ARTIFICIAL INTELLIGENCE

original reports abstract

# Automated NLP Extraction of Clinical Rationale for Treatment Discontinuation in Breast Cancer

Matthew S. Alkaitis, MD, PhD[1,2]; Monica N. Agrawal, MS[1]; Gregory J. Riely, MD, PhD[3,4]; Pedram Razavi, MD, PhD[3,4]; and David Sontag, PhD[1]

**PURPOSE** Key oncology end points are not routinely encoded into electronic medical records (EMRs). We assessed whether natural language processing (NLP) can abstract treatment discontinuation rationale from unstructured EMR notes to estimate toxicity incidence and progression-free survival (PFS).

**METHODS** We constructed a retrospective cohort of 6,115 patients with early-stage and 701 patients with metastatic breast cancer initiating care at Memorial Sloan Kettering Cancer Center from 2008 to 2019. Each cohort was divided into training (70%), validation (15%), and test (15%) subsets. Human abstractors identified the clinical rationale associated with treatment discontinuation events. Concatenated EMR notes were used to train high-dimensional logistic regression and convolutional neural network models. Kaplan-Meier analyses were used to compare toxicity incidence and PFS estimated by our NLP models to estimates generated by manual labeling and time-to-treatment discontinuation (TTD).

**RESULTS** Our best high-dimensional logistic regression models identified toxicity events in early-stage patients with an area under the curve of the receiver-operator characteristic of 0.857 ± 0.014 (standard deviation) and progression events in metastatic patients with an area under the curve of 0.752 ± 0.027 (standard deviation). NLP-extracted toxicity incidence and PFS curves were not significantly different from manually extracted curves ($P = .95$ and $P = .67$, respectively). By contrast, TTD overestimated toxicity in early-stage patients ($P < .001$) and underestimated PFS in metastatic patients ($P < .001$). Additionally, we tested an extrapolation approach in which 20% of the metastatic cohort were labeled manually, and NLP algorithms were used to abstract the remaining 80%. This extrapolated outcomes approach resolved PFS differences between receptor subtypes ($P < .001$ for hormone receptor+/human epidermal growth factor receptor 2– *v* human epidermal growth factor receptor 2+ *v* triple-negative) that could not be resolved with TTD.

**CONCLUSION** NLP models are capable of abstracting treatment discontinuation rationale with minimal manual labeling.

**JCO Clin Cancer Inform 5:550-560. © 2021 by American Society of Clinical Oncology**



## INTRODUCTION

Electronic medical records (EMRs) have rapidly proliferated as a key element of clinical care.[1,2] By aggregating clinical records, imaging, laboratory values, and orders, EMR systems are attractive sources of large-scale real-world evidence (RWE) that could guide development of new biomarkers, therapies, and clinical decision-support tools. However, critical data points are often stored in unstructured formats (eg, free text notes), complicated by unique context-dependent medical terminology, acronyms, and idiosyncrasies such as copy-forwarding. As a result, manual chart abstraction is a time-consuming and expensive exercise. Clinical natural language processing (NLP) addresses this bottleneck via computational techniques that can automate chart abstraction at scale.

Oncology is a data-rich specialty that relies on multiple inputs for clinical decision making, including pathology, radiology, and molecular profiling. Accumulated data from routine clinical care therefore represent a valuable resource for new RWE-based discoveries. Deriving rigorous insights from these data requires access to clinical end points such as progression-free survival (PFS) and treatment-limiting toxicity that mirror end points from randomized controlled trials. Recent studies have supported time-to-treatment discontinuation (TTD) as a surrogate marker for PFS in observational studies.[3,4] However, critical conclusions about patient physiology or tumor biology could be obscured by failure to distinguish between progression and other reasons for discontinuation, including toxicity. In this context, NLP algorithms capable of differentiating treatment discontinuation

ASSOCIATED CONTENT

Data Sharing Statement
Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on April 12, 2021 and published at ascopubs.org/journal/cci on May 14, 2021: DOI https://doi.org/10.1200/CCI.20.00139

## CONTEXT

**Key Objective**

Can natural language processing (NLP) algorithms generate toxicity incidence and progression-free survival estimates by extracting treatment discontinuation rationale from unstructured electronic medical record notes?

**Knowledge Generated**

In a cohort of 6,115 patients with early-stage and 701 patients with metastatic breast cancer, we found that high-dimensional logistic regression and convolutional neural network NLP models can effectively identify clinical rationale for treatment discontinuation. Algorithmically extracted outcomes outperformed time-to-treatment discontinuation as a surrogate estimate for toxicity incidence and progression-free survival in early-stage and metastatic patients, respectively.

**Relevance**

Through algorithmic extraction of treatment discontinuation rationale, NLP could reduce manual labeling burden, which could enable novel biomarker discovery and real-world evidence-based decision support.



rationale could significantly improve the quality and scale of RWE-based studies.

Regarding progression, RECIST is the gold standard for assessing PFS as an end point in oncology clinical trials,[5,6] but routine clinical assessments rarely meet this standard. Algorithmic approaches have been developed to (1) mine radiology report text to characterize progression events, including longitudinal tumor measurements,[7] radiologists' impressions,[8-10] and specific tumor locations[11-13] and (2) extract oncologists' impressions of progression or response to therapy.[14] A study comparing abstraction of progression events from radiology versus medical oncology notes found that medical oncology notes enabled more efficient manual annotation with similar accuracy.[15,16]

Extracting treatment-limiting toxicity has received less attention than progression. Several initial efforts have focused on identifying adverse drug reactions.[17,18] However, automated methods have not previously been shown to reliably abstract real-world treatment-limiting toxicity from EMRs.

Recent NLP studies in oncology illustrate the emerging shift from rules-based to machine learning–based methods that automatically learn associations between phrases in text and the variables of interest.[19-22] High-dimensional logistic regression over bag-of-words text representations[23] is a lightweight but powerful method that can be competitive with deep learning on clinical extraction tasks.[24,25] However, one limitation of bag-of-words representations is that they are sparse and do not leverage relationships between words. By contrast, deep learning architectures, such as convolutional neural networks (CNNs), learn dense embeddings of words and also enable discovery of compositional structure; however, this often requires substantially more data.[26,27] CNNs have emerged as one of the most widely used architectures in clinical NLP[22,28] and are particularly highly represented in studies relevant to our work.[12-14] Comparison of classic machine-learning approaches (eg, logistic regression) and deep learning approaches is a common best practice in related studies.[9,11,12,25]

In this paper, we therefore sought to implement NLP algorithms to abstract the rationale for discontinuing treatment, focusing on toxicity in patients with early-stage breast cancer and progression in patients with metastatic breast cancer. The goal of this design is to demonstrate that the same underlying architectures can be successfully applied to multiple clinical questions of interest in different contexts. To the best of our knowledge, this is the first demonstration of automated abstraction of treatment discontinuation rationale from unstructured clinical EMR text.

