Towards Structuring Real-World Data at Scale: Deep Learning for Extracting Key Oncology Information from Clinical Text with Patient-Level Supervision

Sam Preston,1,* Mu Wei,1,* Rajesh Rao,1 Robert Tinn,1 Naoto Usuyama,1 Michael Lucas,1 Roshanthi Weerasinghe,2 Soohee Lee,2 Brian Piening,3 Paul Tittel,3 Naveen Valluri,1 Tristan Naumann,1 Carlo Bifulco3,∗∗ and Hoifung Poon1,∗∗

1 Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA , 2 Providence Research Network, Providence St Joseph’s Health, Portland, OR, 97213, USA and 3 Providence Genomics & Earle A. Chiles Research Institute, 4805 NE Glisan, Portland, OR, 97213, USA

*These authors contributed equally to this work.
**Corresponding authors: Carlo.Bifulco@providence.org and hoifung@microsoft.com

Abstract

Objective: The majority of detailed patient information in real-world data (RWD) is only consistently available in free-text clinical documents. Manual curation is expensive and time-consuming. Developing natural language processing (NLP) methods for structuring RWD is thus essential for scaling real-world evidence generation.

Materials and Methods: Traditional rule-based systems are vulnerable to the prevalent linguistic variations and ambiguities in clinical text, and prior applications of machine-learning methods typically require sentence-level or report-level labeled examples that are hard to produce at scale. We propose leveraging patient-level supervision from medical registries, which are often readily available and capture key patient information, for general RWD applications. To combat the lack of sentence-level or report-level annotations, we explore advanced deep-learning methods by combining domain-specific pretraining, recurrent neural networks, and hierarchical attention.

Results: We conduct an extensive study on 135,107 patients from the cancer registry of a large integrated delivery network (IDN) comprising healthcare systems in five western US states. Our deep learning methods attain test AUROC of 94-99% for key tumor attributes and comparable performance on held-out data from separate health systems and states.

Discussion and Conclusion: Ablation results demonstrate clear superiority of these advanced deep-learning methods over prior approaches. Error analysis shows that our NLP system sometimes even corrects errors in registrar labels. We also conduct a preliminary investigation in accelerating registry curation and general RWD structuring via assisted curation for over 1.2 million cancer patients in this healthcare network.

Key words: Natural Language Processing (L01.224.050.375.580); Data Mining (L01.313.500.750.280.199); Medical Oncology (H02.403.429.515); Neoplasm Staging (E01.789.625)

INTRODUCTION

Electronic medical records (EMRs) offer an unprecedented opportunity to harness real-world data (RWD) for accelerating progress in clinical research and care [22]. By tracking longitudinal patient care patterns and trajectories, including diagnoses, treatments and clinical outcomes, we can help assess drug efficacy in real-world settings, facilitate post-market surveillance, and speed up clinical trial...
Preston, Wei et al.

Fig. 1: Cancer patients typically have many clinical documents for a tumor diagnosis, with key information scattered around.

Patient clinical document time series

![Diagram of patient clinical document time series]

Patients with various types of clinical reports over time, illustrating the need for structured information extraction.

recruitment. However, pertinent information about patients often resides in clinical text, such as pathology assessments, radiology assessments, and clinical progress notes. Manual curation to structure such text is expensive and hard to scale.

Natural language processing (NLP) can help accelerate manual curation [26]. In recent years, there have been rapid advances in general-domain NLP, where state-of-the-art deep neural networks, such as transformer-based models, have demonstrated remarkable success across a wide range of applications [4, 14]. Training these sophisticated models, however, typically requires a large number of annotated examples. By contrast, prior work in clinical NLP is often limited to annotating small datasets and training simpler methods [10]. Due to the scarcity of qualified domain experts, annotation is usually conducted on a small collection of notes, often from a single institution. Moreover, to make learning easier, these explorations typically restrict annotation to single sentences or single notes. For example, Kehl et al. [10] shows promising results on applying NLP to accelerate real-world evidence generation in oncology. However, while their annotation effort is relatively large among similar prior efforts, their test set contains only 109 patients (1,112 patients in the entire annotated dataset). The notes are limited to radiology reports for lung cancer from a single institution. Their exploration is limited to convolutional neural networks, which does not leverage the latest NLP advances, such as language model pretraining [2, 7].

