Natural language inference for clinical registry curation

Bethany Percha[1,2]*, Kereeti Pisapati[3,4,5], Cynthia Gao[1], Hank Schmidt[4,5]

* Corresponding author: bethany.percha@mssm.edu

[1] Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

[2] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY

[3] Mount Sinai Innovation Partners, Mount Sinai Health System, New York, NY

[4] Breast Surgical Oncology, Icahn School of Medicine at Mount Sinai, New York, NY

[5] Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

Clinical registries - structured databases of demographic, diagnosis, and treatment information for patients with specific diseases or phenotypes - play vital roles in high-quality retrospective studies, operational planning, and assessment of patient eligibility for research, including clinical trials. However, registries are extremely time and resource intensive to curate. Natural language processing (NLP) can help, but standard NLP methods require specially annotated training sets or the construction of separate models for each of dozens or hundreds of different registry fields, rendering them insufficient for registry curation at scale. Natural language inference (NLI), a specific branch of NLP focused on logical relationships between statements, presents a possible solution, but NLI methods are largely unexplored in the clinical domain outside the realm of conference shared tasks and computer science benchmarks. Here we convert registry curation into an NLI problem, applying five state-of-the-art, pretrained, deep learning based NLI models to clinical, laboratory, and pathology notes to infer information about 43 different breast oncology registry fields. We evaluate the models' inferences against a manually curated, 7439 patient breast oncology research database. The NLI models show considerable variation in performance, both within and across registry fields. One model, ALBERT, outperforms the others (BART, RoBERTa, XLNet, and ELECTRA) on 22 out of 43 fields. A detailed error analysis reveals that incorrect inferences primarily arise through models' misinterpretations of temporality--they interpret historical findings as current and vice versa--as well as confusion based on subtle terminology and abbreviation variants common in clinical notes. However, modern NLI methods show promise for increasing the efficiency of registry curation, even when used "out of the box" with no additional training.

Key words: machine learning, natural language processing, electronic health records, clinical research, natural language inference, text mining, entailment recognition

Word count (main text): 3700 **Word count (abstract)**: 273

Tables: 4 (+3 supplemental) **Figures**: 2

References: 27

Introduction

Academic medical centers, pharmaceutical companies, and other healthcare organizations routinely develop clinical registries: structured databases of demographic, treatment, and disease-related information for patients with specific diseases or phenotypes. The creation and maintenance of these registries has traditionally relied upon human curators, who sift through thousands of clinical documents, manually abstracting relevant information into database software **[1]**. The structured data they create form the foundation of high-quality retrospective analyses and ongoing assessment of patient eligibility for research studies, including clinical trials **[2]**. Registries also play an important role in generating "real world evidence" for the pharmaceutical regulatory approval process **[3]**. Manual registry curation, however, is expensive, time-consuming, and prone to human error. As registries grow, the time and effort required to enter new patient information compete with a constant need to update the records of existing patients, causing curators to become bottlenecks and delaying registries relative to the true state of patient care. These delays, in turn, create larger delays in upstream registries that draw data from multiple institutions **[4, 5]**.

Given these challenges, the development of natural language processing (NLP) methods that automatically extract and structure registry information from text would have profound implications, both for clinical research and patient care **[6]**. However, the task is extremely challenging: information pertaining to dozens of different registry fields must be extracted simultaneously, and patients' conditions and treatment status both change over time. The most obvious NLP approach--building a separate text classification model for each registry field **[20]**--would necessitate the development of dozens or even hundreds of different models, each requiring separate annotated data for training. Indeed, any approach requiring annotated training data is likely to fail for one simple, practical reason: instead of hiring annotators, an institution could just as easily hire curators to build the registry itself.

The past few years have seen a reinvention of NLP methods around a class of large language models based on a novel neural network architecture called the transformer **[7]**. These models typically work in two stages. They are first "pretrained" using simple tasks like masked-word prediction, learning to model the regularities of language by

repeating this task millions of times on massive, unlabeled corpora. The pretrained models are then "fine-tuned" to perform a variety of different NLP tasks. Most recently, these models have been applied to the famously difficult task of natural language inference (NLI), the task of deciding whether a given statement (the "hypothesis") is implied to be true or false by another (the "premise") **(Figure 1) [8]**. Also called "entailment recognition", NLI closely resembles the task human curators perform in registry curation: reading the thousands of sentences that comprise a patient's medical record, then deciding whether a series of other statements--corresponding to the structured data fields of the registry--are true. The most important feature of these models is their breadth: the same model can be applied to a potentially infinite number of different statement pairs. One well-performing NLI model could, therefore, be used to curate information for dozens or even hundreds of different registry fields.

In 2020, a team at Facebook released five pretrained NLI models with different transformer-based architectures **[9]**. Here we apply these models to curate a clinical registry for breast oncology, presenting a breakdown of model performance by registry field as well as a global comparison of models. Were this approach successful, it would allow us to build registries automatically from patient medical records, greatly increasing the speed and accuracy of curation and minimizing the effort required of human curators. It would also enable the rapid curation of registries for complex, rare, and underfunded diseases.

Results

Characteristics of patient population and documentation

A graphical summary of the study is shown in **Figure 2**. The study population consisted of 7439 patients who underwent surgery in the Mount Sinai Health System (New York, NY, USA) between January 1, 2010 and December 10, 2020 and had information recorded in an internal clinical research database (henceforth abbreviated as "BSD" for "Breast Surgery Database"), as well as one or more clinical, laboratory, or pathology notes recorded in the Mount Sinai Epic electronic health record (EHR) system within one year of surgery.

Of the 7439 patients, 7386 (99.3%) were female and 53 (0.7%) were male. The median age at the time of surgery was 53.2 years (range: 13.0-97.5; IQR: 44.4-64.0). The patients' self-reported races, in order of frequency, were White/Caucasian (4254; 57.2%), Other (1324; 17.8%), Black or African American (935; 12.6%), Asian (405; 5.4%), Pacific Islander (49; 0.7%), and American Indian or Alaska Native (4; <0.1%); for 468 patients (6.3%) the race was listed as Unknown or Not Reported. Ethnicity information was also self-reported; 4430 patients (59.5%) listed their ethnicity as Not Hispanic or Latino, 897 (12.1%) as Hispanic or Latino, and 2112 (28.4%) did not provide ethnicity information.

The number of unique sentences in documents recorded in clinical, pathology, and laboratory notes in Epic within one year of a patient's surgery date ranged from 2 to 19835. The median number of unique sentences was 374 (IQR: 175-778). Thirty-two patients (0.4%) had more than 5000 unique sentences.

NLI model accuracy and relative performance

True and false hypothesis statements were generated programmatically from structured information in the BSD using the strategies outlined in **Supplementary Tables 1 and 2**. The total number of premise-hypothesis comparisons made by the models was 521,449,945 (approximately 104 million per model x 5 models). Models were evaluated by their ability to score true inferences above false inferences across 43 different registry fields. A summary of the models' performance is shown in **Table 1** and a breakdown by registry field is in **Table 2**.

Five NLI models were evaluated: ALBERT **[10]**, BART **[11]**, RoBERTa **[13]**, XLNet **[14]**, and ELECTRA **[12]**. ALBERT outperformed the other models (scoring true inferences higher than false inferences for a greater fraction of patients) across 22 out of 43 (51.2%) registry fields. RoBERTa ranked second, coming in first in 9/43 fields, followed by ELECTRA (5/43), BART (4/43) and XLNet (3/43). Because the laterality (left or right breast) of certain categories was a potentially important confounder, we also considered model performance among only the 17 fields with no laterality. The ranking differed slightly: ALBERT again came in first, winning 8/17 fields, followed by XLNet and ELECTRA (tied for second with 3/17 fields each), RoBERTa (2/17), and BART (1/17).