## METHODS

### Patients

Our overall methodology is illustrated in Figure 1, and cohort characteristics are reported in Table 1. We constructed a retrospective cohort of 6,565 patients with breast cancer who presented to Memorial Sloan Kettering Cancer Center (MSK) between January 1, 2008, and January 1, 2019, and received all of their adjuvant or neoadjuvant (early-stage patients) or first-line therapy (metastatic patients) at MSK. Patients who were treated elsewhere for early-stage disease, experienced a recurrence, and established care at MSK for first-line treatment of metastatic cancer were included. Patients who transferred care to MSK for second-line or subsequent therapy were not included. All patient records were deidentified with respect to patient names and all dates. This research was reviewed by the MIT Committee on the Use of Humans as Experimental Subjects and was determined to be institutional review board–exempt.

We divided our patient population into two main cohorts consisting of 6,115 patients with early-stage and 701 patients with metastatic breast cancer. For patients who received an early-stage diagnosis and later recurred with metastatic disease, we split the patient's history at the date

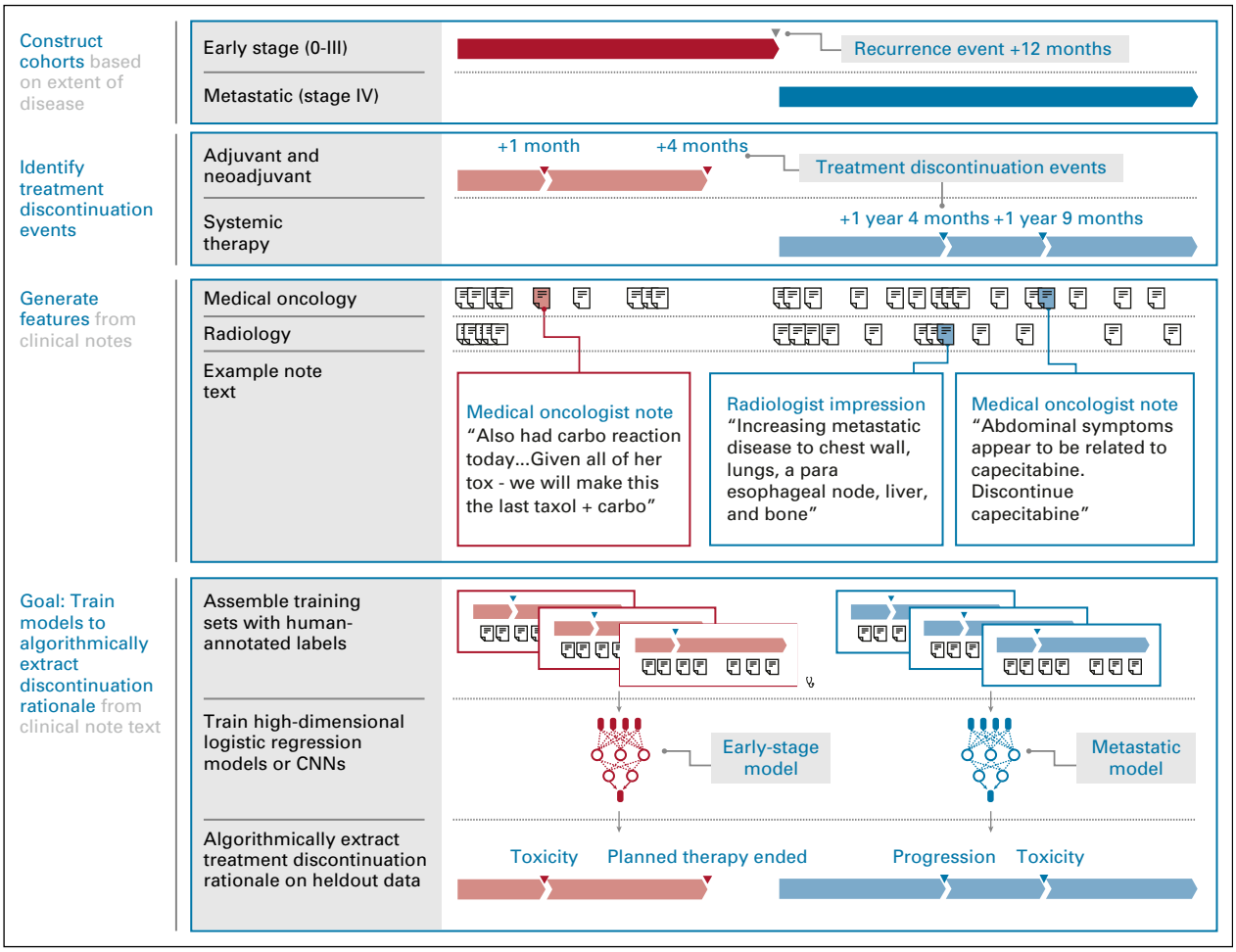


**FIG 1.** Overall study methodology. Treatment duration data were used to identify when a given therapy was discontinued. Feature sets were generated by concatenating unstructured clinical documents, including medical oncology and radiology notes. Examples of text are provided to illustrate the types of data that were interpreted by human abstractors to label the clinician's rationale for discontinuing therapy. Unstructured notes and manually annotated labels were then used to train high-dimensional logistic regression and CNN models to classify treatment discontinuation rationale in validation and test sets. CNN, convolutional neural network.

of metastatic recurrence and included the relevant segments of the patient's history in the early-stage and metastatic cohorts. Treatment discontinuation events were excluded from the analysis if the patient had no clinical notes within 30 days of the event.

### Label Construction and Task Definition

Treatment discontinuation rationale labels were extracted from EMR notes by nonclinician expert data extractors specialized in breast cancer clinical data collection. Progression and toxicity events were annotated primarily based on the assessment of the treating medical oncologist at the time of treatment discontinuation. Machine learning tasks were defined as distinguishing toxicity from other discontinuation reasons in patients with early-stage breast cancer and distinguishing progression from other discontinuation reasons in metastatic patients.