In this paper, we propose to bootstrap deep learning for structuring real-world data (RWD) by using readily available registry data. Medical registries are routinely collected for various diseases, with oncology being a prominent example. In the U.S., cancer is a reportable disease and cancer centers are required to curate patient information per national accreditation and clinical quality requirements. By matching registry entries with their corresponding EMR data, we can assemble a large dataset for training and evaluating state-of-the-art deep NLP methods.

Gao et al. [5, 6] also leverage registry data for supervision. However, like Kehl et al. [10], they restrict classification to individual pathology reports, and exclude tumors associated with multiple reports. Similarly, Percha et al. [19] focuses on classifying individual sentences for breast cancer surgery information. Such methods are not applicable to the prevalent cases where information is scattered across multiple clinical documents and note types (e.g., pathology reports, radiology reports, progress notes). Often, information in a single document (e.g., discussion of a malignant site) is insufficient and additional context is required for identifying the correct diagnosis or staging information.

To the best of our knowledge, our study is the first to explore cross-document medical information extraction, using registry-derived patient-level supervision to train deep NLP methods. Such patient-level supervision is inherently challenging to use as it comprises only annotations associated with a tumor diagnosis, which are not attributable to individual sentences or documents. Each patient may have dozens of clinical documents, yielding very long input text spans that are difficult to process for standard deep learning methods. Additionally, the collection of clinical documents spans decades and varies in completeness. Nevertheless, we found that the scale of such self-supervised data more than compensates for its noise and technical challenges, and our models attain high performance (AUROC 94-99%) for extracting core tumor attributes such as site, histology, clinical/pathological staging.

Unlike in simplistic settings of prior studies [5], sophisticated deep learning methods substantially outperform simplistic approaches, with our top-performing model combining cutting-edge techniques such as transformers [4], domain-specific pretraining [7], recurrent neural networks [3], and hierarchical attention [29]. Our method naturally handles longitudinal information and experiments show that incorporating multiple document types significantly improves performance. Neural attention can be used to pinpoint relevant text spans as extraction rationale and provenance, which facilitate model interpretation and rapid validation by human experts. Our model trained on a health system in one state performed comparably for patients from different states, health systems and EMR configuration, suggesting good generalizability.

While our work is motivated by structuring real-world data, our method can also be used to accelerate registry curation. Our deep learning model not only performs well in abstraction, but also attains high accuracy in case finding (identifying patients for cancer registry), thus paving the way for end-to-end assisted cancer registry curation.
Towards Structuring Real-World Data at Scale

Fig. 2: A general neural architecture for medical abstraction: clinical documents are concatenated by chronological order and converted into a token sequence, which is then transformed into a sequence of neural vectors by the embedding and contextualization modules, before being converted into a fixed-length feature vector by an aggregation module for final classification.

Fig. 3: Relevant information for medical abstraction may manifest in myriad variations, as seen in specification of tumor site in breast cancer with laterality and clockwise position.

MATERIALS AND METHODS

Abstraction

Medical abstraction can be formulated as information extraction in NLP. Given clinical text $T$ for a patient and attribute $A$, the goal is to extract $A$’s value as described in $T$ (or absence thereof), which can be framed as a multiclass classification problem. In most prior work, $T$ is a sentence or a clinical note, and $A$’s value only has a few choices (e.g., presence or absence of active cancer [10]). By contrast, we consider the most general setting, where $T$ comprises all notes for a patient and $A$’s range may number in hundreds. E.g., there are 310 classes for tumor site and 556 for histology in ICD-O-3, and a patient may have many notes (Figure 1).

