with, or cause, the condition(s) in the premise. Contradiction hypotheses feature explicit negation of the premise and implicit negation via assertion of good health. Adversarial filtering demonstrates that performance degrades when evaluated on the *difficult* subset. We provide partition information and recommendations for alternative dataset construction strategies for knowledge-intensive domains.

1 Introduction

In the clinical domain, the ability to conduct natural language inference (NLI) on unstructured, domain-specific texts such as patient notes, pathology reports, and scientific papers, plays a critical role in the development of predictive models and clinical decision support (CDS) systems.

Considerable progress in domain-agnostic NLI has been facilitated by the development of large-scale, crowdworker-constructed datasets, including the Stanford Natural Language Inference corpus (SNLI), and the Multi-Genre Natural Language Inference (MultiNLI) corpus (Bowman et al., 2015; Williams et al., 2017). MedNLI is a similarly-motivated, healthcare-specific dataset created by a small team of physician-annotators in lieu of crowdworkers, due to the extensive domain expertise required (Romanov and Shivade, 2018).

Poliak et al. (2018), Gururangan et al. (2018), Tsuchiya (2018), and McCoy et al. (2019) empirically demonstrate that SNLI and MultiNLI contain lexical and syntactic annotation artifacts that are disproportionately associated with specific classes, allowing a hypothesis-only classifier to significantly outperform a majority-class baseline model. The presence of such artifacts is hypothesized to be partially attributable to the priming effect of the example hypotheses provided to crowdworkers at annotation-time. Romanov and Shivade (2018) note that a hypothesis-only baseline is able to outperform a majority class baseline in MedNLI, but they do not identify specific artifacts.

We confirm the presence of annotation artifacts in MedNLI and proceed to identify their lexical and semantic characteristics. We then conduct adversarial filtering to partition MedNLI into *easy* and *difficult* subsets (Sakaguchi et al., 2020). We find that performance of off-the-shelf `fastText`-based hypothesis-only and hypothesis-plus-premise classifiers is lower on the *difficult* subset than on the *full* and *easy* subsets (Joulin et al., 2016). We provide partition information for downstream use, and conclude by advocating alternative dataset construction strategies for knowledge-intensive domains.[1]

2 The MedNLI Dataset

MedNLI is domain-specific evaluation dataset inspired by general-purpose NLI datasets, including SNLI and MultiNLI (Romanov and Shivade, 2018; Bowman et al., 2015; Williams et al., 2017). Much like its predecessors, MedNLI consists of premise-hypothesis pairs, in which the premises are drawn

---

[1]See https://github.com/crherlihy/clinical_nli_artifacts for code and partition ids.

from the `Past Medical History` sections of a randomly selected subset of de-identified clinical notes contained in MIMIC-III (Johnson et al., 2016; Goldberger et al., 2000 (June 13). MIMIC-III was created from the records of adult and neonatal intensive care unit (ICU) patients. As such, complex and clinically severe cases are disproportionately represented, relative to their frequency of occurrence in the general population.

Physician-annotators were asked to write a *definitely true*, *maybe true*, and *definitely false* set of hypotheses for each premise, corresponding to *entailment*, *neutral* and *contradiction* labels, respectively. The resulting dataset has cardinality: $n_{\text{train}} = 11232$; $n_{\text{dev}} = 1395$; $n_{\text{test}} = 1422$.

3   MedNLI Contains Artifacts

To determine whether MedNLI contains annotation artifacts that may artificially inflate the performance of models trained on this dataset, we train a simple, premise-unaware, `fastText` classifier to predict the label of each premise-hypothesis pair, and compare the performance of this classifier to a majority-class baseline, in which all training examples are mapped to the most commonly occurring class label (Joulin et al., 2016; Poliak et al., 2018; Gururangan et al., 2018). Note that since annotators were asked to create an entailed, contradictory, and neutral hypothesis for each premise, MedNLI is class-balanced. Thus, in this setting, a majority class baseline is equivalent to choosing a label uniformly at random for each training example.