### Models

We then built models that could learn to automatically extract the discontinuation rationale labels identified by human annotators. Input features to both models were constructed by concatenating consecutive EMR notes surrounding the relevant treatment discontinuation event. We investigated (1) high-dimensional logistic regression models over unigram, bigram, and trigram frequencies and (2) one-dimensional CNN models. All training and analyses were performed separately on early-stage and metastatic cohorts. We independently subdivided metastatic and early-stage patients into training (70%), validation (15%), and held-out test (15%) subsets. These splits were performed at the patient level. Resulting data sets included all treatment discontinuation events associated with corresponding patient splits, including multiple discontinuation events per patient if present. This experimental setup

**TABLE 1.** Cohort Characteristics

| Characteristic | Early-Stage Patients (n = 6,115) | | | Metastatic Patients (n = 701) | | |
|---|---|---|---|---|---|---|
| | Train | Validate | Test | Train | Validate | Test |
| No. of patients | 4,280 | 917 | 918 | 490 | 105 | 106 |
| Age, mean (SD), years | 60.1 (12.6) | 59.5 (12.0) | 59.6 (12.4) | 61.7 (13.1) | 59.7 (13.2) | 64.9 (14.7) |
| Sex, % F | 99.5 | 99.2 | 99.5 | 98.5 | 100.0 | 99.0 |
| Years of follow-up, mean (SD) | 4.5 (3.2) | 4.5 (3.2) | 4.5 (3.1) | 3.1 (3.1) | 2.9 (2.4) | 3.2 (2.8) |
| Stage at first presentation | | | | | | |
| 0-I | 1,729 | 360 | 354 | 75 | 15 | 19 |
| II | 1,594 | 323 | 360 | 117 | 23 | 26 |
| III | 335 | 76 | 63 | 57 | 16 | 10 |
| IV | 0 | 0 | 0 | 212 | 48 | 45 |
| Unknown | 622 | 158 | 141 | 29 | 3 | 6 |
| Tumor markers | | | | | | |
| HR+/HER2– | 2,822 | 663 | 603 | 283 | 57 | 61 |
| HER2+ | 745 | 197 | 158 | 90 | 18 | 24 |
| Triple-negative | 663 | 141 | 143 | 89 | 26 | 14 |
| Unknown | 50 | 7 | 14 | 28 | 4 | 7 |
| Histology | | | | | | |
| Invasive ductal | 3,419 | 733 | 732 | 350 | 78 | 80 |
| Invasive lobular | 370 | 81 | 85 | 48 | 11 | 10 |
| Mixed | 215 | 50 | 50 | 28 | 5 | 8 |
| DCIS or LCIS | 133 | 23 | 26 | 12 | 1 | 2 |
| Inflammatory ductal | 30 | 8 | 3 | 27 | 5 | 2 |
| Rare subtype | 72 | 17 | 15 | 12 | 2 | 0 |
| Unknown | 41 | 5 | 7 | 13 | 3 | 4 |
| Treatment discontinuation events | | | | | | |
| Total | 8,107 | 1,788 | 1,758 | 1,568 | 324 | 341 |
| Planned therapy ended | 6,447 | 1,429 | 1,356 | 131 | 37 | 32 |
| Toxicity | 1,236 | 273 | 307 | 166 | 38 | 51 |
| Cancer progression | 199 | 39 | 46 | 1,178 | 231 | 229 |
| Patient or family preference | 176 | 35 | 38 | 21 | 6 | 7 |
| Other | 17 | 8 | 2 | 7 | 0 | 2 |
| Insurance issues | 13 | 0 | 6 | 2 | 2 | 2 |
| Transferred care | 10 | 2 | 1 | 5 | 0 | 5 |
| Patient died | 9 | 2 | 2 | 58 | 10 | 13 |

Abbreviations: DCIS, ductal carcinoma in situ; F, female; HER2, human epidermal growth factor receptor 2; HR, hormone receptor; LCIS, lobular carcinoma in situ; SD, standard deviation.



ensured that all discontinuation events used for validation and test originated from patients unseen during training. We tuned hyperparameters on a per-model basis to maximize the area under the curve (AUC) of the receiver-operating characteristic averaged across training cross-validation and the validation set. Key hyperparameters are described in the Data Supplement, including number of notes included in feature sets and handling of copy-forwarded text. For all models, final assessment was performed by assessing AUC on the held-out test set.

### Clinical Outcomes Estimation

To construct Kaplan-Meier curves from cohort-level clinical outcomes estimates, we identified the time between each patient's diagnosis (early-stage breast cancer) or metastatic disease diagnosis (metastatic cancer) and the next progression- or toxicity-related discontinuation event in the patient's history. Our NLP model estimates were compared with manually labeled data and to TTD estimates at a cohort level and for key subgroups, including hormone receptor–positive and human epidermal growth factor

receptor 2–negative (HR+/HER2–), HER2-positive (HER2+), and triple-negative breast cancer. Differences between Kaplan-Meier curves were assessed with the logrank test, implemented with the python package lifelines, and a $P$ value of $< .05$ was considered significant.[29]

### Extrapolated Outcomes Estimation in Metastatic Patients

To simulate a real-world implementation of our method, we sought to test whether a small labeling investment was sufficient to perform crucial outcome studies. Toward this goal, we designed an extrapolated outcomes approach in which 20% of the metastatic cohort was manually labeled and the remaining 80% of cases were algorithmically extracted and combined with the initial manual labels. This experiment was performed on a new random 20% train and 80% test split of the original underlying data.

## RESULTS

### Cohort Construction

Our initial cohort included 6,565 patients with breast cancer (Fig 2). Within this cohort, 6,115 patients had a history of early-stage breast cancer, with a mean standard deviation (SD) age of 59.9 (12.5) years, 4.5 (3.2) years of follow-up, and 99.5% were female; 701 had a history of metastatic disease, with a mean (SD) age of 61.9 (13.4) years, 3.1 (3.0) years of follow-up, and 98.8% were female (Table 1). Two hundred fifty-one patients first presented with early-stage disease and later experienced metastatic recurrence. These patients were included in both cohorts by splitting each patient's history at the date of metastatic recurrence. The early-stage cohort had 289,260 associated clinical documents and 11,653 associated discontinuation events, whereas the metastatic cohort had 73,103 associated clinical documents and 2,233 associated discontinuation events (Fig 2 and Table 1). In the early-stage cohort, 4,280 (70.0%), 917 (15.0%), and 918 (15.0%) patients were assigned to the train, validate, and test groups, accounting for 8,107, 1,788, and 1,758 discontinuation events, respectively (Table 1). In the metastatic cohort, 490 (69.9%), 105 (15.0%), and 106 (15.1%) patients were assigned to the train, validate, and test groups accounting for 1,568, 324, and 341 discontinuation events, respectively. Characteristics of these cohorts are shown in Table 1, including stage at first presentation, receptor subtype (HR+/HER2–, HER2+, and triple-negative), histology, and annotated treatment discontinuation rationale. In the early-stage breast cancer cohort, 12.2% of treatment discontinuation events were excluded because of lack of clinical notes within 30 days of the treatment discontinuation event. In the metastatic cohort, the proportion of excluded events was 3.6%.

