Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports

Kenneth L. Kehl, MD, MPH; Haitham Elmarakeby, PhD; Mizuki Nishino, MD, MPH; Eliezer M. Van Allen, MD; Eva M. Lepisto, MA, MSc; Michael J. Hassett, MD, MPH; Bruce E. Johnson, MD; Deborah Schrag, MD, MPH

IMPORTANCE A rapid learning health care system for oncology will require scalable methods for extracting clinical end points from electronic health records (EHRs). Outside of clinical trials, end points such as cancer progression and response are not routinely encoded into structured data.

OBJECTIVE To determine whether deep natural language processing can extract relevant cancer outcomes from radiologic reports, a ubiquitous but unstructured EHR data source.

DESIGN, SETTING, AND PARTICIPANTS A retrospective cohort study evaluated 1112 patients who underwent tumor genotyping for a diagnosis of lung cancer and participated in the Dana-Farber Cancer Institute PROFILE study from June 26, 2013, to July 2, 2018.

EXPOSURES Patients were divided into curation and reserve sets. Human abstractors applied a structured framework to radiologic reports for the curation set to ascertain the presence of cancer and changes in cancer status over time (ie, worsening/progressing vs improving/responding). Deep learning models were then trained to capture these outcomes from report text and subsequently evaluated in a 10% held-out test subset of curation patients. Cox proportional hazards regression models compared human and machine curations of disease-free survival, progression-free survival, and time to improvement/response in the curation set, and measured associations between report classification and overall survival in the curation and reserve sets.

MAIN OUTCOMES AND MEASURES The primary outcome was area under the receiver operating characteristic curve (AUC) for deep learning models; secondary outcomes were time to improvement/response, disease-free survival, progression-free survival, and overall survival.

RESULTS A total of 2406 patients were included (mean [SD] age, 66.5 [10.8] years; 1428 female [59.7%]; 2170 [90.2%] white). Radiologic reports (n = 14 230) were manually reviewed for 1112 patients in the curation set. In the test subset (n = 109), deep learning models identified the presence of cancer, improvement/response, and worsening/progression with accurate discrimination (AUC >0.90). Machine and human curation yielded similar measurements of disease-free survival (hazard ratio [HR] for machine vs human curation, 1.18; 95% CI, 0.71-1.95); progression-free survival (HR, 1.11; 95% CI, 0.71-1.71); and time to improvement/response (HR, 1.03; 95% CI, 0.65-1.64). Among 15 000 additional reports for 1294 reserve set patients, algorithm-detected cancer worsening/progression was associated with decreased overall survival (HR for mortality, 4.04; 95% CI, 2.78-5.85), and improvement/response was associated with increased overall survival (HR, 0.41; 95% CI, 0.22-0.77).

CONCLUSIONS AND RELEVANCE Deep natural language processing appears to speed curation of relevant cancer outcomes and facilitate rapid learning from EHR data.

Published online July 25, 2019.

© 2019 American Medical Association. All rights reserved.
The US health care system rapidly incorporated electronic health records (EHRs) into routine clinical practice in the past decade. The large volume of information housed within EHRs can be viewed as evidence with the potential to drive cancer research and optimize oncologic care delivery. However, important clinical endpoints, such as response to therapy and disease progression, are often recorded in the EHR only as unstructured text. For patients in therapeutic clinical trials, standards such as the Response Evaluation Criteria in Solid Tumors (RECIST) provide clear guidelines for defining radiographic disease status. However, RECIST is not routinely applied outside of clinical trials. To create a true learning health system for oncology and facilitate delivery of precision medicine at scale, methods are needed to accelerate curation of cancer-related outcomes from EHRs according to validated standards.

We have developed a structured framework for curation of clinical outcomes among patients with solid tumors using medical records data (PRISSMM). PRISSMM considers each data source—including pathologic reports, radiologic/imaging reports, signs/symptoms, molecular markers, and medical oncologist assessment—independently. By maintaining a record of such data provenance through the curation process, a multidimensional representation of a patient’s clinical trajectory and outcomes, including disease-free survival (DFS) and progression-free survival (PFS), can be constructed.