The micro F1-score achieved by the `fastText` classifier significantly exceeds that of the majority class baseline, confirming the findings of Romanov and Shivade (2018), who report a micro-F1 score of 61.9 but do not identify or analyze artifacts:

	dev	test
majority class	33.3	33.3
`fastText`	**64.8**	**62.6**

Table 1: Performance (micro F1-score) of the `fastText` hypothesis-only classifier.

As the confusion matrix for the test set shown in Table 2 indicates, the `fastText` model is most likely to misclassify entailment as neutral, and neutral and contradiction as entailment. Per-class precision and recall on the test set are highest for contradiction (73.2; 72.8) and lowest for entailment (56.7; 53.8).

	entailment	neutral	contradiction
entailment	**255**	151	68
neutral	126	**290**	58
contradiction	69	60	**345**

Table 2: Confusion matrix for `fastText` classifier.

4   Characteristics of Clinical Artifacts

In this section, we conduct class-specific lexical analysis to identify the clinical and domain-agnostic characteristics of annotation artifacts associated with each set of hypotheses in MedNLI.

4.1   Preprocessing

We cast each hypothesis string in the MedNLI training dataset to lowercase. We then use a `scispaCy` model pre-trained on the `en_core_sci_lg` corpus for tokenization and clinical named entity recognition (CNER) (Neumann et al., 2019a). One challenge associated with clinical text, and scientific text more generally, is that semantically meaningful entities often consist of spans rather than single tokens. To mitigate this issue during lexical analysis, we map each multi-token entity to a single-token representation, where sub-tokens are separated by underscores.

4.2   Lexical Artifacts

Following Gururangan et al. (2018), to identify tokens that occur disproportionately in hypotheses associated with a specific class, we compute token-class pointwise mutual information (PMI) with add-50 smoothing applied to raw counts, and a filter to exclude tokens appearing less than five times in the overall training dataset. Table 3 reports the top 15 tokens for each class.

$$\text{PMI(token, class)} = log_2 \frac{p(\text{token, class})}{p(\text{token}, \cdot)p(\cdot, \text{class})}$$

Entailment   Entailment hypotheses are characterized by tokens about: (1) patient status and response to treatment (e.g., *responsive*; *failed*; *longer* as in *no longer intubated*); (2) medications and procedures which are common among ICU patients (e.g., *broad_spectrum*; *antibiotics*; *pressors*; *steroid_medication*; *underwent*; *removal*); (3) generalized versions of specific words in the premise (e.g., *comorbidities*; *multiple_medical_problems*), which Gururangan et al. (2018) also observe in SNLI; and (4) modifiers related to duration, frequency, or probability (e.g., *frequent*, *possible*, *high_risk*).

entailment	%	neutral	%	contradiction	%
just	0.25%	cardiogenic_shock	0.33%	no_history_of_cancer	0.27%
high_risk	0.26%	pelvic_pain	0.30%	no_treatment	0.27%
pressors	0.25%	joint_pain	0.30%	normal_breathing	0.27%
possible	0.26%	brain_injury	0.32%	no_history_of_falls	0.27%
elevated_blood_pressure	0.26%	delerium	0.30%	normal_heart_rhythm	0.28%
responsive	0.25%	intracranial_pressure	0.30%	health	0.26%
comorbidities	0.26%	smoking	0.42%	normal_head_ct	0.26%
spectrum	0.27%	obesity	0.41%	normal_vision	0.26%
steroid_medication	0.25%	tia	0.32%	normal_aortic_valve	0.27%
longer	0.26%	acquired	0.31%	bradycardic	0.26%
history_of_cancer	0.26%	head_injury	0.31%	normal_blood_sugars	0.27%
broad	0.26%	twins	0.30%	normal_creatinine	0.28%
frequent	0.25%	fertility	0.30%	cancer_history	0.26%
failed	0.26%	statin	0.30%	cardiac	0.33%
medical	0.29%	acute_stroke	0.30%	normal_chest	0.28%

Table 3: Top 15 tokens by PMI(token, class); % of *class* training examples that contain the token.