### Population-Level Model Performance

In early-stage patients, the best high-dimensional logistic regression model demonstrated an AUC of 0.857 ± 0.014 with respect to toxicity. The corresponding CNN model demonstrated an AUC of 0.875 ± 0.018. In metastatic patients, our best high-dimensional logistic regression model achieved an AUC of 0.752 ± 0.027 on the held-out test set and the best CNN model demonstrated an AUC of 0.760 ± 0.031 (Data Supplement). Detailed AUC analyses and confusion matrices, including multiclass outputs, are shown in the Data Supplement. To assess the relative data volume required by each model, we scored each model with cross-validation AUC across a range of training data sizes. Compared to logistic regression models, CNN models demonstrated poorer performance at small training sizes relative to peak performance (Data Supplement). Model introspection was performed by extracting the highest-weighted terms from each high-dimensional logistic regression model (Data Supplement).

### NLP Models Effectively Estimate Clinical Outcomes in Overall Early-Stage and Metastatic Populations

In early-stage patients, the cumulative incidence of toxicity estimated by our best high-dimensional logistic regression model was not significantly different than the manually abstracted incidence of toxicity events ($P = .95$, Fig 3A). By contrast, TTD was largely driven by planned discontinuation of therapy and does not provide an adequate surrogate marker for treatment-limiting toxicity ($P < .001$ *v* manually abstracted toxicity incidence). In metastatic patients, we showed that PFS estimated by our NLP algorithm was indistinguishable from manually abstracted PFS ($P = .70$). By contrast, estimated PFS based solely on TTD was significantly different from manually abstracted PFS ($P < .001$, Fig 3B).

### NLP Models Effectively Estimate Outcomes in HR+/HER2−, HER2+, and Triple-Negative Subtypes

To further assess our estimation ability across key patient subsets, we segmented patients according to major receptor subtypes (HR+/HER2–, HER2+, and triple-negative). In patients with early-stage breast cancer, AUCs were maintained $> 0.8$ across receptor subtypes (Data Supplement) and our estimated cumulative toxicity incidence curves were similar to underlying rates in HR+/HER2–, HER2+, and triple-negative early-stage patients ($P > .05$ for all comparisons, Data Supplement). For metastatic patients, AUCs $> 0.7$ were maintained across receptor subtypes except for triple-negative patients (AUC 0.664 ± 0.038, Data Supplement). Our estimated PFS curves were not significantly different from manually abstracted PFS across all metastatic disease subtypes ($P > .05$ for all comparisons, Data SupplementS3). The Data Supplement shows AUCs and confusion matrices for these receptor subtype analyses.

### Extrapolated PFS Curves Can Distinguish Between Metastatic Subgroups

Next, to simulate a real-world implementation, we tested an extrapolated outcomes approach where human abstractors identified labels for a training subset (20%) and algorithms

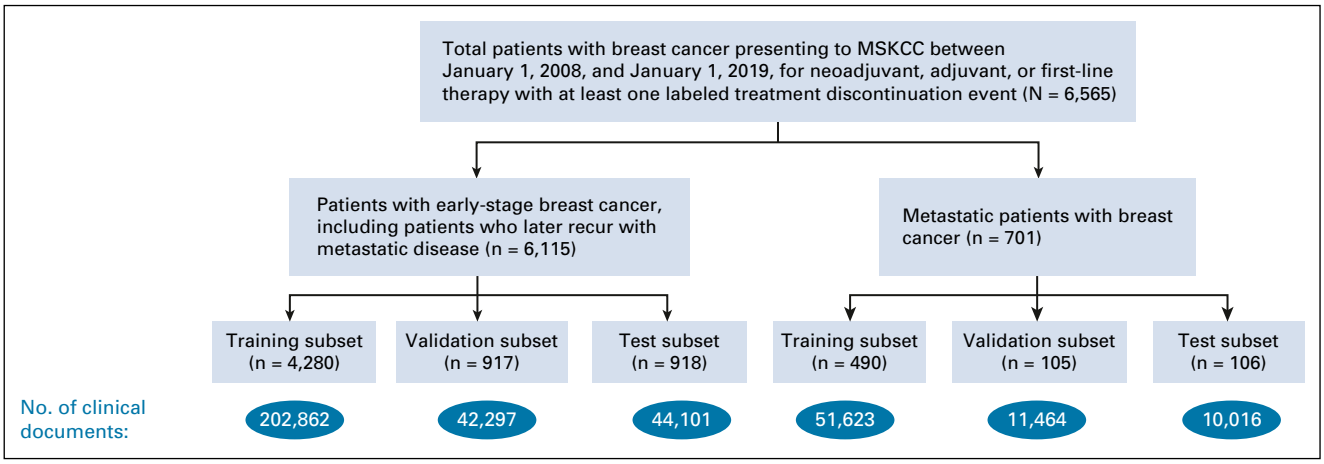


**FIG 2.** Cohort construction. The source cohort for this study was composed of patients with breast cancer who were first seen at MSKCC between the dates of January 1, 2008, and January 1, 2019, who also had EMR notes and documented time-to-treatment discontinuation data available. This cohort was split into early-stage and metastatic cohorts, including 251 early-stage patients who later recurred and were included in both subcohorts by splitting the clinical history at the recurrence date. EMR, electronic medical record; MSKCC, Memorial Sloan Kettering Cancer Center.



were used to extract outcomes from the remaining 80% of the data set (cohort characteristics for this experiment are available in the Data Supplement, and isolated NLP model performance is presented by subgroup in the Data Supplement). As a baseline, we evaluated full manual labeling of the cohort, which demonstrated significant differences among major tumor marker-defined subgroups ($P < .001$ for all comparisons, Fig 4A). We then constructed PFS curves using TTD as a surrogate marker and found that HR+/HER2− and HER2+ curves were not significantly different from one another ($P = .47$, Fig 4B). In isolation, PFS curves constructed from the 20% manually labeled training set showed no significant difference between triple-negative patients and HER2+ ($P = .08$) or HR+/HER2− ($P = .82$) because of large confidence intervals (Fig 4C). By contrast, PFS curves constructed from both the manual labels (20%) and algorithmically extracted labels (80%) showed that all groups were significantly different from one another ($P < .001$ for all comparisons Fig 4D), consistent with findings in the fully manually labeled baseline (Fig 4A). Taken together, these results indicate that the PFS estimates generated by our extrapolated outcomes approach were comparable to manually extracted data.

## DISCUSSION

In this study, we demonstrate that high-dimensional logistic regression NLP algorithms are capable of abstracting treatment discontinuation rationale from unstructured clinical oncology notes. Our models maintained their performance across HR+/HER2−, HER2+, and triple-negative breast cancer subtypes. We also demonstrated the utility of an extrapolated outcomes approach in which a focused manual labeling investment can be used to train NLP algorithms to extract outcomes in remaining cases. This approach enabled resolution of outcomes differences that could not be assessed from focused manual labels alone or by using TTD as a surrogate marker.