There was considerable variation in model performance across registry fields. The registry field curated most easily by NLI models was patient age; RoBERTa, the best-performing model for that field, assigned a higher maximum entailment score (MES; see **Figure 1**) to true statements of age vs. false statements for 94.4% of patients. This was followed by site of surgery (RoBERTa; 93.8% correct), whether a breast reconstruction had been performed (ALBERT; 89.5% correct), and whether the patient had received neo-adjuvant therapy (ALBERT; 92.9% correct). The most difficult-to-curate fields were related to the estrogen receptor (ER), progesterone receptor (PR), and HER2 status of the breast masses. ALBERT outperformed the other models in all these cases. For ER status, ALBERT assigned a higher MES to true statements for 26.9% (right side) and 29.4% (left side) of patients; for PR, it was correct in 11.9% (right side) and 13.5% (left side) of patients; for HER-2, it was correct in 26.3% (right side) and 30.7% (left side) of patients.

There was also considerable variation in model performance across the five models within individual registry fields. For example, ELECTRA completely failed to identify true statements about whether sentinel node biopsies and axillary dissections were performed (**Table 2**, rows 5, 7, 27, and 33), even though the other four models performed well in these categories. ALBERT consistently outperformed the other models among the most difficult-to-infer fields (**Table 2**, rows 26 and below); for example, ALBERT was able to identify the correct number of sentinel lymph nodes biopsied in approximately 50% of cases, while performance among the other four models ranged from 22-40% (**Table 2**, rows 29 and 32). Among the "sided" fields, model rankings tended to be the same for both left and right breasts; for example, XLNet best identified whether a sentinel lymph node biopsy had been performed on the left side (82.9%; **Table 2**, row 5) and it was also the highest-ranking model for the same field on the right side (81.7%; **Table 2**, row 7).

Error analysis and characteristic patterns

Examples of true and false hypothesis statements with high entailment scores, as well as the premise statements from patient EHRs that generated those scores, are shown in **Table 3**. A detailed error analysis by registry field is shown in **Table 4**.

The most common errors arose from models' misinterpretations of temporality; they interpreted prior or planned surgeries as current and vice versa. In part, this was a product of our experimental design; since notes were extracted from within a year of surgery, descriptions of the index surgery from before or after it occurred were almost always included. For example, a patient may have had no prior surgery on the index date (and hence be listed in the BSD as having had no prior surgery) but the surgery performed on the index date could be referred to as a "prior surgery" in a note written shortly thereafter. Manual review (**Table 4**) revealed that approximately a quarter of all "errors"--statements with high entailment scores that did not match the BSD--were indeed correct inferences based on what was stated explicitly in the text.

A second set of errors arose from models' tendency to fixate on certain clinical terms. For example, the models often failed to distinguish between references to "palpable" breast masses and "palpable" lymph nodes (**Table 4**, rows 4-5); they also tended to interpret any reference to "biopsy" as an indication that a patient had undergone a sentinel lymph node biopsy (**Table 4**, rows 12-13). Lymphovascular invasion was often inferred incorrectly because notes contained references to other words beginning with "lymph", such as lymphedema, lymphadenopathy, and lymphoma (**Table 4**, rows 17-18). Clinical terms like "papilloma" or "fibroadenoma", which share a suffix with "carcinoma" (cancer), were also sometimes misinterpreted as indicating a diagnosis of breast cancer (**Table 4**, row 6).

Finally, there were a few cases in which what was written in the note directly contradicted the database, most likely because notes contained text that was copy/pasted over from earlier notes. One of these was age: patients' age as written in notes sometimes did not match what one would obtain by subtracting the patient's birth date from the date the note was written (**Table 4**, row 2). The patient might be described as a "30-year-old female" but be 32 years of age, for instance. A second case related to smoking status. Over half the patients in the BSD were listed as nonsmokers at the time of surgery (or had no smoking status listed) but the NLI models sometimes picked up instances where a supposedly nonsmoking patient was described as a former or occasional smoker in notes (**Table 4**, row 9).

Discussion

Our goals in this project were (1) to establish a baseline for NLI performance in registry curation and (2) to uncover characteristic error patterns that could help guide the development of future NLI models. Hundreds of papers have applied NLP to clinical text **[6]**, but most of these approaches are inappropriate for the registry curation problem, either because they rely on the existence of specially annotated training data or because they would necessitate building a separate model for each registry field. NLI presents a possible path forward.

Factors affecting NLI model performance

Although the five NLI models shared many architectural similarities, the ALBERT-based model outperformed the others on a majority (22 out of 43) registry fields. The details of the ALBERT model offer clues to its enhanced performance: it incorporates factorized embeddings and cross-layer parameter sharing, which reduce the total number of parameters learned by the model **[10]**. These choices were designed to permit pretraining on much larger datasets: in other words, ALBERT sees much more data during pretraining for a similar amount of computational cost. This appears to provide the model with a significant performance lift on the registry curation task and suggests that the size of the pretraining corpus is a significant factor in clinical NLI.

A second important consideration is domain specificity. We were initially concerned that the high frequency of specialized terms and abbreviations in clinical text would cause the pretrained NLI models to fail e.g., select random and irrelevant premise statements and essentially act as random number generators. However, this turned out not to be the case. The models, which were trained using general-domain corpora (text from news articles, blogs, books, Wikipedia, etc.), made correct inferences for at least 50% of patients on over half of the registry fields tested (**Table 2**). This observation has important implications for the focus of future clinical NLI model development. For example, to date there have been a few attempts to pretrain BERT on clinical text **[24, 25]** and create NLI training sets specific to clinical text **[23]**. However, the corpora involved in these studies are orders of magnitude smaller than the general-domain corpora used to pretrain and fine-tune the models used in our experiments **[9]**. If corpus

size is the determining factor in the quality of NLI models for registry curation, it may make sense to combine corpora or NLI training sets from the clinical and general domains, rather than train clinically specific models.

Finally, we observed that even when the NLI models made incorrect inferences, they almost always chose premise sentences from the patient's record that contained the information needed to determine the truth or falsehood of a given hypothesis (**Table 3**). This suggests that even if NLI models are not sufficient for end-to-end registry curation, they could be effective in software that filters patient records to identify relevant sentences and paragraphs. Whether this could also be accomplished through simpler methods like term matching remains to be seen.

Computational considerations and model deployment

Modern NLP methods based on pretrained transformer models are extremely computationally expensive. The pretraining process for BERT **[21]** and related architectures is so resource-intensive that most active development is happening at large technology companies, such as Facebook and Google, and well-funded research labs at institutions like OpenAI and Stanford. Other groups wishing to use these models typically apply a transfer learning approach **[22]**, in which pretrained models are downloaded and then fine-tuned to perform downstream tasks, such as NLI. Even the fine-tuning step is likely to be prohibitive for most healthcare institutions, however; most will lack access to GPUs and the expertise needed to train and deploy these models. For this reason, we opted to deploy the five pretrained NLI models "out of the box", even though they had been fine-tuned on general-purpose NLI corpora (SNLI, MNLI, FEVER, and ANLI) and might have benefited from further training on a clinical NLI corpus like MedNLI **[23]**. Although we performed no additional training, even the inference process was expensive on our 7439-patient dataset: it took approximately one month to apply the five pretrained models to all ~521 million premise-hypothesis pairs in our experiments, running between 10 and 30 jobs in parallel at any given time. Of course, most registries are built over a period of years, and there would rarely be a need to batch process an entire registry at once as we did.

Deploying an NLI model for registry curation requires the thoughtful design of hypothesis statements for all of the different registry fields, as we have outlined in **Supplemental Tables 1 and 2.** This process takes time and should be checked by a domain expert. Ideally, multiple forms of each hypothesis would be included to reduce the impact of noise and/or models' fixation on particular terms.