In general, abstraction presents substantial challenges for NLP systems. Relevant information may manifest in myriad variations (Figure 3). Name entity recognition (NER) is not enough, as abstraction is more about extracting underlying relations. E.g., abstracting tumor site is not about recognizing site mentions, but to determine if the patient has malignancy at the given site on a given date (Figure 4, a). Moreover, abstraction may require information integration across multiple clinical documents (Figure 4, b).

With patient-level supervision from medical registries, our machine learning setting can be regarded as a form of distant supervision or more generally self-supervision [33], as the labels cannot be attributed to a sentence or even a clinical document. However, given the aforementioned complex linguistic phenomena in medical abstraction, we do not generate noisy training examples by associating a label with a specific text span (e.g., individual sentences with the presence of relevant entities), as in standard distant supervision. Instead, we combine all clinical documents for a patient as input and rely on the deep-learning method to automatically identify pertinent sentences and notes.
Deep Learning for Medical Abstraction

Traditional clinical NLP systems are often rule-based, e.g., by leveraging regular expressions and domain lexicons from ontologies [15]. They require significant efforts to build and may be vulnerable to linguistic variations and ambiguities. Consequently, machine-learning methods have seen increasing adoption [30]. Traditional learning-based NLP methods require users to provide feature templates for classification, whereas modern deep learning methods forgo this requirement and can automatically transform input text into a neural feature representation (a real-number vector) [27, 2, 5, 6, 10].

Figure 2 shows a general deep learning architecture for medical abstraction. Medical documents are ordered temporally and converted into a sequence of sentences. They are tokenized and converted into a neural representation by an embedding module where each token is turned into a real-number vector. The vectors are then updated by a contextualization module and combined into a fixed-length feature vector by an aggregation module, which the classification module uses as input to produce the final classification.

In prior work applying deep learning to medical abstraction, the embedding module generally uses simple context-free embedding such as word2vec [16] or GLoVE [18]. Contextualization is usually done by convolutional neural network (CNN), which runs a sliding window over the tokens and generates output vectors using a shared neural network, with aggregation done by pooling.

Recently, there has been substantial progress in deep NLP methods. Transformer [24], with its multi-layer, multi-head self-attention mechanism, has proven very effective in modeling long-range dependencies and leveraging GPU parallelism. Contextualized embedding from language model pretraining [4, 20] is much more powerful than context-free embedding such as Word2Vec and GLoVe in extracting semantic information from unlabeled text and modeling variations/ambiguities. While the bulk of pretraining work focuses on general domains such as newswire and Web, domain-specific pretraining has proven beneficial for specialized domains such as biomedicine, by prioritizing learning of biomedical terms in relevant biomedical contexts [17, 13, 7].

In this paper, we conduct a thorough study of advanced deep NLP techniques in medical abstraction (Figure 2, blue). Some prior work investigated deep NLP in simplistic settings (e.g., classifying individual pathology reports) and concluded that advanced techniques such as transformer does not help their tasks [5]. By contrast, we find that in the real-world setting of cross-document medical abstraction, advanced NLP techniques can confer significant benefit in combating the prevalent noises and linguistic complexities.