Our approach incorporates toxicity as an end point of interest to supplement prior efforts focused on progression or prognosis.[30,31] We highlight the importance of detecting toxicity in the adjuvant or neoadjuvant setting where large numbers of patients are exposed to potential adverse effects that must be optimized against the benefit of risk reduction. Other critical clinical questions require the ability to extract both toxicity and progression end points, such as assessing whether premature discontinuation of immune checkpoint inhibitors because of adverse events affects long-term outcomes.[32]

Prior studies have assessed the minimum expert labeling required for adequate performance for tasks such as annotation of breast pathology features.[33] In our study, we found that clinically meaningful outcomes assessment could be accomplished with a focused manual labeling effort of 20% of our data set combined with algorithmically abstracted outcomes for the remaining 80%. Consequently, our approach could significantly reduce the bottleneck for abstracting outcomes from large volumes of clinical text, accelerating discovery of novel biomarkers to identify subpopulations with different likelihood of response, duration of response, or toxicity.

Recently, deep learning methods have been shown to replicate or exceed human experts on a range of clinical tasks, including radiology-based diagnoses[34-37] or risk stratification,[38] assessing ophthalmology imaging,[39-41] classifying skin lesions,[42-44] and deriving diagnoses from pathology images.[45-47] Deep learning methods have also demonstrated promise in predicting clinical outcomes, such as prognosis or response to therapy. For example, pathology images were used to directly predict 5-year

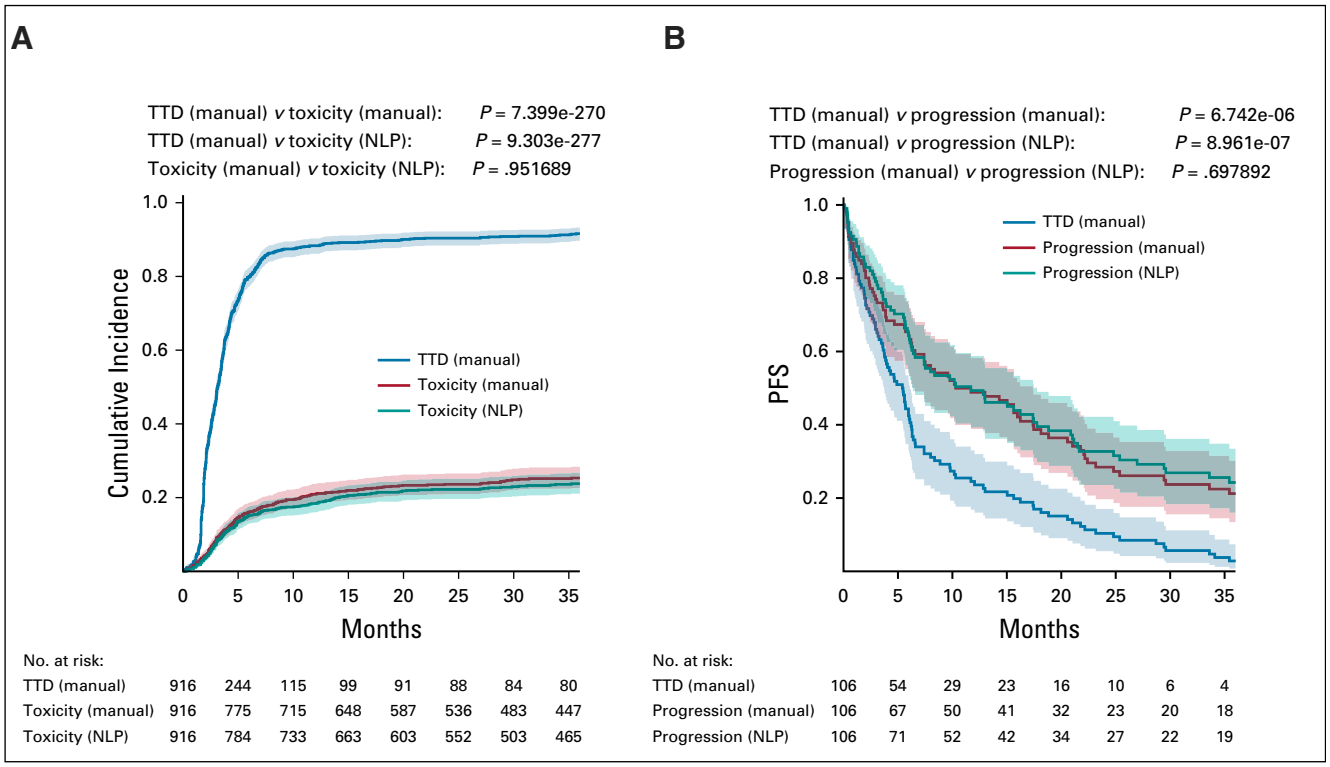


**FIG 3.** Estimated treatment-limiting toxicity in early-stage breast cancer and PFS in metastatic breast cancer compared with TTD as a surrogate estimate. (A) In early-stage patients, cumulative incidence curves were constructed based on toxicity events algorithmically extracted by a logistic regression NLP model. Estimated toxicity incidence curves (dark blue curve) were not significantly different from manually abstracted toxicity incidence (light blue). Gray curves indicate TTD for any reason, largely driven by planed discontinuation of therapy. For early-stage analyses, recurrent metastatic patients were censored at the time of metastatic diagnosis, such that any toxicity events attributable to therapy for distant metastases are not included. (B) In metastatic patients, Kaplan-Meier curves were constructed based on progression events algorithmically extracted by a logistic regression-based NLP model. Estimated PFS curves (dark blue curve) were not significantly different from manually abstracted PFS curves (light blue). PFS estimates based only on TTD (gray) inaccurately suggest faster accumulation of progression events than the manually abstracted underlying PFS curve. *P* values determined by logrank test. NLP, natural language processing; PFS, progression-free survival; TTD, time-to-treatment discontinuation.

survival in patients with colorectal cancer in one recent study[48] and non–small-cell lung cancer prognosis in another.[49] Furthermore, proteomic profiles,[50] genomic sequencing,[51] and molecular markers[52] can be used to enrich feature sets and further improve predictive performance.

Set against this context, our approach could help establish a new paradigm in which algorithmically extracted outcomes are subsequently used as labels to train deep learning models at a scale intractable for manually generated labels. As a result, this paradigm could accelerate predictive biomarker discovery for prognosis, therapeutic response, or risk of adverse events (Fig 5).