Study limitations

Our study faced several data extraction and preprocessing challenges, our responses to which may have affected model performance and/or limited the applicability of these methods. First, the NLI models we used can only handle single-sentence inference, meaning that both premise and hypothesis can only be single sentences. More complex patterns of reasoning, e.g., by combining information from multiple notes in a patient's record, would require different methods. Second, because the patients in the BSD had surgery at different times over a span of a decade, they had different amounts of text available in Epic. Through our experimental design, we were effectively assuming that the information needed to make inferences about all our BSD-derived hypothesis statements was present in the text extracts we obtained from Epic; in reality, this was not guaranteed to be the case. Because our results somewhat conflate data availability and model performance, we present a detailed breakdown of model performance by patient data availability in **Supplementary Table 3**. Finally, the models' struggles with temporal reasoning might have been alleviated by applying heuristics or treating the different registry fields differently. Information about different registry fields can be found in multiple locations within the patient's EHR, and some registry fields are more likely to be found in specific places - pathology reports, post-operative reports, etc. We could have applied restrictions to the models to enable them to access text from only these sources. We could likewise have applied date restrictions, allowing certain fields access to a narrower set of notes than others to avoid problems stemming from confusion of past, present, and planned surgeries. Since this study was designed to establish a baseline for NLI performance on this task, however, we chose to employ the simplest approach possible (allowing all fields access to all notes) even if it hurt model performance.

Conclusions and future work

As the volume of EHR data in the United States and around the world grows **[26]**, we are in increasingly greater need of systems that can structure and curate this information automatically. Clinical registries are a fundamental component of high-quality clinical research and operational planning, yet the quality and timeliness of registry data still hinge on human curators' ability to interpret vast amounts of clinical text. This is exactly the sort of detailed, repetitive task that most benefits from computational approaches. However, the structure of the task, particularly the need to curate dozens of different registry fields simultaneously, puts it beyond the reach of today's most popular clinical NLP systems, most of which are limited to simpler tasks like named entity recognition and text classification **[6, 19, 27]**.

This study is, to our knowledge, the first to deploy modern NLI methods on a practical, clinical task that is not part of a computer science shared task or benchmark. We hope it provides encouragement to the vibrant and growing NLI community to continue to develop NLI technology in a way that will permit healthcare institutions--even those without computer science expertise or sophisticated computational resources--to use NLI for registry curation and related problems that directly impact patient lives.

Methods

Data access, storage, and processing

Raw clinical, laboratory, and pathology notes were obtained by querying the Mount Sinai Epic Caboodle database (Epic Systems Corp., Verona, WI, USA). Notes were included if they were written within one year of a patient's surgery date, as recorded in the BSD. Notes were broken into sentences using the spaCy **[19]** default NLP pipeline. The BSD itself is stored in a RedCap database **[1]** and was obtained via direct export. If a patient had more than one surgery, we restricted our study to information from the first surgery date. This was there is no reason why the methods described here could not be applied to all surgeries, but we did not want patients with multiple surgeries to dominate the evaluation.

This study was approved by the Mount Sinai Institutional Review Board. Hypothesis statement generation and interpretation of results were overseen by a breast surgery attending (HS). All data storage and analysis were performed on Minerva, the Icahn School of Medicine's HIPAA-compliant supercomputing cluster. This project benefited from our use of Minerva's NVIDIA V100 GPU compute nodes, which greatly sped up the computations required to apply the pretrained NLI models to hundreds of millions of sentence pairs.

Translation of structured database fields into hypothesis statements

A full description of how hypothesis statements were generated and labeled for each patient based on BSD information can be found in **Supplementary Tables 1 and 2**. For numeric fields, like age, multiple statements were generated, only one of which contained the correct age; typically, 4-5 false statements were generated for each true statement. For categorical fields with few levels, like smoking status, one statement was generated for each level (e.g., "current", "former", "never", "unknown"), only one of which contained the correct value. For string fields, such as principal finding, which could have hundreds of levels, 4-5 strings randomly sampled from the rest of the database were used to create false statements, while the patient's true string was used to generate a true statement.

Application of pre-trained NLI models

Five pre-trained NLI models were used. We denote each by the name of its pretraining approach: ALBERT **[10]**, BART **[11]**, ELECTRA **[12]**, RoBERTa **[13]**, and XLNet **[14]**. These models were released by a team at Facebook in 2020 **[9]** and stored on the HuggingFace **[15]** model repository. All had been fine-tuned for NLI using a combination of NLI datasets, including SNLI **[16]**, MNLI **[17]**, FEVER-NLI **[18]**, and ANLI (R1, R2, R3) **[9]**.

After converting the structured information from the BSD into hypothesis statements for each patient, we applied these pretrained NLI models to the Cartesian product of hypothesis statements and all "premise" sentences from the patient's clinical, laboratory, and pathology notes. Each of these comparisons produced an entailment score: the

model's level of certainty that the premise implied the hypothesis. The maximum entailment score (MES) was, for each hypothesis statement, its highest entailment score across all possible premise statements for that patient. This approach also identified the specific premise statement that produced the MES.

Our measure of performance for the NLI models was the fraction of patients for whom the MES for true statements (MES-T) was higher than the MES for false statements (MES-F) for statements pertaining to a particular registry field. We summarized this measure for each model-field combination (**Table 4**) and created a global summary of each model's performance across fields (**Table 3**).

Acknowledgments

This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. We would particularly like to thank Lili Gai, who provided extensive assistance with the Minerva GPU nodes and helped troubleshoot several failed job submissions.

Author Contributions

KP and HS created the initial problem specification for the project. BP chose NLI as a potential solution, designed and built the software for registry and note preprocessing and NLI model application, ran the NLI jobs on Minerva, and drafted the paper. CG extracted the raw clinical, laboratory, and pathology notes from Epic. CG and KP manually evaluated 860 sentence pairs each for the error analysis in Table 4. HS oversaw the project, provided technical advice and feedback, quality checked the auto-generated hypothesis statements, and served as our source of access to the BSD. All of the authors reviewed, edited, and proofread the final manuscript.

Funding

Funding for this project was provided by the Icahn School of Medicine at Mount Sinai.

Competing Interests

The authors declare no competing interests.

References

[1] Harris, P.A. et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*. **42(2),** 377-81 (2009).

[2] Hickey, G.L. et al. Clinical registries: governance, management, analysis and applications. *European Journal of Cardio-Thoracic Surgery*. **44(4),** 605-14 (2013).

[3] Beaulieu-Jones, B.K. Examining the Use of Real-World Evidence in the Regulatory Process. *Clinical Pharmacology & Therapeutics*. **107(4),** 843-52 (2020).

[4] Midthune, D.N. Fay, M.P. Clegg, L.X. Feuer, E.J. Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data. *Journal of the American Statistical Association*. **100(469)**, 61-70 (2005).

[5] Bray, F. & Parkin, D.M. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. *European Journal of Cancer.* **45(5),** 747-55 (2009).

[6] Percha, B. Modern Clinical Text Mining: A Guide and Review. *Annual Review of Biomedical Data Science.* **Vol. 4.** Preprint at https://doi.org/10.1146/annurev-biodatasci-030421-030931 (2021).

[7] Vaswani, A. et al. Attention Is All You Need. Preprint at https://arxiv.org/abs/1706.03762 (2017).

[8] Dagan, I. Roth, D. Sammons, M. Zanzotto, F.M. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies.* **6(4),** 1-220 (2013).

[9] Nie, Y. et al. Adversarial NLI: A New Benchmark for Natural Language Understanding. Preprint at https://arxiv.org/abs/1910.14599 (2019).

[10] Lan, Z. et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. Preprint at https://arxiv.org/abs/1909.11942 (2019).

[11] Lewis, M. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Preprint at https://arxiv.org/abs/1910.13461 (2019).

[12] Clark, K. Luong, M.T. Le, Q.V. Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. Preprint at https://arxiv.org/abs/2003.10555 (2020).

[13] Liu,Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).

[14] Yang, Z. et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Preprint at https://arxiv.org/abs/1906.08237 (2019).