Even within a structured framework, human medical records curation is resource intensive and often impractical. To accelerate the process of curating relevant clinical outcomes for patients with cancer, we applied deep learning to EHR curated by abstractors according to PRISSMM. We focused on text reports of imaging studies, which are ubiquitous and central to ascertainment of cancer progression and treatment response for patients with advanced solid tumors. Our hypothesis was that deep learning algorithms could use routinely generated radiologic text reports to identify the presence of any cancer, and, if cancer was present, identify shifts in disease burden and involved sites.

**Methods**

**Cohort**

The data for this analysis were derived from the EHRs of patients with lung cancer who had genomic profiling performed through the Dana-Farber Cancer Institute (DFCI) PROFILE precision medicine effort from June 26, 2013, to July 2, 2018. PROFILE participants consented to medical records review and genomic profiling of their tumor tissue. PROFILE was approved by the DFCI Institutional Review Board; this supplemental retrospective analysis was declared exempt from review also by the DFCI Institutional Review Board. Patients were included regardless of cancer stage at diagnosis, histologic findings (ie, small cell vs non–small cell lung cancer), or treatment history. Information on self-reported race/ethnicity is reported to inform assessment of generalizability.

The cohort was restricted to PROFILE patients with lung cancer who had at least 1 report of a computed tomographic, magnetic resonance imaging, positron emission tomographic-computed tomographic, or nuclear medicine bone scan in our EHR. Manual curation for the cohort according to PRISSMM is ongoing; we conducted this analysis after records for approximately one-half of the PROFILE patients with lung cancer had been curated. Patients who had already undergone manual medical record review at the time of this analysis constituted the curation set. For deep learning model training, patients were further divided randomly, at the patient level, into training (80%), validation (10%), and test (10%) subsets (Figure 1); the number of patients in each curation subset approximated the proportions, because splitting was originally performed among all PROFILE patients. Patients who had not already undergone manual medical record review constituted the reserve set, which was used to conduct independent indirect validation of deep natural language processing models.

**Medical Record Curation**

Eight curators were trained to abstract data for the curation set according to the PRISSMM framework; details regarding curator training and interrater reliability are provided in eTable 1A.
in the Supplement.9 The present analysis focused on imaging reports. For each imaging study performed on or after the first pathologic report documenting lung cancer, curators evaluated whether each radiologic report indicated the presence of cancer; if cancer was present, whether the cancer was improving/responding, stable/no change, mixed, progressing/worsening/enlarging, or not stated/indeterminate in comparison with the previous scan; and the extent of cancer, focusing on common metastatic sites, including liver, adrenal gland, bone, lymph nodes, and brain/spine14 (eTable 1B in the Supplement). Data were collected using a REDCap database.15

Statistical Analysis
Within the curation set, deep natural language processing models were trained to determine whether imaging reports would be interpreted by human curators as demonstrating any evidence of cancer; improving/responding disease; worsening/progressive disease; and cancer present in specific extrapulmonary metastatic sites, including liver, bone, brain and/or spine, lymph nodes, and adrenal glands. Each of these 8 end points was treated as a binary outcome for a distinct model (eTable 1B in the Supplement). The model architecture treated each imaging report as a sequence of tokens, embedded into a vector space,16 and then fed into a 1-dimensional convolutional layer (eFigure 1 in the Supplement) to capture local interactions among nearby words.17 The dimensionality of each convolutional layer was then reduced using max-pooling.18 Pooling layer output was fed into fully connected layers to model global interactions among input words; this higher-level model architecture was designed to reflect intrinsic interdependence among outcomes. For example, the presence of cancer is required for response, progression, or disease at specific sites to be present, so the hidden layer for estimation of any cancer was concatenated to a specific hidden layer for each other outcome (eFigure 1B in the Supplement). Dropout was applied to hidden layers to improve generalizability.19 The final hidden layer output was fed into a sigmoid layer to generate a score that can be interpreted as the probability that a given imaging report was in a class of interest (eg, representing any evidence of cancer).