Neutral Neutral hypotheses feature tokens related to: (1) chronic and acute clinical conditions (e.g., *obesity*; *joint_pain*; *brain_injury*); (2) clinically relevant behaviors (e.g., *smoking*; *alcoholic*; *drug_overdose*); and (3) gender and reproductive status (e.g., *fertility*; *pre_menopausal*). Notably, the most discriminative conditions tend to be commonly occurring within the general population and generically stated, rather than rare and specific. This presumably contributes to the relative difficulty that the hypothesis-only `fastText` model has distinguishing between the entailment and neutral classes.

Contradiction Contradiction hypotheses are characterized by tokens that convey normalcy and good health. Lexically, such sentiment manifests as: (1) explicit negation of clinical severity, medical history, or in-patient status (e.g., *denies_pain*; *no_treatment*; *discharged_home*), or (2) affirmation of clinically unremarkable findings (e.g., *normal_heart_rhythm*; *normal_blood_sugars*), which would generally be rare among ICU patients. This suggests a heuristic of inserting negation token(s) to contradict the premise, which Gururangan et al. (2018) also observe in SNLI.

4.3 Syntactic Artifacts

Hypothesis Length In contrast to Gururangan et al. (2018)'s finding that entailed hypotheses in SNLI tend to be shorter while neutral hypotheses tend to be longer, hypothesis sentence length does not appear to play a discriminatory role in MedNLI, regardless of whether we consider merged- or separated-token representations of multi-word entities, as illustrated by Table 4:

	entailment		neutral		contradiction	
	mean	median	mean	median	mean	median
separate	5.6	5.0	5.2	5.0	5.6	5.0
merged	5.3	5.0	4.9	5.0	5.3	5.0

Table 4: Average and median hypothesis length by class and entity representation.

5 Physician-Annotator Heuristics

In this section, we re-introduce premises to our analysis to evaluate a set of hypotheses regarding latent, class-specific annotator heuristics. If annotators *do* employ class-specific heuristics, we should expect the semantic contents, φ, of a given hypothesis, $h \in \mathcal{H}$, to be influenced not only by the semantic contents of its associated premise, $p \in \mathcal{P}$, but also by the target class, $c \in \mathcal{C}$.

To investigate, we identify a set of heuristics parameterized by $\varphi(p)$ and c, and characterized by the presence of a set of heuristic-specific Medical Subject Headings (MeSH) linked entities in the premise and hypothesis of each heuristic-satisfying example. These heuristics are described below; specific MeSH features are detailed in the Appendix.

Hypernym Heuristic This heuristic applies when the premise contains clinical condition(s), medication(s), finding(s), procedure(s) or event(s), the target class is *entailment*, and the generated hypothesis contains term(s) that can be interpreted as super-types for a subset of elements in the premise (e.g., clindamycin <: antibiotic).

Probable Cause Heuristic This heuristic applies when the premise contains clinical condition(s), the target class is *neutral*, and the generated hypothesis provides a plausible, often subjective

or behavioral, causal explanation for the condition, finding, or event described in the premise (e.g., associating altered mental status with drug overdose).

Everything Is Fine Heuristic This heuristic applies when the premise contains condition(s) or finding(s), the target class is *contradiction*, and the generated hypothesis negates the premise or asserts unremarkable finding(s). This can take two forms: repetition of premise content plus negation, or inclusion of modifiers that convey good health.

Analysis We conduct a χ^2 test for each heuristic to determine whether we are able to reject the null hypothesis that pattern-satisfying premise-hypothesis pairs are uniformly distributed over classes.

heuristic	χ^2	p-value	top class
hypernym	59.15	1.4e-13‡	entail. (45.2%)
probable cause	111.05	7.7e-25‡	neutral (57.8%)
everything fine	874.71	1.1e-190‡	contradict. (83.8%)

Table 5: Results of χ^2 test statistic by heuristic, computed using the combined MedNLI dataset (‡ $p < 0.001$, † $p < 0.01$, * $p < 0.5$). Top class presented with % of heuristic-satisfying pairs.