[15] Wolf, T. et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. Preprint at arXiv:1910.03771 (2019).

[16] Bowman, S.R. Angeli, G. Potts, C. Manning, C.D. A large, annotated corpus for learning natural language inference. Preprint at https://arxiv.org/abs/1508.05326 (2015).

[17] Williams, A. Nangia, N. Bowman, S.R. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Preprint at https://arxiv.org/abs/1704.05426 (2017).

[18] Thorne, J. Vlachos, A. Christodoulopoulos, C. Mittal, A. FEVER: a large-scale dataset for Fact Extraction and VERification. Preprint at https://arxiv.org/abs/1803.05355 (2018).

[19] Honnibal, M. Montani, I. Van Landeghem, S. Boyd, A. spaCy: Industrial-Strength Natural Language Processing in Python. https://github.com/explosion/spaCy

[20] Kowsari, K. et al. Text Classification Algorithms: A Survey. *Information.* **10(4),** 150 (2019).

[21] Devlin, J. Chang, M.W. Lee, K. Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://arxiv.org/abs/1810.04805 (2018).

[22] Pan, S.J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering.* **22(10),** 1345-59 (2009).

[23] Romanov, A. & Shivade, C. Lessons from Natural Language Inference in the Clinical Domain. Preprint at https://arxiv.org/abs/1808.06752 (2018).

[24] Alsentzer, E. et al. Publicly Available Clinical BERT Embeddings. Preprint at https://arxiv.org/abs/1904.03323 (2019).

[25] Huang, K. Altosaar, J. Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Preprint at https://arxiv.org/abs/1904.05342 (2019).

[26] Office of the National Coordinator for Health Information Technology. "Office-based Physician Electronic Health Record Adoption," Health IT Quick-Stat #50. dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php. January 2019.

[27] Reátegui, R. & Ratté, S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*. **18(3),** 13-9 (2018).

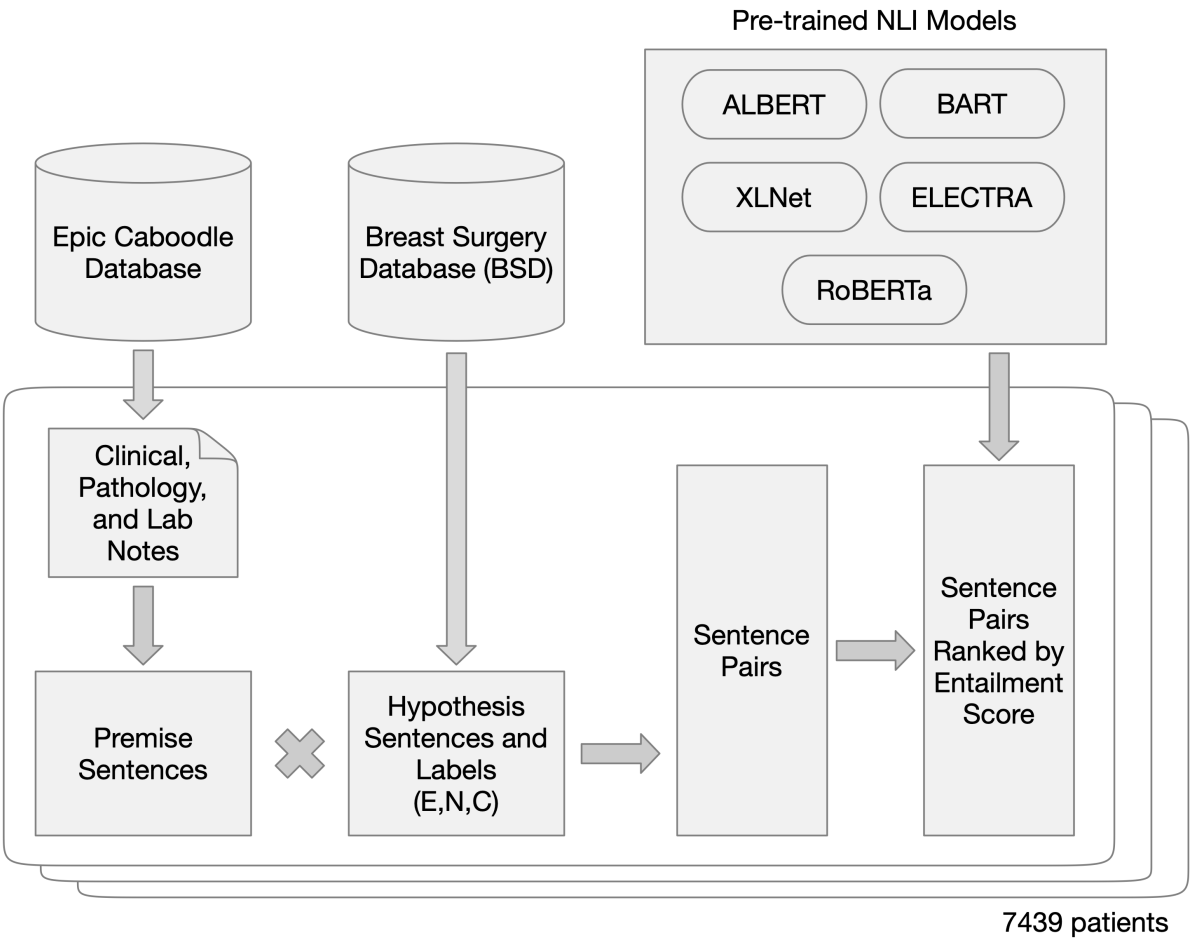
Figures

Figure 1: Natural language inference (NLI) model evaluation. An NLI model considers each premise sentence from a patient's record (shaded blue) and compares it to a series of hypothesis statements, producing an entailment score ("Score") for each comparison. Some of the hypothesis statements are true based on information from the BSD (labeled "E" for "entailment") and some are false (labeled "C" for "contradiction"). The maximum entailment score (MES) for true statements is compared to the MES for false statements to assess the quality of model inferences. Shown here is the process for a single registry field: site of surgery. The real experiments usually involved hundreds of premise sentences and dozens of hypothesis sentences based on 43 different registry fields. Note: The sentences shown here were invented for the purposes of this example and do not refer to any real patient.

(1) The patient is a 44 y.o female G3P2 with right invasive ductal cancer (ER+, PR+, HER2-). (2) Planned procedure is a right total mastectomy with reconstruction.

Premise	Hypothesis	Label	Score	
(1) The patient is a 44 y.o female G3P2 with right invasive ductal cancer (ER+, PR+, HER2-).	The patient is undergoing surgery on the right breast.	E	0.402	
(1) The patient is a 44 y.o female G3P2 with right invasive ductal cancer (ER+, PR+, HER2-).	The patient is undergoing surgery on the left breast.	C	0.030	
(2) Planned procedure is a right total mastectomy with reconstruction.	The patient is undergoing surgery on the right breast.	E	0.957	**MES-T**
(2) Planned procedure is a right total mastectomy with reconstruction.	The patient is undergoing surgery on the left breast.	C	0.318	**MES-F**

Figure 2: Graphical overview of the study. Premise sentences were extracted from clinical, pathology, and laboratory notes. Hypothesis sentences were generated from structured data in the BSD following the protocols in Tables 1 and 2. All possible combinations of premise and hypothesis were evaluated by one of five pre-trained NLI models. The top-ranked sentence for each hypothesis was extracted along with its entailment score. The process was repeated for all 7439 patients and all 5 NLI models.



Tables

Table 1: Summary of model performance. For each of the 43 possible hypothesis statement categories, the models were ordered according to the percent of patients for whom the maximum entailment score for true statements, MES-T, was greater than the MES for false statements, MES-F.