Model stability was assessed, and the optimal threshold probability for a positive outcome was defined by applying cross-validation to models trained within the 80% curation training subset; each of 5 cross-validation models was trained with a random 80% sample of the training data (ie, 64% of total curation data) and evaluated using the remaining 20% of the training data (ie, 16% of total curation data). The sigmoid activation layer output score for defining a predicted outcome as positive was defined as half of the mean best F1 scores calculated by cross-validation, where the F1 score is defined as the harmonic mean between precision and recall.20 For each outcome, an ensemble of the cross-validation models was constructed by taking the simple mean of the sigmoid activation layer outputs from each cross-validation model. The ensemble models were applied to the 10% validation subset for evaluation and iterative tuning. At the end of the analysis, the models were applied to the 10% final held-out test data set, after which no further training was performed. Using each imaging report as the unit of analysis and considering test subset human-curated outcomes to represent true labels, model performance was evaluated using the area under the receiver operating characteristic curve.21 The area under the precision-recall curve, which measures the tradeoff between precision (positive predictive value) and recall (sensitivity) as well as the F1 score at the best F1 threshold, were also calculated. The local-interpretable model-agnostic explanation framework22 was applied to generate human-interpretable explanations of each classification.

Models were constructed using the Keras framework23 with the TensorFlow backend.24 Our free-text imaging reports constitute protected health information and are not publicly available, but the code used to train and evaluate the models is available on GitHub.25

To ascertain the face validity (clinical relevance) of machine-curated outcomes, we compared human and machine measurements of DFS, PFS, and time to improvement/response for curation set patients. Subgroups were defined as those who had early-stage disease at diagnosis, for evaluation of DFS, or received palliative-intent systemic therapy, for evaluation of PFS and time to improvement/response. Among patients with early-stage disease, we defined DFS as the time from initial diagnosis (regardless of upfront therapy) until death (ascertained via an institutional linkage to the National Death Index26 through December 31, 2017, and institutional administrative data thereafter) or the first imaging report indicating worsening/progression; patients were censored at the date of medical record abstraction or 180 days after their last imaging report, whichever came first. Among patients who received palliative-intent systemic therapy, we defined PFS as the time from initiation of first palliative-intent therapy until death or an imaging report indicating cancer worsening/progression, and time to improvement/response as the time from initiation of first palliative-intent systemic therapy to the first imaging report indicating improvement/response, subject to competing risks of death; patients were censored at the date of medical record abstraction or 90 days after their last imaging report, whichever came first. Time to these events using human vs machine curation was analyzed using Cox proportional hazards regression models, stratified by patient, with analysis of the cause-specific hazard in the competing risks case of time to improvement/response.

Next, we measured associations between cancer outcomes derived from machine curation of radiologic and overall survival. Because these analyses did not rely on manually curated medical record data, they were conducted among both curation set and reserve set patients. The intermediate outcomes serving as predictors in these models included the presence of any cancer, worsening/progression, and improvement/response. Analyses of cancer response or progression were restricted to imaging reports following initiation of a palliative-intent treatment plan. In addition, we used the curation set to measure the association between human abstractions of these end points in the curation set and overall survival. These associations were quantified using Cox proportional hazards regression models in which the curated intermediate outcomes on the most recent imaging report represented time-varying
covariates with respect to the hazard of mortality; patients were censored on the data of data download or 90 days after each scan, whichever came first. Because genomic testing was required for inclusion in the cohort, a sensitivity analysis was performed in which only imaging reports that followed the genomic test order date were analyzed. Nominal, 2-sided P-values <.05 were considered statistically significant.

Results

Cohort

A total of 2406 patients were included (mean [SD] age, 66.5 [10.8] years; 1428 female [59.7%]; 2170 [90.2%] white). Radiologic reports (n = 14230) were manually reviewed for 1112 patients in the curation set who underwent genomic sequencing from June 26, 2013, through July 2, 2018. These patients had 14230 radiologic reports from our institution and affiliated sites available for analysis. A total of 884 patients (79.5%) were assigned to the training subset, 119 patients (10.7%) to the validation subset, and 109 patients (9.8%) to the final test subset for machine learning. The mean (SD) number of imaging reports per curated patient was 12.8 (12.0), over a mean of 26.0 (22.7) months. The reserve set, for whom manual curation has not yet been performed, included an additional 1294 patients, with 15 000 imaging reports. Patient and radiologic report characteristics in the curated and reserve sets were similar. For example, 71% to 74% of patients in the curation subsets and 75% of patients in the reserve set had adenocarcinoma (Table 1; eTable 2 in the Supplement).