The results support our hypotheses regarding each of the three heuristics. Notably, the percentage of heuristic-satisfying pairs accounted for by the top class is lowest for the HYPERNYM hypothesis, which we attribute to the high degree of semantic overlap between entailed and neutral hypotheses.

6 Adversarial Filtering

To mitigate the effect of clinical annotation artifacts, we employ AFLite, an adversarial filtering algorithm introduced by Sakaguchi et al. (2020) and analyzed by Bras et al. (2020), to create *easy* and *difficult* partitions of MedNLI.

AFLite requires distributed representations of the full dataset as input, and proceeds in an iterative fashion. At each iteration, an ensemble of n linear classifiers are trained and evaluated on different random subsets of the data. A score is then computed for each premise-hypothesis instance, reflecting the number of times the instance is correctly labeled by a classifier, divided by the number of times the instance appears in any classifier's evaluation set. The top-k instances with scores above a threshold, τ, are filtered out and added to the *easy* partition; the remaining instances are retained. This process continues until the size of the filtered subset is $< k$,

or the number of retained instances is $< m$; retained instances constitute the *difficult* partition.

To represent the full dataset, we use fastText_{MIMIC-III} embeddings, which have been pretrained on deidentified patient notes from MIMIC-III (Romanov and Shivade, 2018; Johnson et al., 2016). We represent each example as the average of its component token vectors. We proportionally adjust a subset of the hyperparameters used by Sakaguchi et al. (2020) to account for the fact that MedNLI contains far fewer examples than WINOGRANDE[2]: specifically, we set the training size for each ensemble, m, to 5620, which represents $\approx \frac{2}{5}$ of the MedNLI combined dataset. The remaining hyperparameters are unchanged: the ensemble consists of $n = 64$ logistic regression models, the filtering cutoff, $k = 500$, and the filtering threshold $\tau = 0.75$.

We apply AFLite to two different versions of MedNLI: (1) $\mathcal{X}_{h,m}$: hypothesis-only, multi-token entities merged, and (2) $\mathcal{X}_{ph,m}$: premise and hypothesis concatenated, multi-token entities merged. AFLIte maps each version to an *easy* and *difficult* partition, which can in turn be split into training, dev, and test subsets. We report results for the fastText classifier trained on the original, hypothesis-only (hypothesis + premise) MedNLI training set, and evaluated on the *full*, *easy* and *difficult* dev and test subsets of $\mathcal{X}_{h,m}$ ($\mathcal{X}_{ph,m}$), and observe that performance decreases on the *difficult* partition:

	model	eval dataset	full	easy (Δ)	difficult (Δ)
no premise	majority class	dev	0.33	0.34 (+0.01)	0.35 (+0.02)
no premise	majority class	test	0.33	0.35 (+0.02)	0.37 (+0.04)
no premise	fastText	dev	0.65	0.67 (+0.02)	0.46 (-0.19)
no premise	fastText	test	0.63	0.65 (+0.02)	0.4 (-0.23)
with premise	majority class	dev	0.33	0.45 (+0.12)	0.36 (+0.03)
with premise	majority class	test	0.33	0.48 (+0.15)	0.37 (+0.04)
with premise	fastText	dev	0.53	0.6 (+0.07)	0.43 (-0.1)
with premise	fastText	test	0.51	0.55 (+0.04)	0.4 (-0.11)

Table 6: Performance (micro F1-score) for the majority class baseline and fastText classifiers, with and without premise, by partition (e.g., *full, easy, difficult*).

7 Discussion

7.1 MedNLI is Not Immune from Artifacts

In this paper, we demonstrate that MedNLI suffers from the same challenge associated with annotation artifacts that its domain-agnostic predecessors have

[2]MedNLI's training dataset contains 14049 examples when the training, dev, and test sets are combined, while WINOGRANDE contains 47K after excluding the 6K used for fine-tuning.

encountered: namely, NLI models trained on {Med, S, Multi}NLI can perform well even without access to the training examples' premises, indicating that they often exploit shallow heuristics, with negative implications for out-of-sample generalization.