Model	Overall (n=43)			No Laterality (n=17)			Laterality (n=26)		
	Rank Sum	Times Ranked 1st	Times Ranked 1st or 2nd	Rank Sum	Times Ranked 1st	Times Ranked 1st or 2nd	Rank Sum	Times Ranked 1st	Times Ranked 1st or 2nd
ALBERT	**90**	**22**	**28**	**42**	**8**	**10**	**48**	**14**	**18**
BART	121	4	18	49	1	8	70	6	12
RoBERTa	126	9	17	50	2	5	71	2	13
XLNet	142	3	15	56	3	5	93	2	7
ELECTRA	166	5	8	58	3	6	108	2	2

Table 2: Model performance by hypothesis category. Hypothesis categories are ordered according to the median fraction of patients for whom the maximum entailment score for true statements, MES-T, was greater than the MES for false statements, MES-F.

				Fraction with MES-T > MES-F				
	Hypothesis Category	Laterality	N	ALBERT	BART	ELECTRA	RoBERTa	XLNet
1	age		7439	0.895	0.889	0.942	**0.944**	0.935
2	site		7439	0.933	0.922	0.912	**0.938**	0.934
3	reconst		7294	**0.895**	0.792	0.629	0.828	0.875
4	neo_adj		5185	**0.929**	0.735	0.370	0.643	0.831
5	sln_bx_left	X	2646	0.732	0.669	0.000	0.731	**0.829**
6	palpable_right	X	3955	0.670	0.700	0.695	**0.753**	0.562
7	sln_bx_right	X	2576	0.691	0.683	0.000	0.747	**0.817**
8	path_dx		7439	0.639	0.686	**0.775**	0.672	0.691
9	palpable_left	X	4096	0.656	0.711	0.663	**0.735**	0.605
10	genetics		7439	0.640	0.648	**0.704**	0.691	0.598
11	fhx_breast		7439	0.614	**0.681**	0.597	0.641	0.664
12	gravida		7439	0.638	0.599	0.561	0.446	**0.671**
13	hormone_use		7439	**0.634**	0.596	0.598	0.449	0.549
14	smoking		7439	**0.799**	0.771	0.569	0.498	0.587
15	margins_right	X	1786	0.516	0.516	**0.536**	0.410	0.209
16	margins_left	X	1893	0.499	0.549	**0.556**	0.443	0.225
17	grade_left	X	2123	0.496	0.559	0.426	**0.563**	0.432
18	grade_right	X	2024	0.495	0.506	0.428	**0.533**	0.344
19	image_biop		7439	**0.690**	0.467	0.675	0.169	0.116
20	para		7439	0.218	**0.593**	0.426	0.464	0.538
21	lvi_right	X	1627	**0.536**	0.307	0.451	0.458	0.487
22	lvi_left	X	1706	**0.519**	0.338	0.457	0.447	0.514
23	status		7439	0.423	0.316	0.346	**0.518**	0.378
24	prin_finding_left	X	4438	0.367	0.532	0.074	**0.547**	0.340
25	prin_finding_right	X	4281	0.364	0.517	0.075	**0.555**	0.349
26	surgery_type_left	X	7439	**0.374**	0.361	0.343	0.357	0.249
27	ax_dissect_right	X	2529	**0.634**	0.208	0.000	0.540	0.348
28	fertility_rx		7439	**0.542**	0.537	0.347	0.233	0.256
29	sln_bx_left_n	X	1769	**0.499**	0.383	0.291	0.343	0.239
30	fhx_ovar		7439	**0.429**	0.346	0.237	0.261	0.338
31	surgery_type_right	X	7439	**0.350**	0.336	0.327	0.346	0.227
32	sln_bx_right_n	X	1741	**0.504**	0.402	0.286	0.318	0.221
33	ax_dissect_left	X	2615	**0.630**	0.203	0.000	0.514	0.315
34	prior_surg		7439	0.209	0.226	**0.266**	0.212	0.228
35	ax_dissect_right_n	X	296	0.274	**0.351**	0.223	0.166	0.169
36	personal_hx		7439	**0.208**	0.205	0.131	0.207	0.192
37	ax_dissect_left_n	X	361	0.285	**0.335**	0.180	0.188	0.199

38	er_right	X	2211	**0.269**	0.191	0.029	0.114	0.200
39	er_left	X	2280	**0.294**	0.242	0.028	0.117	0.186
40	her2_left	X	1708	**0.307**	0.166	0.042	0.248	0.123
41	her2_right	X	1678	**0.263**	0.141	0.043	0.243	0.121
42	pr_left	X	2277	**0.135**	0.080	0.013	0.024	0.093
43	pr_right	X	2207	**0.119**	0.063	0.011	0.026	0.102

Table 3: Top ten sentences with maximum entailment scores for the ALBERT model for hypothesis statements that were (a) true and (b) false based on information from the BSD. All entailment scores for the sentences shown here were greater than 0.99. To be included in the table, a sentence needed to be less than 300 characters in length. Only the top statement from each hypothesis category was included. All PHI elements and other potentially identifying information, such as physician names, were replaced by bracketed generic terms.

(a)

	Hypothesis Category	Hypothesis Statement	Top Candidate Premise
1	age	"The patient is <AGE IN YEARS> years old."	"All Imaging Tests Reviewed: No Assessment and Plan: <AGE IN YEARS>-year old with locally advanced breast cancer, ER -, PR +, HER 2 -; getting neoadjuvant dd AC-->T. Breast cancer - completed AC 4/4--> T4/4 - BRCA 1/2 without mutation - s/p left lumpectomy with Dr. <SURNAME> with CPR." *Note: The age listed in the hypothesis statement matched the age that was in the note.*
2	sln_bx_right	"A sentinel lymph node (SLN) biopsy was performed on the right side."	"IDC- ER+, PR-, Her2-neg Procedure and Date: On <DATE> she had bilateral total mastectomies nipple sparing, right sentinel lymph node biopsy, right axillary lymph node dissection, and prepectoral tissue expander placement."
3	neo_adj	"The patient has not received neo-adjuvant chemotherapy."	"========== DIAGNOSIS: L breast cancer PROCEDURE AND DATE: L breast lumpectomy and SLNBx <DATE> PATHOLOGY: L breast IDC, + / + / - ; 0/3 LN's postiive for carcinoma STAGE: pT1b pN POST-OPERATIVE TREATMENT: Chemotherapy: None."
4	palpable_right	"The right-side mass was palpable."	"Cancer history: On <DATE>, Ms. <SURNAME> underwent bilateral ultrasound-guided core biopsies of a right palpable mass and a left-sided mammographically detected lesion, which revealed right invasive ductal carcinoma (ER+/PR+, HER2 pending) and left ductal carcinoma in situ (ER+/PR+, HER2 pending)."
5	genetics	"The patient has a genetic susceptibility to breast cancer."	"IMPLANTS: Sientra Dermaspan smooth surfaced tissue expander, full height, reference #<REFNUM>, serial #<SERIALNUM>, initial fill 400mL. INDICATIONS: This is a <AGE IN YEARS>-year-old

			female with a history of the CHEK2 gene mutation with genetic predisposition for development of breast cancer."
6	prin_finding_left	"Principal finding, left side: ILC."	"The Mount Sinai Hospital New York, NY 10029-6574 PATIENT NAME: <SURNAME>, <FIRSTNAME> MRN: <MRN> ACCOUNT NUMBER: <ACCNUM> ADMIT DATE: <DATE> DATE OF PROCEDURE: <DATE> SURGEON: Dr. <FIRSTNAME> <SURNAME>, MD PREOPERATIVE DIAGNOSIS: Invasive lobular cancer of the left breast."
7	site	"The patient is undergoing surgery on the right breast."	"Received a phone call from <SURNAME>, <FIRSTNAME> MRN <MRN> who has a hx a <SPAN lang=EN>T2N2M0 ER+PR+HER2- breast cancer s/p right mastectomy <DATE> with 4/30 LN+, adjuvant AC x 4 and weekly taxol x 12 followed by radiation completed on <DATE>."
8	surgery_type_left	"The patient is undergoing a lumpectomy on the left side."	"======== INITIAL CONSULT NOTE: <DATE> ========== TREATMENT: -SURGERY: <DATE> left breast lumpectomy and SLNB with Dr. <SURNAME> -CHEMOTHERAPY: Oncotype Dx score of 32."
9	gravida	"Gravida (number of times pregnant): 2."	"Hormonal History: Age of Menarche: 10 Number of Years OCP: 20 Total number of pregnancies: 2 Number of live births: 1 Age at 1st child: <AGE IN YEARS> Review of systems: Constitutional: Negative for weight loss and fever/chills."
10	image_biop	"The patient has not had a prior ultrasound-guided biopsy."	"Gender: female CC: BRCA 1+ Referral Source: <FIRSTNAME> <SURNAME> Accompanied by: husband, <FIRSTNAME>, and mother in law HPI: <DATE> BRCA1+ Never had mammo/sono/MRI Past Medical History: Patient has a past medical history of H/O menorrhagia."