NLP-enabled outcomes abstraction could also aid clinical decision-support tools that draw on RWE to provide patient-tailored risk assessment or prognostic predictions. Such systems could be envisioned to update in real time as patients are treated in massively parallel natural experiments, given interprovider, interinstitutional, regional-, and national-scale variation in clinical practice. Our methodology could also serve to complement current efforts in applying machine learning to identifying eligible patients for clinical trial recruitment.[53-55]

Prior studies have demonstrated AUC $>$ 0.8 for identification of progression in metastatic patients[13-16] compared with our models' performance in the range of 0.75-0.76. This difference can be attributed to the increased difficulty of our task; our work separates progression events from other complications leading to treatment discontinuation, whereas other works contrast progression from a patient's entire history, a simpler discriminative task. Furthermore, the labels in this work were generated at a per-event level, and not at a per-note level. This approach mirrors the real-world scenario in which it is unknown a priori which note contains the rationale, requiring concatenation of multiple notes. With sufficient data, future exploration of other approaches such as recurrent neural networks may mitigate signal dilution across multiple notes.

Another limitation of our approach is the assumption that the timing of each treatment discontinuation event is

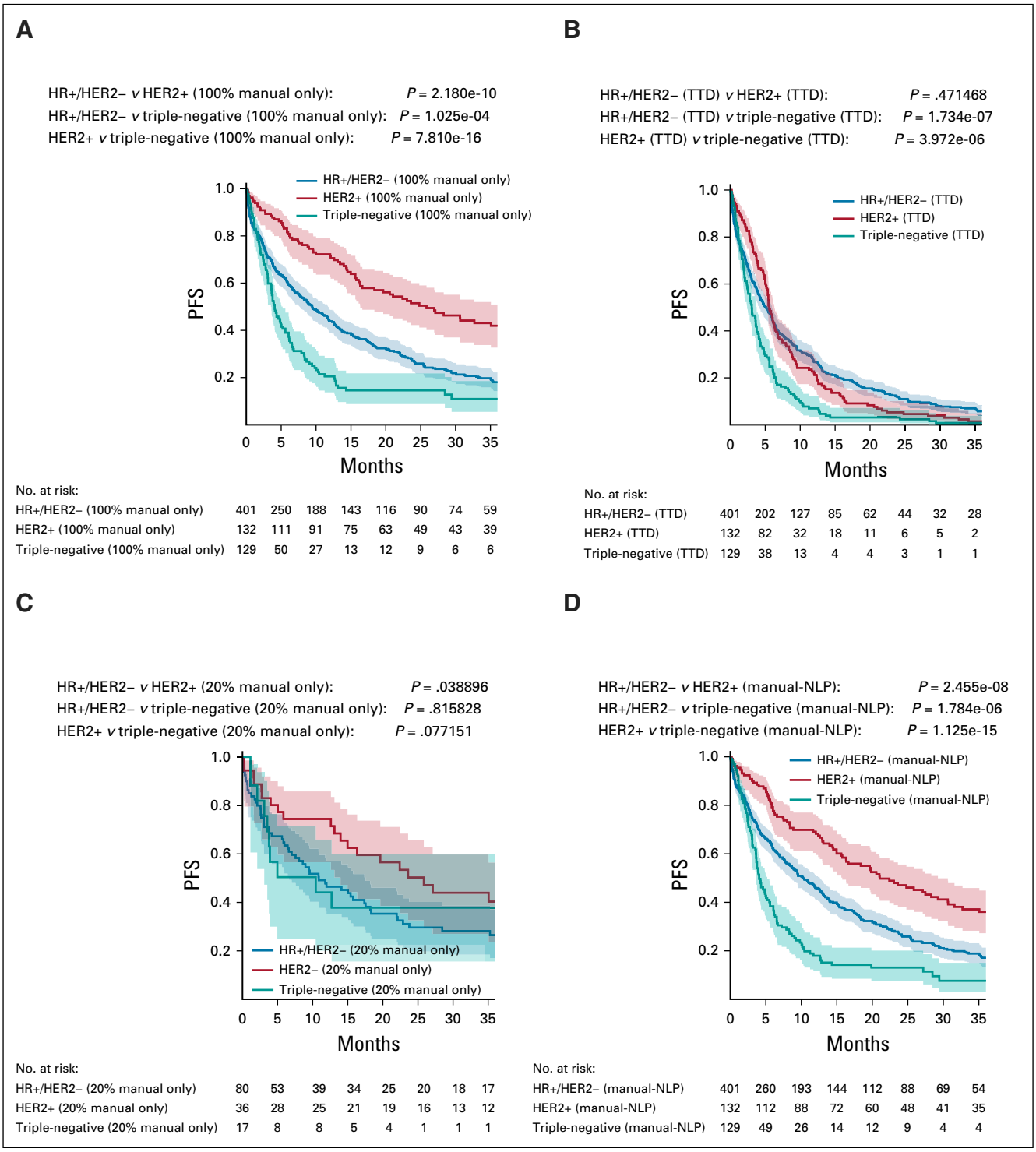


**FIG 4.** Comparison of PFS in major tumor marker-defined metastatic subgroups using an extrapolated outcomes approach of focused labeling plus automated abstraction. (A) Full manual labeling of the cohort demonstrated significant differences among major tumor marker-defined subgroups. (B) PFS curves constructed using TTD as a surrogate marker. (C) Focused labeling, defined as 20% of the total cohort was performed and used to construct PFS curves in HR+/HER2–, HER2+, and triple-negative subcohorts. (D) A logistic regression model was trained on the labels accounting for 20% of the cohort and tested on the remaining 80%. PFS curves were constructed from both the manual labels (20%) and algorithmically extracted labels (80%). *P* values determined by logrank test. HER2, human epidermal growth factor receptor 2; HR, hormone receptor; NLP, natural language processing; PFS, progression-free survival; TTD, time-to-treatment discontinuation.