Interestingly, many of the high-level lexical characteristics identified in MedNLI can be considered domain-specific variants of the more generic, class-specific patterns identified in SNLI. This observation suggests that a set of abstract design patterns for inference example generation exists across domains, and may be reinforced by the prompts provided to annotators. Creative or randomized priming, such as Sakaguchi et al. (2020) 's use of anchor words from WikiHow articles, may help to decrease reliance on such design patterns, but it appears unlikely that they can be systematically sidestepped without introducing new, "corrective" artifacts.

7.2 A Prescription for Dataset Construction

To mitigate the risk of performance overestimation associated with annotation artifacts, Zellers et al. (2019) advocate adversarial dataset construction, such that benchmarks will co-evolve with language models. This may be difficult to scale in knowledge-intensive domains, as expert validation of adversarially generated benchmarks is typically required. Additionally, in high-stakes domains such as medicine, information-rich inferences should be preferred over correct but trivial inferences that time-constrained expert annotators may be rationally incentivized to produce, because entropy-reducing inferences are more useful for downstream tasks.

We advocate the adoption of a mechanism design perspective, so as to develop modified annotation tasks that reduce the cognitive load placed on expert annotators while incentivizing the production of domain-specific NLI datasets with high downstream utility (Ho et al., 2015; Liu and Chen, 2017). An additional option is to narrow the generative scope by defining a set of inferences deemed to be useful for a specific task. Annotators can then map (premise, relation) tuples to relation-satisfying, potentially fuzzy subsets of this pool of useful inferences, or return partial functions when more information is needed.

8   Ethical Considerations

When working with clinical data, two key ethical objectives include: (1) the preservation of pa-

tient privacy, and (2) the development of language and predictive models that benefit patients and providers to the extent possible, without causing undue harm. With respect to the former, MedNLI's premises are sampled from de-identified clinical notes contained in MIMIC-III (Goldberger et al., 2000 (June 13; Johnson et al., 2016), and the hypotheses generated by annotators do not refer to specific patients, providers, or locations by name. MedNLI requires users to complete Health Insurance Portability and Accountability Act (HIPAA) training and sign a data use agreement prior to being granted access, which we have complied with.

Per MedNLI's data use agreement requirements, we do not attempt to identify any patient, provider, or institution mentioned in the de-identified corpus. Additionally, while we provide AFLite *easy* and *difficult* partition information for community use in the form of split-example ids and a checksum, we do not share the premise or hypothesis text associated with any example. Interested readers are encouraged to complete the necessary training and obtain credentials so that they can access the complete dataset (Romanov and Shivade, 2018; Goldberger et al., 2000 (June 13).

With respect to benefiting patients, the discussion of natural language artifacts we have presented is intended to encourage clinical researchers who rely on (or construct) expert-annotated clinical corpora to train domain-specific language models, or consume such models to perform downstream tasks, to be aware of the presence of annotation artifacts, and adjust their assessments of model performance accordingly. It is our hope that these findings can be used to inform error analysis and improve predictive models that inform patient care.

Acknowledgments

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000 (June 13). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 419–429, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yang Liu and Yiling Chen. 2017. Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 63–80, New York, NY, USA. Association for Computing Machinery.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019a. 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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019b. 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.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. *CoRR*, abs/1808.06752.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *CoRR*, abs/1704.05426.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Hypothesis-only Baseline Analysis

To conduct the analysis presented in Section 3, we take the MedNLI training dataset as input, and exclude the premise text for each training example. We cast the text of each training hypothesis to lowercase, but do not perform any additional preprocessing. We use an off-the-shelf `fastText` classifier, with all model hyperparameters set to their default values with the exception of `wordNgrams`, which we set equal to 2 to allow the model to use bigrams in addition to unigrams (Joulin et al., 2016). We evaluate the trained classifier on the hypotheses contained in the MedNLI dev and test datasets, and report results for each split.