(b)

	Hypothesis Category	Hypothesis Statement	Top Candidate Premise
1	sln_bx_right	"No sentinel lymph node (SLN) biopsy was performed on the right side."	"Gender: female Planned Procedure: SURGICAL HYSTEROSCOPY WITH BIOPSY OF ENDOMETRIUM AND POLYPECTOMY AND DILATION AND CURETTAGE, N/A - Uterus PELVIC EXAMINATION UNDER ANESTHESIA, N/A - Pelvis HTN, HLD, NIDDM, right breast ca s/p excision <DATE> without LN, obese, >4 METS ET."

2	age	"The patient is <AGE IN YEARS> years old."	"The Mount Sinai Hospital New York, NY 10029 PATIENT NAME: <SURNAME>, <FIRSTNAME> MRN: <MRN> VISIT NUMBER: <VISITNO> DATE OF SERVICE: <DATE> DATE OF BIRTH: <DATE> <FIRSTNAME> <SURNAME> is a <AGE IN YEARS>-year-old woman who presents having been diagnosed with a radial scar of the right breast." *Note: The age listed in the note matched the hypothesis statement, but both were one year older than the age listed in the database.*
3	prin_finding_right	"Principal finding, right side: ILC/LCIS."	"The Mount Sinai Hospital New York, NY 10029-6574 PATIENT NAME: <SURNAME>, <FIRSTNAME> MRN: <MRN> ACCOUNT NUMBER: <ACCNUM> ADMIT DATE: <DATE> DATE OF PROCEDURE: <DATE> SURGEON: <FIRSTNAME> <SURNAME>, M.D. PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ of the right breast."
4	prior_surg	"The patient has had prior surgery."	"Dear Dr. <SURNAME>, I had the pleasure of seeing Ms. <SURNAME> for a postop visit at the <FACILITY NAME> Breast Center today after a left lumpectomy on <DATE>." *Note: The note is referring to the surgery that happened on the index date as though it is prior.*
5	para	"Para (number of pregnancies reaching viable gestational age, including live births and stillbirths): 1."	"ASSESSMENT / PLAN: <AGE IN YEARS> y.o F PMH HTN, breast cancer s/p Bilateral total mastectomy, left axillary lymph node dissection, with TE reconstruction - regular diet - PO pain control - dc IVF - continue clinda - lovenox dvt ppx - SW for VNS - home tomorrow"
6	genetics	"The patient is BRCA (+)."	"female with PMHx of Lt Breast invasive ductal cancer (ER +, PR +, Her 2 -ve) s/p Left mastectomy on <DATE> s/p chemo (AC-T) and radiation last on <DATE>, now on exemestane, IDA , Migraine and Hypothyroid presenting with acute febrile illness with associated myalgia and chronic headaches."
7	personal_hx	"The patient has a history of breast cancer."	"Describes two days of feeling a disconnect Between head and feet when walking... Past Medical History: Ms. <FIRSTNAME> <SURNAME> has a past medical history of Thyroid dysfunction; Breast cancer (<DATE>); Insomnia; Arrhythmia (x 1 episode); and Cancer (<DATE>)." *Note: Breast cancer from index date referred to as though it is in the past.*
8	palpable_left	"The left-side mass was palpable."	"**** THIS IS AN ADDENDUM REPORT **** CASE: <CASENUM> PATIENT: <FIRSTNAME>

			<SURNAME> **Addendum ** TISSUE SUBMITTED: LEFT BREAST 10 O'CLODK 9 CM FN PALPABLE 1 X 0.5 X 1 CM MASS FINAL DIAGNOSIS: LEFT BREAST, PALPABLE MASS @10:00 9 CM FN, ULTRASOUND GUIDED CORE BIOPSY - Angiolipoma."
9	ax_dissect_left	"An axillary dissection was performed on the left side."	"Gender: female Planned Procedure: Side: Bilateral Procedure: BL mastectomy+reconstruction and L axillary dissection <AGE IN YEARS>F BrCa, GERD Denies PONV, OSA, former smoker Mets>4 Active Hospital Problems: There are no active hospital problems to display for this patient."
10	surgery_type_right	"The patient is undergoing a mastectomy on the right side."	"DIAGNOSIS: Right breast cancer PROCEDURE AND DATE: right breast NL partial mastectomy with tissue transfer, and right SLNBx on <DATE> PATHOLOGY: 2.5 cm IDC, with negative margins, and 1/3 LN with carcinoma."

Table 4: Top causes of error for each hypothesis category. The top 20 error sentences from each category (defined as sentences for which the MES was high but the hypothesis was false) were examined by two different reviewers to establish the most common cause(s) of error. Reviewer 1 (CG) was knowledgeable about the project but had no specific training in breast oncology. Reviewer 2 (KP) had a deeper knowledge of breast oncology and was part of the team that curated the BSD.

	Hypothesis Category	Laterality	Number of "Errors" Extracted Correctly from Text (n=20)	Most Common Cause(s) of Error
1	site		14, 8	Surgery only performed on one breast, but historical procedure (mammogram, reconstruction, or past surgery) described for the other breast. Former biopsy mistaken for surgery.
2	age		12, 13	Incorrect patient age stated in note.
3	reconst		9, 4	Reconstruction referenced in note but not found in database; often due to planned reconstruction that had not yet been carried out.
4	palpable_right	X	13, 13	Mass described as palpable in note but not in database (may refer to different mass). Palpable lymph nodes mistaken for palpable breast masses.
5	palpable_left	X	16, 14	See "palpable_right".
6	path_dx		6, 5	Breast cancer described in historical context but interpreted as current. Equally common: terms like "papilloma" and "carcinoma in situ" interpreted as "cancer". Other type of cancer mistaken for breast cancer.
7	neo_adj		13, 10	Neoadjuvant therapy described in historical or planned/future context but interpreted as current.
8	genetics		3, 1	Statements including terms "positive" or "+" used in other contexts (e.g., "HER2+" or "positive nodes") interpreted as BRCA+.
9	smoking		19, 19	Smoking status extracted correctly in nearly all cases from notes but doesn't match what is in database.