For embedding, we use the state-of-the-art biomedical neural language model PubMedBERT [7]. The input to a neural language model consists of text spans, such as sentences, separated by special tokens [SEP]. To address the problem of out-of-vocabulary words, neural language models generate a vocabulary from subword units [11, 23], by greedily identifying a small set of subwords that can compactly form all words in a given corpus. BERT [4] is a state-of-the-art language model based on transformer [24], which is pretrained by predicting held-out words in unlabeled text. While most BERT models were pretrained on general-domain text [4, 14], PubMedBERT instead uses a biomedicine-specific vocabulary and was pretraining on biomedical literature from scratch. We also pretrain an oncology-specific OncoBERT on EMRs from over one million cancer patients and explore its use in oncology abstraction.
Self-attention requires pairwise computation among tokens, which scales quadratically in input text length. Consequently, standard BERT models typically limit input length (e.g., 512 tokens). This is not a problem for restricted settings such as sentence-level or document-level abstraction in prior work, but it poses a substantial challenge in the general setting, as patient-level cross-document input has a median length of over 4,000 tokens. To handle such long text, we use gated recurrent unit (GRU) [3] for contextualization and hierarchical attention network (HAN) [29] for aggregation. GRU helps propagate information beyond BERT’s default length limit and HAN provides better aggregation than pooling by weighing relevant tokens higher. The classification module is a standard linear layer followed by \text{softmax}, which produces multinomial probabilities among possible labels.

Case Finding

In medical abstraction, we are given cancer patients and asked to extract key tumor attributes. By contrast, the goal of case finding is to determine if a patient should be included for cancer registry. This can be framed as binary classification over a patient’s clinical documents from a given day. We use the same architecture as in Figure 2 and find it similarly effective. (The models are learned separately for case finding vs. abstraction. We conducted preliminary experiments on multi-task learning but didn’t find significant difference in performance, as each task has abundant training data.)

Case finding poses a distinct self-supervision challenge. We can easily identify positive examples from the registry (patients with their diagnosis dates). However, it is less clear how to identify negative examples. We explore two self-supervision schemes. Initially, we randomly sample non-cancer patients and days from their medical history with pathology reports. This yields a classifier with good sensitivity (recall), but often incorrectly flags pre-diagnosis days for a cancer patient, causing a high false-positive rate. To address this problem, we experiment with adding hard negative examples from cancer patients by sampling days before diagnosis. The resulting classifier not only distinguishes cancer patients from non-cancer patients, but also identifies the time of initial diagnosis, as required for case finding. Together with abstraction, we can thus help accelerate cancer registry curation end-to-end.

Human Subjects / IRB, Data Security & Patient Privacy

This work was performed under the auspices of an Institutional Review Board (IRB)-approved research protocol (Providence protocol ID 2019000204) and was conducted in compliance with Human Subjects research and clinical data management procedures—as well as cloud information security policies and controls—administered within Providence St. Joseph Health. All study data were integrated, managed and analyzed exclusively and solely on Providence-managed cloud infrastructure. All study personnel completed and were credentialed in training modules covering Human Subjects research, use of clinical data in research, and appropriate use of IT resources and IRB-approved data assets.

Results

We conduct experiments using data from a large integrated delivery network (IDN) with over 28 distinct cancer care centers across U.S. states. We assemble a dataset with patient-level supervision by matching comprehensive EMR records (including all free-text clinical documents in scope here) and cancer registry records. Patients without a digitized pathology report within 30 days of diagnosis are skipped. This yields a total of 135,107 patients spanning multiple U.S. states, between 2000-2020. We use patients in Oregon for the initial exploration (\( n = 39,064 \), 29% of patients). We divide patients into ten random folds. We use six folds for training and development (\( n = 23,438 \)), two folds for test (\( n = 7,740 \)), and two folds for additional held-out test set (\( n = 7,881 \)). We reserve patients from Washington (\( n = 36,900 \)), as well as the remaining states (\( n = 59,143 \)) for further generalizability test, with a distinct health system being used in each state.

We use the ICD-O-3 ontology for tumor site and histology. For staging, we focus on solid tumors and follow AJCC guidelines for clinical and pathological staging. Both represent cancer progression using TNM classification (T: tumor size/location; N: lymph node status; M: metastasis). Clinical staging is based on initial diagnosis using medical imaging, clinical assessment and/or biopsy, whereas pathological staging incorporates more definitive assessment of the tumor size and spread. For simplicity and based on practical utility, we focus on classifying coarse categories (T: 0-4, in situ; N: 0 vs 1+; M: 0 vs 1).