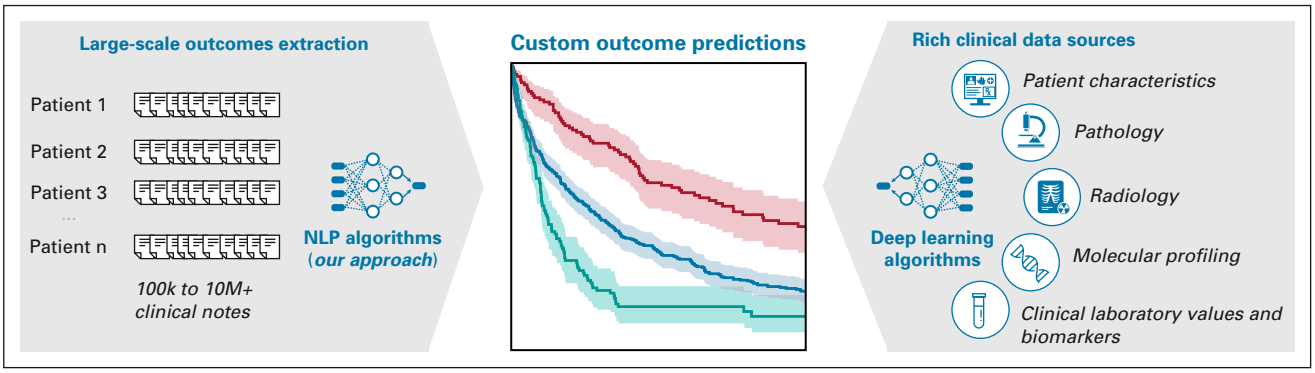


**FIG 5.** Proposed paradigm for achieving large-scale clinical machine learning algorithms predicting outcomes from rich clinical data sources. The approach described in this paper is focused on deriving outcomes for hundreds of thousands to millions of clinical documents. These abstracted outcomes could then be used to train deep learning algorithms trained to predict key outcomes from up-front data sources such as patient characteristics, pathology, radiology, molecular profiling, and clinical laboratory values and biomarkers. k, thousand; M, million; NLP, natural language processing.

known. In real-world settings, additional inference may be required to identify the occurrence of a discontinuation event before identifying the rationale. Our approach also benefits from manual abstraction of metastatic status for each discontinuation event. In undifferentiated data sets, additional methods would be required to extract metastatic status and identify metastatic recurrence.[56,57] Since we excluded treatment discontinuation events without a note within 30 days, future studies could attempt to incorporate non-note data to address events without tightly associated clinical notes.

Generalizability is critical to advance machine learning methods into clinical practice, but transfer learning across tumor type, for example, has proven challenging.[58-60] We have demonstrated that our overall methods are applicable to multiple clinically relevant questions; however, we note that our specific models include highly weighted terms that are breast cancer–specific, which would limit out-of-the-box generalizability to other tumor types. Further work would also be required to evaluate whether our approach can effectively transfer to other institutions with different practice patterns.

In conclusion, this work demonstrates that NLP is capable of automated abstraction of clinician rationale for treatment discontinuation events. Adopting this approach could lead to a paradigm shift in how large-scale outcomes are aggregated to support biomarker discovery and clinical decision support.


## AFFILIATIONS

[1]CSAIL & IMES, Massachusetts Institute of Technology, Cambridge, MA
[2]Harvard Medical School, Boston, MA
[3]Memorial Sloan Kettering Cancer Center, New York, NY
[4]Weill-Cornell Medical College, New York, NY

## CORRESPONDING AUTHOR

David Sontag, PhD, CSAIL & IMES, Massachusetts Institute of Technology, 45 Carleton St, Building E25, Cambridge, MA 02139; e-mail: dsontag@mit.edu.


## DISCLAIMER

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.


## SUPPORT

Supported by the NIH Oxford-Cambridge Scholars Program (M.S.A.), Award No. T32GM007753 from the National Institute of General Medical Sciences (M.S.A.), Memorial Sloan Kettering Cancer Center (M.N.A. and D.S.), and the NCI Cancer Center (Grant No. P30 CA08748; G.J.R. and P.R.).


## DATA SHARING STATEMENT

A data sharing statement provided by the authors is available with this article at DOI https://doi.org/10.1200/CCI.20.00139.

## AUTHOR CONTRIBUTIONS

**Conception and design:** Matthew S. Alkaitis, Monica N. Agrawal, Gregory J. Riely, David Sontag
**Financial support:** David Sontag
**Administrative support:** David Sontag
**Collection and assembly of data:** Gregory J. Riely, Pedram Razavi, David Sontag
**Data analysis and interpretation:** Matthew S. Alkaitis, Monica N. Agrawal, Pedram Razavi, David Sontag
**Manuscript writing:** All authors
**Final approval of manuscript:** All authors
**Accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by the authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's

conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

**Matthew S. Alkaitis**
**Employment:** Advanced Clinical, Beacon Biosignals
**Stock and Other Ownership Interests:** Beacon Biosignals, Moderna Therapeutics

**Monica N. Agrawal**
Uncompensated relationship: Flatiron Health

**Gregory J. Riely**
**Research Funding:** Novartis, Roche/Genentech, GlaxoSmithKline, Pfizer, Infinity Pharmaceuticals, Mirati Therapeutics, Merck, Takeda
**Patents, Royalties, Other Intellectual Property:** Patent application submitted covering pulsatile use of erlotinib to treat or prevent brain metastases
**Travel, Accommodations, Expenses:** Merck Sharp & Dohme
**Other Relationship:** Pfizer, Roche/Genentech, Takeda

**Pedram Razavi**
**Honoraria:** Epic Sciences, Inivata
**Consulting or Advisory Role:** Novartis, AstraZeneca, Foundation Medicine, Epic Sciences, Tempus, Natera, Inivata
**Research Funding:** Grail, Illumina, Novartis, Epic Sciences, Archer
**Travel, Accommodations, Expenses:** Epic Sciences, Guardant Health

**David Sontag**
**Employment:** ASAPP, Repertoire Immune Medicines (I)
**Stock and Other Ownership Interests:** Curai
**Consulting or Advisory Role:** Curai, GNS Healthcare
**Research Funding:** Takeda, Genentech

No other potential conflicts of interest were reported.

## ACKNOWLEDGMENT

The authors would like to thank Lior Gazit, Alex Grigorenko, and Iker Huerga for their help in data preparation and research advice.