Deep Learning to Classify Imaging Reports

Deep learning model performance characteristics are provided in Table 2. Within the test subset of the curation set, the AUC for detection of any cancer was 0.92; for worsening/progression, 0.94; for improvement/response, 0.95; for liver lesions, 0.98; for bone lesions, 0.95; for brain/spine lesions, 0.97; for lymph node lesions, 0.89; and for adrenal lesions, 0.97. Graphic depictions of the receiver operating characteristic curve and the precision-recall curve are provided in eFigure 2 in the Supplement. Sample automated model explanations for high-probability individual predictions are provided in eFigure 3 in the Supplement. For example, words such as mass, tumor, and burden tended to indicate that cancer was present; words such as increasing and metastatic tended to indicate progressive disease; and words such as decrease tended to indicate improving disease in specific reports.

Face Validity of Machine-Curated End Points

Among patients with early-stage lung cancer in the curation set, machine and human curations of the time from diagnosis to disease recurrence or death (DFS) yielded similar

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Curation Set (n = 1112 Patients)</th>
<th>Validation (n = 119)</th>
<th>Test (n = 109)</th>
<th>Reserve Set (n = 1294 Patients)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Histologic diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>658 (74.4)</td>
<td>84 (70.6)</td>
<td>78 (71.6)</td>
<td>968 (74.8)</td>
<td></td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>113 (12.8)</td>
<td>17 (14.3)</td>
<td>15 (13.8)</td>
<td>136 (10.5)</td>
<td></td>
</tr>
<tr>
<td>Small cell lung cancer</td>
<td>45 (5.1)</td>
<td>3 (2.5)</td>
<td>5 (4.6)</td>
<td>37 (2.9)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>68 (7.7)</td>
<td>15 (12.6)</td>
<td>11 (10.1)</td>
<td>153 (11.8)</td>
<td></td>
</tr>
<tr>
<td><strong>Disease extent at original diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early (stage I-III)</td>
<td>544 (61.5)</td>
<td>78 (65.5)</td>
<td>70 (64.2)</td>
<td>NA^</td>
<td></td>
</tr>
<tr>
<td>Metastatic (stage IV)</td>
<td>340 (38.5)</td>
<td>41 (35.4)</td>
<td>39 (35.8)</td>
<td>NA^</td>
<td></td>
</tr>
<tr>
<td><strong>Age at genomic sequencing, y</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;50</td>
<td>55 (6.2)</td>
<td>8 (6.7)</td>
<td>8 (7.3)</td>
<td>87 (6.7)</td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>189 (21.4)</td>
<td>27 (22.7)</td>
<td>22 (20.2)</td>
<td>244 (18.9)</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>301 (34.0)</td>
<td>42 (35.3)</td>
<td>39 (35.8)</td>
<td>442 (34.2)</td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td>250 (28.3)</td>
<td>34 (28.6)</td>
<td>31 (28.4)</td>
<td>390 (30.1)</td>
<td></td>
</tr>
<tr>
<td>≥80</td>
<td>89 (10.1)</td>
<td>8 (6.7)</td>
<td>9 (8.3)</td>
<td>131 (10.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>359 (40.6)</td>
<td>51 (42.9)</td>
<td>47 (43.1)</td>
<td>521 (40.3)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>525 (59.4)</td>
<td>68 (57.1)</td>
<td>62 (56.9)</td>
<td>773 (59.7)</td>
<td></td>
</tr>
<tr>
<td><strong>Self-reported race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>795 (89.9)</td>
<td>106 (89.1)</td>
<td>94 (86.2)</td>
<td>1175 (90.8)</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>28 (3.2)</td>
<td>4 (3.4)</td>
<td>6 (5.5)</td>
<td>48 (3.7)</td>
<td></td>
</tr>
<tr>
<td>Black/African American</td>
<td>30 (3.4)</td>
<td>1 (0.8)</td>
<td>6 (5.5)</td>
<td>39 (3.0)</td>
<td></td>
</tr>
<tr>
<td>Unknown/declined/other</td>
<td>31 (3.5)</td>
<td>8 (6.7)</td>
<td>3 (2.8)</td>
<td>32 (2.5)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: NA, not available. * Staging information was not available for reserve set patients, whose records have not been manually reviewed. The other available clinical and demographic data for the reserve set were collected through the Dana-Farber Cancer Institute precision medicine effort (PROFILE).
results (hazard ratio [HR] in the test subset for machine vs human curation, 1.18; 95% CI, 0.71-1.95). Among patients who received palliative-intent systemic therapy, machine and human curations in the time from initiation of first palliative therapy to disease progression or death (PFS; HR in the test subset, 1.11; 95% CI, 0.71-1.71) or from initiation of first palliative therapy to improvement/response (HR in the test subset, 1.03; 95% CI, 0.65-1.64) also yielded similar results (Figure 2).