For each attribute, we report standard area under the receiver operating characteristic curve (AUROC). For system comparison, however, AUROC might obscure key performance difference in the presence of imbalanced distribution (e.g., some sites appear much more frequently), so we evaluate area under the precision-recall curve (AUPRC). We also report accuracy for completeness. Precision and recall are also known as positive predictive value and sensitivity, respectively.

Main Results

Table 1 shows test results for extracting key oncology attributes. By incorporating state-of-the-art advances such as PubMedBERT, our deep learning system attains high performance across the board, even for tumor site and histology where the system has to distinguish among hundreds of labels. Despite the large parameter space, our system is robust in experiments, with standard deviation across two random runs smaller than 1% for all tasks.

Generalizability

To assess the generalizability, we evaluate on the held-out set and find that model performance is nearly identical. We further evaluate our model trained on the Oregon training set on patients from Washington and other states. See Table 2. The results are
Table 1. Test results for oncology abstraction by our deep learning system based on PubMedBERT.

| Tumor Site | Histology | Clinical T | Clinical N | Clinical M | Pathological T | Pathological N | Pathological M |
|------------|-----------|------------|------------|------------|----------------|----------------|----------------|
| 99.3       | 76.7      | 93.9       | 97.2       | 98.7       | 96.1           | 95.2           | 98.6           |
| 99.4       | 87.2      | 79.3       | 97.2       | 98.7       | 87.2           | 95.3           | 98.6           |
| 93.9       | 79.3      | 70.1       | 91.6       | 94.9       | 78.6           | 88.9           | 95.1           |

Table 2. Generalizability test: all results were obtained using our deep-learning models (based on PubMedBERT) trained on Oregon Training. Washington (WA) and other states all use different health systems. There is only slight degradation for most results, which bodes well for generalizability of our models. A notable exception is histology, with up to 9 points drop. Upon close inspection, this stems from divergence in curation standards on ambiguous cases, with registrars using different labeling granularity (e.g., non small-cell lung cancer vs. lung adenocarcinoma).

Table 3. Comparison of test AUPRC scores for oncology abstraction by various NLP systems. Ontology: ontology-aware rule-based system. BOW: logistic regression with bag-of-word features. OncoGloVe: 100-dimensional GloVe embedding pretrained on oncology notes.

| Tumor Site | Histology | Clinical T | N | M | Pathological T | N | M |
|------------|-----------|------------|---|---|----------------|---|---|
| Ontology   | 19.4      | 19.2       | - | - | -              | - | - |
| BOW        | 62.8      | 76.6       | 70.4 | 96.6 | 98.4         | 72.1 | 90.7 |
| OncoGloVe + CNN | 72.0   | 84.4       | 74.2 | 96.5 | 98.6         | 83.9 | 93.1 |
| OncoGloVe + HAN/GRU | 74.0   | 85.9       | 76.2 | 97.1 | 98.7         | 86.4 | 94.2 |
| BERT + HAN/GRU     | 75.1      | 86.2       | 77.0 | 96.6 | 98.4         | 86.4 | 94.4 |
| PubMedBERT + HAN/GRU (ours) | 76.7    | 87.2       | 79.3 | 97.2 | 98.7         | 87.2 | 95.2 |
| OncoBERT + HAN/GRU (ours) | 77.1    | 87.6       | 81.4 | 97.5 | 99.0         | 87.6 | 95.5 |

Table 3 compares our deep learning systems with prior approaches for medical abstraction. An ontology-aware rule-based system (matching against class lexicon and known aliases) performs poorly, demonstrating that entity recognition alone is inadequate for such challenging tasks. Deep-learning methods perform substantially better, with BERT-based approaches outperforming CNN, especially for the most challenging tasks such as site, histology, and clinical/pathological tumor (T) staging. HAN/GRU and transformer-based language models each contribute significantly, with our best system gaining 5.1 points for site, 3.2 points for histology, and 7.2 points for clinical T over GloVe-CNN.