A.2 Lexical Artifact Analysis

To perform the analysis presented in Section 4, we cast each hypothesis string in the MedNLI training dataset to lowercase. We then use a `scispaCy` model pre-trained on the `en_core_sci_lg` corpus for tokenization and clinical named entity recognition (CNER) (Neumann et al., 2019a). Next, we merge multi-token entities, using underscores as delimiters—e.g., "brain injury" → "brain_injury".

When computing token-class pointwise mutual information (PMI), we exclude tokens that appear less than five times in the overall training dataset's hypotheses. Then, following Gururangan et al. (2018), who apply add-100 smoothing to raw counts to highlight particularly discriminative token-class co-occurrence patterns, we apply add-50 smoothing to raw counts. Our approach is similarly motivated; our choice of 50 reflects the smaller state space associated with a focus on the clinical domain.

A.3 Semantic Analysis of Heuristics

To perform the statistical analysis presented in Section 5, we take the premise-hypothesis pairs from the MedNLI training, dev, and test splits, and combine them to produce a single corpus. We use a `scispaCy` model pre-trained on the `en_core_sci_lg` corpus for tokenization and entity linking (Neumann et al., 2019b), and link against the Medical Subject Headings (MeSH) knowledge base. We take the top-ranked knowledge base entry for each linked entity. Linking against MeSH provides a unique concept id, canonical name, alias(es), a definition, and one or more MeSH tree numbers for each recovered entity. Tree numbers convey semantic type information by embedding each concept into the broader MeSH hierarchy [3]. We operationalize each of our heuristics with a set of MeSH-informed semantic properties, which are defined as follows:

1. **Hypernym Heuristic:** a premise-hypothesis pair satisfies this heuristic if specific clinical concept(s) appearing in the premise appear in a more general form in the hypothesis. Formally: $\{(p,h)|\varphi(p) \subsetneq \varphi(h)\}$. MeSH tree numbers are organized hierarchically, and increase in length with specificity. Thus, when a premise entity and hypothesis entity are left-aligned, the hypothesis entity is a hypernym for the premise entity if the hypothesis entity is a substring of the premise entity. To provide a concrete example: *diabetes mellitus* is an *endocrine system disease*; the associated MeSH tree numbers are C19.246 and C19, respectively.

2. **Probable Cause Heuristic:** a premise-hypothesis pair satisfies this heuristic if: (1) the premise contains one or more MeSH entities belonging to high-level categories C (diseases), D (chemicals and drugs), E (analytical, diagnostic and therapeutic techniques, and equipment) or F (psychiatry and psychology); and (2) the hypothesis contains one or more MeSH entities that can be interpreted as providing a plausible causal or behavioral explanation for the condition, finding, or event described in the premise (e.g., smoking, substance-related disorders, mental disorders, alcoholism, homelessness, obesity).

3. **Everything Is Fine Heuristic:** a premise-hypothesis pair satisfies this heuristic if the hypothesis contains one or more of the same MeSH entities as the premise (excluding the *patient* entity, which appears in almost all notes) and also contains: (1) a negation word or phrase (e.g., *does not have*, *no finding*, *no*, *denies*); or (2) a word or phrase that affirms the patient's health (e.g., *normal*, *healthy*, *discharged*).

For each heuristic, we subset the complete dataset to find pattern-satisfying premise-heuristic pairs. We use this subset when performing the χ^2 tests.

---

[3] https://meshb.nlm.nih.gov/treeView

A.4 Adversarial Filtering

When implementing `AFLite`, we follow Sak-aguchi et al. (2020). We use a smaller training set size of $m = 5620$, but keep the remaining hyperparameters unchanged, such that the ensemble consists of $n = 64$ logistic regression models, the filtering cutoff, $k = 500$, and the filtering threshold $\tau = 0.75$.