Within the separate reserve set of 1294 patients, whose records have not been manually curated, the HR for mortality associated with a report indicating active cancer vs no active cancer was 6.77 (95% CI, 4.89-9.38). Among reserve set patients with radiologic reports for studies following initiation of palliative-intent systemic therapy, reports indicating worsening/progression per machine curation were associated with an increased risk of mortality (HR, 4.04; 95% CI, 2.78-5.85) vs reports indicating no worsening/progression, and reports indicating cancer improvement/response per machine curation were associated with a decreased risk of mortality (HR, 0.41; 95% CI, 0.22-0.77) vs reports indicating no improvement/response (Figure 2). Results were similar in a sensitivity analysis restricted to imaging reports following a genomic testing order date (HR for any cancer, 6.15; 95% CI, 4.16-9.09; HR for worsening/progression, 3.43; 95% CI, 2.32-5.08; HR for improvement/response, 0.40; 95% CI, 0.20-0.79).

### Discussion

These results suggest that deep natural language processing, trained using the structured PRISMM framework for curation of real-world cancer outcomes, can rapidly extract clinically relevant outcomes from the text of radiologic reports. Our human curators can annotate imaging reports for approximately 3 patients per hour; at that rate, it would take 1 curator approximately 6 months to annotate all imaging reports for the patients with lung cancer in our cohort. Once trained, the models that we developed could annotate imaging reports for the entire lung cancer cohort in approximately 10 minutes. By reducing the time and expense necessary to review medical records, this technique could substantially accelerate efforts to use real-world data from all patients with cancer to generate...
Figure 2. Comparison of Human and Machine Curation of Imaging Reports

No. at risk

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>51</td>
<td>41</td>
</tr>
<tr>
<td>B</td>
<td>41</td>
<td>34</td>
</tr>
<tr>
<td>C</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>D</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>E</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Shaded bands represent 95% CIs. The cumulative rate of response was calculated subject to competing risks of death. Disease-free survival was measured among patients with early-stage disease at diagnosis. The end point was defined as the time from diagnosis until death or the first imaging report curated as representing disease worsening/progression, whichever came first. Progression-free survival was measured among patients who received systemic therapy documented as palliative in intent. The end point was defined as the time from initiation of first palliative-intent systemic therapy to the date of first imaging report indicating disease worsening/progression, whichever came first. Response rate was measured among patients who received systemic therapy documented as palliative in intent. The end point was defined as the time from initiation of first palliative-intent systemic therapy at the Dana-Farber Cancer Institute until the first imaging report indicating improving/responding disease, subject to competing risks of death.

Figure 3. Association Between Machine-Curated Imaging Report Outcomes and Overall Survival

<table>
<thead>
<tr>
<th>Curation Type</th>
<th>HR for Mortality (95% CI)</th>
<th>Favors Presence of Finding</th>
<th>Favors Absence of Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine; reserve set</td>
<td>6.77 (4.89-9.38)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Machine; curation set</td>
<td>7.50 (5.60-10.03)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Human; curation set</td>
<td>6.48 (4.95-8.49)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

All results were derived from Cox proportional hazards regression models in which the imaging finding was included as a time-varying covariate. Hazard ratios therefore represent the relative hazard of mortality at any given time if a particular imaging finding was present on the most recent prior scan vs if it was not. In models analyzing the outcome of any cancer, the presence of cancer on an imaging report was the sole independent variable. In models analyzing the outcome of response and progression, each was included as a binary variable.

A Cohort included 1294 reserve set patients (with imaging reports on 11,946 dates predating survival cutoff) who did not undergo human curation but for whom overall survival data were available. Survival time was indexed at the date of the first imaging report available.

B Cohort included 1112 curation set patients (with imaging reports on 11,091 dates predating survival cutoff) who did not undergo human curation but for whom overall survival data were available. Survival time was indexed at the date of the first imaging report available.

C Cohort included 395 reserve set patients (with imaging reports on 4,092 dates predating survival cutoff) who did not undergo human curation but for whom overall survival data were available and initiated a treatment plan recorded as palliative in intent. Survival time was indexed at the date of first palliative-intent systemic therapy.