Domain-specific pretraining is especially impactful. By pretraining entirely on oncology notes, OncoBERT further improves over PubMedBERT, which is already pretrained on biomedical text. Compared to general-domain BERT, our best system with OncoBERT gains 2.0 points for site and 4.4 points for clinical tumor (T) staging.
Ablation Study

We incorporate three types of clinical documents as input: pathology report, radiology reports, and operative notes. In ablation study, we find that having all three helps, presumably because this increases robustness in case some relevant notes are missing or not yet digitized (e.g., scanned PDFs). E.g., adding radiology reports on top of pathology reports increased AUPRC by 3.4 absolute points for tumor-site extraction, with the inclusion of operative notes providing an additional one point gain. E.g., using [-30, 30] days as input improves AUPRC by over four absolute points for pathological tumor (T) staging (87.2 vs 91.8).

Case Finding

|               | Train Positive Instances | Train Negative Instances | Test F1 |
|---------------|--------------------------|--------------------------|---------|
| Default       | 37,207                   | 13,123                   | 91.4    |
| + Hard Negatives | 37,207                   | 22,959                   | 97.3    |

Table 4. Comparison of test results in case finding with two self-supervision schemes.

We assemble a case-finding dataset using patients in the cancer registry. For positive cases, we identify cancer patients with at least a pathology report on the day of diagnosis. For negative cases, we randomly sample non-cancer patients. This yields 62,090 positive and 8,460 negative patients. We divide them into train/development/test by 60%/20%/20%, with 12,418 positive and 1,692 negative patients in the test set.

A patient may have clinical documents in multiple days. In case finding, a classification instance comprises a patient’s clinical documents in a given day, and the ultimate goal is to identify the moment of cancer diagnosis (when registry curation starts). For evaluation, we adopt a patient-level metric that mirrors real-world applications. For each patient, we return the first day with positive classification. For cancer patients, the case-finding decision is deemed correct if the first day of positive classification is within [-7, 30] days of diagnosis. For non-cancer patients, the case-finding decision is deemed correct if all classifications are negative. We report the F1 score which is the harmonic mean of precision (positive predictive value) and recall (sensitivity). Specifically, $F1 = 2/(1/\text{precision} + 1/\text{recall})$.

For self-supervision, we explore the two settings as described in the Method Section. In both cases, positive instances comprise cancer patients on the diagnosis date. By default, negative instances comprise of randomly chosen days among non-cancer patients. Additionally, we randomly sample days before diagnosis among cancer patients and add 9,836 instances as hard negative examples. With the base setting, we attain test F1 91.4. By incorporating hard negative examples, we substantially improve test F1 to 97.3, gaining six absolute points.

DISCUSSION AND CONCLUSION

Error Analysis

We conduct manual analysis on sample errors. Some stem from annotation inconsistency, where registrars actually agree with our system classifications upon close inspection. Others stem from missing notes. After adjusting for annotation inconsistency and missing input, the real test performance of our deep-learning system is even higher. For example, based on estimate from analyzing fifty error examples, the real test AUPRC for tumor site is about 91.6 (vs. 76.7).