10	fhx_breast		11, 12	Family history of another type of cancer mistaken for breast cancer. Patient's history of breast cancer mistaken for family history.
11	gravida		7, 7	Number corresponding to hypothesis statement found elsewhere in note.
12	sln_bx_right	X	12, 2	Other types of biopsy described in pathology reports are mistakenly interpreted as referring to a sentinel lymph node biopsy. Previous surgery described instead of current.
13	sln_bx_left	X	15, 6	See "sln_bx_right".
14	hormone_use		11, 3	Other drug names interpreted as referring to hormone use - chemotherapy drug names often caused this problem. The field "Hormone Use:" was often written in the note but left blank yet interpreted as positive.
15	grade_left	X	8, 7	Statements describing left-side breast mass in general are interpreted as specifying a grade, even when no grade information is present. Radiology results describing mass interpreted as grade.
16	grade_right	X	5, 5	See "grade_left".
17	lvi_left	X	4, 2	References to lymphedema, lymphadenopathy, and lymph node metastases interpreted as specifying lymphovascular invasion (LVI); these are indeed related conditions.
18	lvi_right	X	3, 1	See "lvi_left".
19	para		0, 0	Unclear in most cases. Sometimes other numbers, such as dates, cancer stage, etc. mistaken for para.
20	margins_left	X	16, 16	Surgery results in negative margins but is follow-up from earlier surgery with positive margins. Reference to earlier positive margins in note causes error.
21	margins_right	X	16, 14	See "margins_left".
22	image_biop		19, 2	Other references to biopsies, some historical, interpreted as relevant to current surgery. These included non-breast biopsies, such as lymph node biopsies.
23	prin_finding_right	X	3, 2	Different specific diagnoses, especially abbreviations (e.g., "ILC" vs. "DCIS") mistaken for each other.
24	fertility_rx		3, 3	Other types of patient history ("history of...") interpreted as patient's having a history of fertility treatment. Adjuvant or previous tamoxifen/aromatase inhibitor therapies mistaken for fertility drugs.

25	status		6, 6	Abbreviations such as "D.O.S." or "D.D." interpreted as "D.O.D. (dead of disease)". Some of these were also extracted correctly but did not match database.
26	prin_finding_left	X	4, 4	See "prin_finding_right".
27	ax_dissect_right	X	15, 5	Model confused sentinel lymph node biopsy or lymph node core biopsy for axillary dissection. Planned procedure interpreted as axillary dissection.
28	sln_bx_left_n	X	4, 3	Misinterpretation of numbers and terms positive/negative from elsewhere in sentence. Difficulty interpreting fractions like 0/3. Confused core biopsy for SLN biopsy.
29	surgery_type_left	X	13, 3	Historical surgeries misinterpreted as referring to current surgery. Confused partial mastectomy for full mastectomy.
30	sln_bx_right_n	X	3, 3	See "sln_bx_left_n".
31	ax_dissect_left	X	19, 3	See "ax_dissect_right".
32	fhx_ovar		12, 15	Many of these were extracted correctly but did not match what was recorded in the database. Models sometimes confused genetic susceptibility for ovarian cancer for history of ovarian cancer.
33	surgery_type_right	X	15, 3	See "surgery_type_left".
34	prior_surg		4, 7	Current surgery misinterpreted as prior surgery. (This was technically correct in some cases due to records of follow-up visits' being included in the analysis.) Models also picked up surgeries unrelated to breast.
35	ax_dissect_left_n	X	0, 0	Similar to "sln_bx_left_n", errors most often occurred due to misinterpretation of numbers and the terms "positive" and "negative" from elsewhere in the sentence.
36	ax_dissect_right_n	X	0, 0	See "ax_dissect_left_n".
37	her2_left	X	5, 5	References to "HER2" interpreted as positive by default. Shorthand terms like +/+/+, which refer to ER/PR/HER2, ignored.
38	personal_hx		17, 12	Family history mistaken for personal history. Current cancer mistaken for personal history.
39	er_left	X	9, 7	Similar to "her2_left", references to "ER" interpreted as positive by default. Model unable to infer that "triple negative" implies ER negative.
40	her2_right	X	4, 4	See "her2_left".
41	er_right	X	12, 12	See "er_left".

42	pr_left	X	5, 0	Similar to "her_left", references to "PR" interpreted as positive by default. Reference to ER interpreted as PR.
43	pr_right	X	8, 0	See "pr_left".

Supplementary Information

Supplementary Table 1: Generation of hypothesis statements based on structured registry data, Part 1 (statements referring to the whole patient). The abbreviation "E" refers to an entailed statement, "C" refers to a contradictory statement, and "N" refers to a neutral statement.

Registry Field	Description	Statement Format	Generation Process
age	Age in years	"The patient is <age> years old."	One E was generated by filling in <age> with the correct age, and five Cs were generated using random incorrect ages between 18 and 90 years.
path_dx	Final post-operative diagnosis, based on pathology report	"Patient's final post-operative diagnosis: <dx>."	Four statements were generated by filling in <dx> with the terms "benign", "benign high-risk", "cancer", and "malignancy". If the true database finding was "benign", the four statements were labeled ECCC. If it was "benign-high risk", the statements were labeled EECC. If it was "cancer", the four statements were labeled CCEE. If it was blank, meaning there was no finding (for example, if no biopsy was performed), the statements were labeled CCCC.
fhx_breast	Family history of breast cancer	(1) "There is a family history of breast cancer." (2) "There is no family history of breast cancer." (3) "The family history of breast cancer is unknown since the patient was adopted."	Statements 1-3 at left were generated for each patient. If the database field was "yes", they were labeled ECC. If it was "no" or was left blank, they were labeled CEC. If it was "none-adopted", they were labeled NNE.
fhx_ovar	Family history of ovarian cancer	(1) "There is a family history of ovarian cancer." (2) "There is no family history of ovarian cancer." (3) "The family history of ovarian cancer is unknown since the patient was adopted."	See fhx_breast.
gravida	Number of times pregnant to date	"Gravida (number of times pregnant): <n>."	One E was generated by filling in <n> with the correct number, and three Cs were

			generated using random incorrect numbers between 0 and 10.
para	Number of pregnancies reaching viable gestational age to date	"Para (number of pregnancies reaching viable gestational age, including live births and stillbirths): <n>."	See gravida.
hormone_use	History of hormone use	(1) "The patient has a history of hormone use (i.e. hormone-containing birth control and/or hormone replacement therapy)." (2) "The patient has no history of hormone use (i.e. hormone-containing birth control and/or hormone replacement therapy)."	Statements 1 and 2 were generated for each patient. If the database field contained a "1", the two statements were labeled EC. If it contained a "-1", they were labeled CE. If it contained a "0", they were labeled NN.
fertility_rx	History of fertility treatment	(1) "The patient has a history of fertility treatment." (2) "The patient has no history of fertility treatment."	See hormone_use.
smoking	Smoking history	(1) "The patient is a current smoker." (2) "The patient is a former smoker." (3) "The patient has never smoked." (4) "The patient's smoking history is unknown."	Statements 1-4 were generated for each patient. If the database field contained the term "current", the statements were labeled ECCC. If it contained "former", they were labeled CECC. If it contained "no", they were labeled CCEC. If it was blank, they were labeled NNNE.
image_biop	Prior ultrasound-guided biopsy	(1) "The patient has had a prior ultrasound-guided biopsy." (2) "The patient has not had a prior ultrasound-guided biopsy."	See hormone_use.
prior_surg	Prior surgery	(1) "The patient has had prior surgery." (2) "The patient has had no prior surgery."	See hormone_use.
personal_hx	Personal history	(1) "The patient has no relevant prior history." (2-13) "The patient has a history of <x>"	Statements 1-13 were generated for each patient. A patient could have zero, one, or more than one true statement. For statements 1-9, if the relevant term was