D Cohort included 501 curation set patients (with imaging reports on 4,941 dates predating survival cutoff) who underwent human and machine curation and initiated a treatment plan recorded as palliative in intent. Survival time was indexed at the date of first palliative-intent systemic therapy.

Evidence regarding effectiveness of treatment approaches and guide decision support.

For example, rapid ascertainment of cancer outcomes using machine learning could specifically improve the utility of large-scale genomic data for precision medicine efforts. In trials of broad next-generation sequencing, most participants in such studies have not enrolled in genotype-matched therapeutic trials. However, oncology clinical trials often close without reaching their accrual goals. Methods are needed to
identify patients who are eligible for trials at the time when they have progressive disease and are in need of a new therapy. Scalable techniques for extracting outcome data from EHRs could be incorporated into decision support tools for matching patients to targeted therapies at appropriate times in their disease trajectories.

**Strengths and Limitations**

Strengths of this study included the application of deep natural language processing to the challenge of extracting clinically relevant end points from EHRs using techniques that have previously been applied more often to sentiment analysis problems in nonmedical fields. Prior efforts to apply natural language processing techniques to measure oncologic outcomes have not leveraged recent rapid innovation in deep learning. Deep learning confers specific advantages, including facilitating subsequent transfer learning in which components of previously trained models can be repurposed to perform related tasks using fewer data than might otherwise be required. We also incorporated a method for explaining individual deep learning model predictions into our analysis. This method evaluates the effect on local model predictions of slight alterations to input data for individual observations, enabling the text most important for a specific prediction, which may vary among observations, to be highlighted. This model explanation framework addresses a common criticism of deep learning models, particularly in health care—that nonlinear, black box models fail to yield the interpretability necessary for clinicians and researchers to trust model predictions. In addition, the labeled data for the analysis were generated using the open PRISMM framework, which relies on publicly available, tumor site-agnostic data standards to improve reproducibility. This technique enabled the derivation of structured outcome data without requiring that radiologists generate information directly by clicking additional buttons in the EHR, which can be a tedious, error-prone process.

This study has limitations, including the single-institution design, conducted among patients with one type of cancer. Models trained using EHR data from a single tertiary care institution may not perform as well in other institutions or practice settings. However, if multiple institutions and practices were to share their unstructured data for additional model training, or even implement distributed training without requiring that raw protected health information be shared, generalizability could be evaluated and optimized. The present study also ascertained cancer outcomes using only imaging reports. Some clinically important end points may be evaluable only within other components of the EHR, including laboratory test results, physical examination, symptoms, and oncologists’ impressions. Further research is needed to extract reproducible end points based on these additional EHR data sources. In addition, real-world clinical reports often communicate true clinical uncertainty about a finding at any given time. This underlying uncertainty likely explains the somewhat lower performance of our models for certain end points, such as the presence or absence of cancer in lymph nodes. Further work is needed to explicitly incorporate measures of uncertainty into model predictions.

Our real-world end points specifically do not attempt to replicate the RECIST criteria traditionally used to evaluate responses to treatment in clinical trials. Using the PRISMM framework, disease response and progression do not require attainment of specific relative changes in tumor diameter, as would be the case for RECIST measurements. Raw response rates as defined by PRISMM therefore cannot be directly compared with objective response rates defined by RECIST criteria. However, in our study, PRISMM response and progression were associated with expected shifts in mortality risk along the disease trajectory.

Deep learning techniques for image analysis are advancing rapidly; an alternative approach to this work could be to train models to replicate human labels using radiologic images rather than radiologist-generated text interpretations of those images. However, training such a model to accurately discern malignant disease and identify changes in that disease might require substantially more data and computing resources. Analyzing the radiologic reports effectively leverages the clinical expertise of radiologists to reduce the dimensionality of raw images into a text report for analysis, constituting a form of human to machine transfer learning.

**Conclusions**

Automated collection of clinically relevant, real-world cancer outcomes from unstructured EHRs appears to be feasible. This technique has the potential to augment capacity for learning from the large population of patients with cancer who receive care outside the clinical trial context. Key next steps will include testing this approach in other health care systems and clinical contexts and applying it to evaluate associations among tumor profiles, therapeutic exposures, and oncologic outcomes.

**ARTICLE INFORMATION**

Accepted for Publication: April 3, 2019.
Published Online: July 25, 2019.