Assisted Curation

We envision that NLP extraction can serve as candidates to help accelerate curation. The attention mechanism in transformer-based model provides a straightforward approach to identify extraction rationale. Effectively, the aggregate representation of the input text is the weighted sum of token representations in the top layer, with the weights (derived from self-attention to the special [CLS] token) signifying relative importance of individual tokens in the final classification decision. While there is no guarantee that attentions provide explanation [9], in practice we find that tokens with the highest attention weights are conducive to assisted curation and generally conform with what human experts would consider as extraction rationale. As an example, Figure 5 highlights tokens with high attention weights for the example text in Figure 4a and two variations. While the attention may not entirely align with individual human intuition, it broadly conforms with the extraction rationale and enables quick verification. Figure 6 shows a prototype we have developed for assisted curation, which is in beta test by selected clinical users. For each attribute, the interface displays the extraction rationale by highlighting individual note and text span with the highest neural attention weight for final classification. In preliminary studies, tumor registrars can verify a candidate extraction in 1-2 minutes, either ascertaining its correctness or fixing the label in the interface.
Fig. 5: Examples of observed attention patterns and predictions from the tumor site model. (a) shows the attention pattern for the example shown in Figure 4a, with darker color signifying higher attention weight. The tumor site model correctly identifies C22.1: Intrahepatic bile duct due to the cholangiocarcinoma histology (indicating cancer of the bile duct). To probe the model understanding further, inference was run on modified text. In (b), the description was changed to a generic “carcinoma” diagnosis. While the attention is more diffuse, the model places the highest attention on the “liver” section, and correctly identifies C22.0: Liver as the tumor site. In (c), the “carcinoma” diagnosis was moved to the “gallbladder” section, and the model now correctly identifies the site as C23.9: Gallbladder, with attentions now focusing on this section.

Limitations

Our study focuses on medical abstraction of key diagnosis information as curated in cancer registry. Future work should explore extraction of treatment and outcome information, as well as other diagnostic information such as biomarkers. Cancer registry focuses on complete curation of “analytical cases”, i.e., patients with both initial diagnosis and treatment occurring within a given healthcare system. The models may perform less well for patients who are initially diagnosed elsewhere and then referred to the given network, e.g., due to missing digitized reports. In many such cases, PDFs or scanned documents are still available. We are exploring the use of state-of-the-art document image understanding methods, such as LayoutLM [28], with initial promising results. Our immediate exploration of assisted curation focuses on accelerating case identification and medical abstraction, but it also opens up opportunities for interactive learning to continuously improve machine reading based on user feedback. In addition to improving abstraction accuracy, this can potentially help calibrate attention weights for extraction rationale [1]. Pretraining can also be further improved by incorporating domain knowledge such as UMLS [8, 31].

Towards Scaling Real-World Data Curation

Manual curation of complex clinical records and EHR data is expensive and time-consuming. The healthcare network represented in this study hires several dozens of full-time registrars for cancer registry abstraction. Curation is limited to analytic cases (i.e. those first treated in a given cancer center), which are legally required for reporting, thus skipping a large swath of patients. Despite such restrictions and significant investment, there is still significant delay for majority of the patients. To estimate the extent of curation backlog, we analyze two snapshots of cancer registry that are eight months apart. Among newly curated cases in the second snapshot, 23,670 are diagnosed before the first snapshot ends. They have a median of 324 days between diagnosis and the first snapshot end date. Many cases are curated over a year after diagnosis. By leveraging assisted curation with candidate abstractions generated by our deep NLP system, we can accelerate cancer registry abstraction and reduce backlog. Given promising results in the preliminary study, we are now exploring integration of assisted curation to the registry abstraction workflow.

NLP-based machine reading also helps scale real-world data curation. The healthcare network in our study has over 1.2 million cancer patients with digitized pathology reports within 30 days of diagnosis. However, only 135,107 of them have been curated in the cancer registry. By applying our NLP system to all patients, we instantly expand structured real-world data for the network by an order of magnitude. In future work, we plan to expand the scope of curation by applying self-supervised learning to extracting other key information for real-world evidence, such as treatments and key clinical outcomes [21, 25, 12, 32]. A particularly exciting research frontier lies in studying response to immunotherapy, such as check-point inhibitors (CPI). In preliminary study, we find...
that self-supervised NLP methods can immediately identify and abstract over an order of magnitude more CPI patients, compared to prior manual efforts that took many months.

**COMPETING INTEREST**

There is NO Competing Interest.

**References**

1. Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving machine attention from human rationales. In *Proceedings of EMNLP*, 2018.