## REFERENCES

1. Henry J, Pylypchuk Y, Searcy T, et al: Adoption of Electronic Health Record Systems Among U.S. Non-Federal Acute Care Hospitals: 2008-2015, Office of the National Coordinator for Health Information Technology, Washington DC. 2016 (ONC Data Brief No. 35)
2. Cohen MF: Impact of the HITECH financial incentives on EHR adoption in small, physician-owned practices. Int J Med Inform 94:143-154, 2016
3. Cazzaniga ME, Pronzato P, Meattini I, et al: Validation of time to treatment change (TTC) as a surrogate end-point of progression free survival (PFS) for observational trials in metastatic breast cancer patients (MBC): The GIM-13 AMBRA study. J Clin Oncol 36, 2018 (suppl; abstr e13081)
4. Blumenthal GM, Gong Y, Kehl K, et al: Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. Ann Oncol 30:830-838, 2019
5. Eisenhauer EA, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247, 2009
6. Schwartz LH, Litière S, de Vries E, et al: RECIST 1.1—Update and clarification: From the RECIST committee. Eur J Cancer 62:132-137, 2016
7. Sevenster M, Bozeman J, Cowhy A, et al: A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. J Biomed Inform 53:36-48, 2015
8. Cheng LTE, Zheng J, Savova GK, et al: Discerning tumor status from unstructured MRI reports—Completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 23:119-132, 2010
9. Chen P-H, Zafar H, Galperin-Aizenberg M, et al: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging 31:178-184, 2018
10. Arbour KC, Luu AT, Luo J, et al: Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. Cancer Discov 11:59-67, 2021
11. Gao S, Young MT, Qiu JX, et al: Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 25:321-330, 2018
12. Qiu JX, Yoon H-J, Fearn PA, et al: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform 22:244-251, 2018
13. Kehl KL, Elmarakeby H, Nishino M, et al: Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol 5:1421-1429, 2019
14. Kehl KL, Xu W, Lepisto E, et al: Natural language processing to ascertain cancer outcomes from medical oncologist notes. JCO Clin Cancer Inform:680-690, 2020
15. Griffith SD, Tucker M, Bowser B, et al: Generating real-world tumor burden endpoints from electronic health record data: Comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. Adv Ther 36:2122-2136, 2019
16. Griffith SD, Miksad RA, Calkins G, et al: Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non–small-cell lung cancer data set. JCO Clin Cancer Inform:1-13, 2019 doi:10.1200/CCI.19.00013
17. Taira RK, Soderland SG, Jakobovits RM: Automatic structuring of radiology free-text reports. Radiographics 21:237-245, 2001
18. Sarker A, Gonzalez G: Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 53:196-207, 2015
19. Nadkarni PM, Ohno-Machado L, Chapman WW: Natural Language processing: An introduction. J Am Med Inform Assoc 18:544-551, 2011
20. Yim W, Yetisgen M, Harris WP, et al: Natural language processing in oncology: A review. JAMA Oncol 2:797-804, 2016
21. Sheikhalishahi S, Miotto R, Dudley JT, et al: Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Med Inform 7:e12239, 2019
22. Wu S, Roberts K, Datta S, et al: Deep learning in clinical natural language processing: A methodical review. J Am Med Inform Assoc 27:457-470, 2020
23. Manning C, Schütze H: Foundations of Statistical Natural Language Processing. Cambridge, MA, MIT Press, 1999
24. Shao Y, Taylor S, Marshall N, et al: Clinical Text Classification With Word Embedding Features vs. Bag-of-Words Features. 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, pp 2874-2878
25. Oleynik M, Kugic A, Kasáč Z, et al: Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc 26:1247-1254, 2019

26. Kim Y: Convolutional Neural Networks for Sentence Classification [Internet]. arXiv:14085882 [cs], 2014. http://arxiv.org/abs/1408.5882
27. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521:436-444, 2015
28. Dernoncourt F, Lee JY, Uzuner O, et al: De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 24:596-606, 2017
29. Davidson-Pilon C: CamDavidsonPilon/lifelines [Internet], 2020. https://github.com/CamDavidsonPilon/lifelines
30. Yang Y, Fasching PA, Tresp V: Modeling Progression Free Survival in Wreast Cancer with Tensorized Recurrent Neural Networks and Accelerated Failure Time Models. Proceedings of the 2nd Machine Learning for Healthcare Conference, PMLR 68, pp 164-176, 2017
31. Gensheimer MF, Henry AS, Wood DJ, et al: Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. J Natl Cancer Inst 111:568-574, 2019
32. Shoushtari AN, Friedman CF, Navid-Azarbaijani P, et al: Measuring toxic effects and time to treatment failure for nivolumab plus ipilimumab in melanoma. JAMA Oncol 4:98-101, 2018
33. Yala A, Barzilay R, Salama L, et al: Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 161:203-211, 2017
34. Cheng J-Z, Ni D, Chou Y-H, et al: Computer-Aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 6:24454, 2016
35. Kooi T, Litjens G, van Ginneken B, et al: Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 35:303-312, 2017
36. Cicero M, Bilbily A, Colak E, et al: Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest Radiol 52:281-287, 2017
37. McKinney SM, Sieniek M, Godbole V, et al: International evaluation of an AI system for breast cancer screening. Nature 577:89-94, 2020
38. Castro SM, Tseytlin E, Medvedeva O, et al: Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 69:177-187, 2017
39. Gulshan V, Peng L, Coram M, et al: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316:2402, 2016
40. Ting DSW, Cheung CY-L, Lim G, et al: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 318:2211-2223, 2017
41. Fauw JD, Ledsam JR, Romera-Paredes B, et al: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24:1342-1350, 2018
42. Esteva A, Kuprel B, Novoa RA, et al: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115-118, 2017
43. Haenssle HA, Fink C, Schneiderbauer R, et al: Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29:1836-1842, 2018
44. Tschandl P, Codella N, Akay BN, et al: Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. Lancet Oncol 20:938-947, 2019
45. Bejnordi BE, Veta M, van Diest PJ, et al: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318:2199-2210, 2017
46. Nagpal K, Foote D, Liu Y, et al: Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digital Med 2:1-10, 2019
47. Cruz-Roa A, Gilmore H, Basavanhally A, et al: Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Sci Rep 7:46450, 2017
48. Bychkov D, Linder N, Turkki R, et al: Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 8:3395, 2018
49. Yu K-H, Zhang C, Berry GJ, et al: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 7:1-10, 2016
50. Yu K-H, Berry GJ, Rubin DL, et al: Association of omics features with histopathology patterns in lung adenocarcinoma. Cell Syst 5:620-627.e3, 2017
51. Loh P-R, Tucker G, Bulik-Sullivan BK, et al: Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47:284-290, 2015
52. Mobadersany P, Yousefi S, Amgad M, et al: Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci USA 115:E2970-E2979, 2018
53. Bustos A, Pertusa A: Learning eligibility in cancer clinical trials using deep neural networks. Appl Sci 8:1206, 2018
54. Shivade C, Hebert C, Regan K, et al: Automatic data source identification for clinical trial eligibility criteria resolution. AMIA Annu Symp Proc 2016:1149-1158, 2017
55. Zhang K, Demner-Fushman D: Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. J Am Med Inform Assoc 24:781-787, 2017
56. Carrell DS, Halgrim S, Tran D-T, et al: Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. Am J Epidemiol 179:749-758, 2014
57. Banerjee I, Bozkurt S, Caswell-Jin JL, et al: Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. JCO Clin Cancer Inform 3:1-12, 2019
58. D'Avolio LW, Nguyen TM, Farwell WR, et al: Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc 17:375-382, 2010
59. Bethard S, Savova G, Palmer M, et al: SemEval-2017 Task 12. Clinical TempEval, 2017
60. Santus E, Li C, Yala A, et al: Do neural information extraction algorithms generalize across institutions? JCO Clin Cancer Inform 3:1-8, 2019