		where <x> was (2) breast cancer (3) ovarian cancer (4) DCIS (5) ADH (6) ALH (7) radial scar (8) LCIS (9) cancer (not breast or ovarian) (10) benign breast masses (11) fibroadenoma (12) phyllodes tumor (13) papilloma.	present in the field, the corresponding statement was labeled E; if not, it was labeled C. Once the terms corresponding to statements 1-9 were accounted for, if the field was empty, statements 10-13 were labeled CCCC. If it was not empty, statement 10 was labeled E and statements 11-13 were labeled E or C depending on whether the string for each statement was present.
genetics	Genetic susceptibility to breast cancer	(1) "The patient is BRCA (+)." (2) "The patient is BRCA positive." (3) "The patient is BRCA (-)." (4) "The patient is BRCA negative." (5) "The patient's BRCA status is unknown." (6) "The patient has a genetic susceptibility to breast cancer." (7) "The patient has no known genetic susceptibility to breast cancer."	Statements 1-7 were generated for each patient. Depending on the patient's BRCA status, the first five statements could be labeled CCEEC (if the patient was BRCA-negative), EECCC (if the patient was BRCA-positive), or NNNNE (if the patient's BRCA status was unknown). Statements 6 and 7 could be labeled EC (if the patient had a documented genetic susceptibility to breast cancer in any of the three database fields described below) or CE (if there was no documented susceptibility). Note: There were three database fields that together summarized a patient's genetic status. The first summarized BRCA status (1=positive, 0=negative, NA=unknown). The second summarized genetic susceptibility via a panel test (1=positive, 0=negative, NA=unknown). The third summarized the presence of a variant of uncertain significance (VUS) (1=mentioned, 0=not mentioned).
site	Side of surgery (right, left, or both)	(1) "The patient is undergoing surgery on the right breast." (2) "The patient is undergoing surgery on the left breast." (3) "The patient is undergoing surgery on both breasts."	Statements 1-3 were generated for each patient. If the database field contained "Left", they were labeled CEC. If it contained "Right", they were labeled ECC. If it contained "Bilateral", they were labeled EEE. If it was blank, they were labeled NNN.
neo_adj	Neo-adjuvant chemotherapy	(1) "The patient received neo-adjuvant	Statements 1 and 2 were generated for each patient. If the database field

		chemotherapy." (2) "The patient has not received neo-adjuvant chemotherapy."	contained "Yes", they were labeled EC; if "No", they were labeled CE.
reconst	Breast reconstruction status	(1) "The patient has undergone a breast reconstruction." (2) "The patient has not undergone a breast reconstruction."	See neo_adj.
status	Patient's vital status at time of documentation	"Patient status: <status>" where <status> could be "NED (no evidence of disease)", "LWD (living with disease)", "DOC (diagnosed with other cancer)", "DOD (dead of disease)", or "unknown".	A statement incorporating each of the five possible status strings was generated for each patient. The statement containing the patient's true status was labeled E; the other four were labeled C.

Supplementary Table 2: Generation of hypothesis statements based on structured registry data, Part 2 (statements generated separately for the left and right breasts). The variable <side> was filled in with either "left" or "right". Information from the database field for "site" (Table 1) was used to determine which statements to generate. If a patient was having surgery on both sides, statements about both sides were generated. Abbreviations: "E" refers to an entailed statement, "C" refers to a contradictory statement, and "N" refers to a neutral statement.

Registry Field	Description	Statement Format	Generation Process
surgery_type	Surgery type	(1) "The patient is undergoing a mastectomy on the <side> side." (2) "The patient is undergoing a lumpectomy on the <side> side." (3) "The patient is undergoing breast conserving therapy on the <side>	Statements 1-6 were generated for each patient. The database field could contain one of the following strings, in which case the statements were labeled as shown: "Mastectomy" (ECCCCC), "BCT" (CEECCC), "Needle Biopsy" (CCCECC), "Axillary Excision" (CCCCEC), or "Chest Wall Excision" (CCCCCE). If the field was empty, the statements were labeled CCCCCC.

		side." (4) "The patient is undergoing a needle biopsy on the <side> side." (5) "The patient is undergoing an axillary excision on the <side> side." (6) "The patient is undergoing a chest wall excision on the <side> side."	
prin_finding	Principal finding based on surgery	"Principal finding, <side> side: <finding>."	The "principal finding" field was the most difficult to parse because it was a string field: hundreds of different strings were stored there. For each patient, we included one statement in which <finding> was the correct finding and labeled it E. We also generated five additional statements in which <finding> was a randomly chosen string from among all strings in the database, ensuring that the randomly chosen strings did not match the patient's true finding. Those were labeled C.
grade	Grade of mass	"The <side> side mass is <grade>."	Different types of masses use different grading systems. There were six possible descriptors in this field: "well differentiated", "moderately differentiated", "poorly differentiated", "low grade", "moderate grade", and "high grade". Six statements were generated for each patient, replacing the <grade> token by each of these six strings. If the string matched what was written for the patient, it was labeled E; if it did not, it was labeled C. If no grade information was present in the database, no statements were created.
lvi	Lymphovascular invasion	(1) "There is definite lymphovascular invasion on the <side> side." (2) "There is no lymphovascular invasion on the <side> side." (3) "Suspicious for lymphovascular invasion on the <side> side."	Statements 1-3 were generated for each patient. If the database field contained "yes", they were labeled ECE. If it contained "no", they were labeled CEC. If it contained "suspicious", "indeterminate", or "cannot be ruled out", they were labeled CCE.

er	Estrogen receptor status	(1) "The <side> side breast mass is ER (estrogen receptor) positive." (2) "The <side> side breast mass is ER (estrogen receptor) negative." (3) "The ER (estrogen receptor) status of the <side> side breast mass is indeterminate/unknown."	Statements 1-3 were generated for each patient. If the database field contained "Positive", they were labeled ECC. If it contained "Negative", they were labeled CEC. If it contained "Indeterminate", they were labeled CCE.
pr	Progesterone receptor status	(1) "The <side> side breast mass is PR (progesterone receptor) positive." (2) "The <side> side breast mass is PR (progesterone receptor) negative." (3) "The PR (progesterone receptor) status of the <side> side breast mass is indeterminate/unknown."	See er.
her2	HER2 status	(1) "The <side> side breast mass is HER2 positive." (2) "The <side> side breast mass is HER2 negative." (3) "The HER2 status of the <side> side breast mass is indeterminate / unknown."	Statements 1-3 were generated for each patient. If the database field contained "Positive", they were labeled ECC. If it contained "Negative", they were labeled CEC. If it contained "Indeterminate" or "Equivocal", they were labeled CCE.
margins	Margin status post-surgery	(1) "The margins on the <side> side were negative." (2) "The margins on the <side> side were close." (3) "The margins on the <side> side were positive."	Statements 1-3 were generated for each patient. If the database field contained "Negative", they were labeled ECC. If it contained "Close", they were labeled CEC. If it contained "Positive", they were labeled CCE.
sln_bx	Sentinel node biopsy performed	(1) "A sentinel lymph node (SLN) biopsy was performed on the <side> side." (2) "No sentinel lymph node (SLN) biopsy was performed on the <side> side."	See neo_adj.

sln_bx_n	Sentinel node biopsy information	"<n_nodes> <side>-side sentinel lymph <s1> biopsied, of which <s2> positive."	Multiple statements were generated. <n_nodes> was replaced by either the true number or that number plus one. <s1> was replaced by "node was" if <n_nodes> was 1 and "nodes were" otherwise. <s2> was replaced by a phrase corresponding to the number of nodes found to be positive: "none were", "1 was", or "<n> were" (if the number of positive nodes, n, was greater than one). The number of positive nodes could be the true number of positive nodes, the true number of nodes biopsied, or zero. A statement was labeled E only if both the number of nodes biopsied and the number of positive nodes were correct; otherwise it was labeled C.
ax_dissect	Axillary dissection performed	(1) "An axillary dissection was performed on the <side> side." (2) "No axillary dissection was performed on the <side> side."	See neo_adj.
ax_dissect_n	Axillary dissection information	"<n_nodes> <side>-side axillary lymph <s1> biopsied, of which <s2> positive."	See sln_bx_n.
palpable	Palpability of mass	(1) "The <side>-side mass was palpable." (2) "The <side>-side mass was not palpable."	Statements 1 and 2 were generated for each patient. If the database field contained "Yes", they were labeled EC. If it contained "No", they were labeled CE.

Supplementary Table 3: Dependence of model performance on availability of patient data. Patients were divided into three groups: those with fewer than 100 sentences, those with 100-1000 sentences, and those with more than 1000 sentences.

(Full table available in attached csv file.)