2. Oliver J Bear Don’t Walk IV, Tony Sun, Adler Perotte, and Noémie Elhadad. Clinically relevant pretraining is all you need. *Journal of the American Medical Informatics Association*, 28(9):1970–1976, 06 2021.

3. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

5. Shang Gao, Mohammed Alawad, Michael Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, and Georgia D. Tourassi. Limitations of Transformers on Clinical Text Classification. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2021.

6. Shang Gao, John X. Qiu, Mohammed Alawad, Jacob D. Hinkle, Noah Schaefferkoetter, Hong Jun Yoon, Blair Christian, Paul A. Fearn, Lynne Penberthy, Xiao Cheng Wu, Linda Coyle, Georgia Tourassi, and Arvind Ramanathan. Classifying cancer pathology reports with hierarchical self-attention networks. *Artificial Intelligence in Medicine*, 101(September), 2019.

7. Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

8. Boran Hao, Henghui Zhu, and Ioannis Paschalidis. Enhancing clinical BERT embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

9. Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL*, 2019.
10. Kenneth L. Kehl, Haitham Elmarakeby, Mizuki Nishino, Eliezer M. Van Allen, Eva M. Lepisto, Michael J. Hassett, Bruce E. Johnson, and Deborah Schrag. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes from Radiology Reports. *JAMA Oncology*, 5(10):1421–1429, 2019.

11. Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

12. Hunter Lang and Hoifung Poon. Self-supervised self-attention by combining deep learning and probabilistic logic. In *Proceedings of AAAI*, 2021.

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

14. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

15. Stéphane M Meystre, Paul M Heider, Youngjun Kim, Matthew Davis, Jihad Obeid, James Madory, and Alexander V Alekseyenko. Natural language processing enabling COVID-19 predictive analytics to support data-driven patient advising and pooled testing. *Journal of the American Medical Informatics Association*, 29(1):12–21, 10 2021.

16. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

17. Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.

18. Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

19. Bethany Percha, Kereeti Pisapati, Cynthia Gao, and Hank Schmidt. Natural language inference for curation of structured clinical registries from unstructured text. *Journal of the American Medical Informatics Association*, 29(1):97–108, 11 2021.

20. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

21. Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, 2016.

22. Vivek A. Rudrapatna and Atul J. Butte. Opportunities and challenges in using real-world data for health care. *J Clinical Investigation*, 130(2), 2020.

23. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

24. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

25. Hai Wang and Hoifung Poon. Deep probabilistic logic: A unifying framework for indirect supervision. In *Proceedings of EMNLP*, 2018.

26. Chih-Hsuan Wei, Bethany R Harris, Donghui Li, Tanya Z Berardini, Eva Huala, Hung-Yu Kao, and Zhiyong Lu. Accelerating literature curation with text mining tools: a case study of using pubtator to curate genes in pubmed abstracts. *Database*, 2012.

27. Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. Deep learning in clinical natural language processing: a methodical review. *JAMIA*, 27(3), 2020.

28. Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of KDD*, 2020.

29. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

30. Wen-Wai Yim, Meliha Yetisgen, William P Harris, and Sharon W Kwan. Natural language processing in oncology: A review. *JAMA Oncology*, 2(6), 2016.

31. Sheng Zhang, Hao Cheng, Shikhar Vashisht, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Knowledge-rich self-supervised entity linking. In *arxiv*, 2021.

32. Sheng Zhang, Cliff Wong, Naoto Usuyama, Sarthak Jain, Tristan Naumann, and Hoifung Poon. Modular self-supervision for document-level relation extraction. In *Proceedings of EMNLP*, 2021.

33. Tiantian Zhu, Yang Qin, Yang Xiang, Baotian Hu, Qingcai Chen, and Weihua Peng. Distantly supervised biomedical relation extraction using piecewise attentive convolutional neural network and reinforcement learning. *Journal of the American Medical Informatics Association*, 28(12):2571–2581, 09 2021.