**Author Contributions:** Dr Kehl had full access to all the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis. Concept and design: Kehl, Elmarakeby, Van Allen, Lepisto, Hassett, Johnson, Schrag. Acquisition, analysis, or interpretation of data: Kehl, Nishino, Van Allen, Lepisto, Hassett, Johnson, Schrag. Drafting of the manuscript: Kehl, Van Allen, Schrag. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: Kehl, Elmarakeby, Van Allen. Obtained funding: Schrag. Administrative, technical, or material support: Lepisto, Hassett, Johnson, Schrag. Supervision: Van Allen, Hassett, Johnson, Schrag.

**Conflict of Interest Disclosures:** Dr Kehl reports serving as a consultant to Aetion Inc. Dr Nishino reports serving as a consultant to Toshiba Medical Systems, WorldCare Clinical, and Daiichi Sankyo; holding research grants from the Merck Investigator Studies Program, Toshiba Medical Systems, and AstraZeneca; and receiving honoraria from Bayer and Roche. Dr Van Allen reports serving as a consultant/advisor to Tango Therapeutics, Invitae, Genome Medical, Dynamo, Foresite Capital, and Illumina; holding research support from Novartis and Bristol-Myers Squibb; and holding...
equity in Sayse, Genome Medical, Tango, and Microsoft Corp. Dr Johnson reports receiving postmarketing royalties for epidermal growth factor receptor testing, paid by Dana-Farber Cancer Institute, and receiving research support from Novartis and Toshiba. Dr Schrag reports serving as a consultant/advisor to Cureus and Pfizer, receiving research funding from the American Association of Cancer Research, serving as an editor and receiving personal fees from JAMA outside the related work, holding equity in mixed biotech funds including Invesco QQQ and holding a trademark and having financial interest in the PRISMmM phenomic data standard. The other authors have no conflicts of interest to disclose.

**Funding/Support:** The study was supported by the National Cancer Institute (K08CA169384 [Dr Schrag], RO1CA203636, [Dr Nishino], and U01CA209414 [Dr Nishino]); the Claude Daupehlin Philanthropic Research Fund (Ors Johnson and Schrag); and the Simeon J. Fortin Foundation.

**Role of the Funder/Sponsor:** The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Additional Contributions:** The Dana-Farber Cancer Institute Oncology Data Retrieval System (OncDRS) was used for the aggregation, management, and delivery of the clinical and operational research data used in this project (Orechia J, Pathak A, Shi Y, et al. OncDRS: an integrative clinical and genomic data platform for enabling translational research and precision medicine. Appl Transl Genom. 2015;6:e18-25). The content is solely the responsibility of the authors.

**Additional Information:** Study data were collected and managed using REDCap electronic data capture tools hosted at Partners Healthcare. REDCap (Research Electronic Data Capture) is a secure, web-based application designed to support data capture for research studies, providing (1) an intuitive interface for validated data entry, (2) audit trails for tracking data manipulation and export procedures, (3) automated export procedures for seamless data downloads to common statistical platforms, and (4) procedures for importing data from external sources. Electronic health records were curated using the PRISMmM, version 2.0 data standard for defining clinical cancer end points from unstructured text. The PRISMmM 2.0 data standard, training modules, and curation directives are freely available to academic and governmental entities under the terms of a Creative Commons license. They will be made available to commercial entities at cost. Questions and requests for a PRISMmM license and data curation tools should be sent to PRISMmM@dfci.harvard.edu.

**REFERENCES**


**Mach Learn Knowl Discov Databases.** 2014;8725:224-239. doi:10.1007/978-3-662-44851-9_15
The Tipping Point for Deep Learning in Oncology

Andrew Daniel Trister, MD, PhD

**Approaches to artificial intelligence** have rapidly transformed many elements in our lives. Many have claimed that a new revolution fueled by artificial intelligence will bring even greater benefits and risks. Most of the common approaches leverage deep neural networks—algorithmic structures that process data through multiple layers of mathematical operations to result in a prediction or classification. These deep learning techniques require large amounts of well-annotated data and access to fast computing resources to ensure reasonable performance.

There are many promising applications of deep networks that are well positioned to affect diagnosis, research, and clinical decision-making. Two limitations of deep learning techniques to drive this paradigm shift in clinical care relate to interpretability of the neural network prediction and generalizability of the results from a network to populations that were not included in the original training. To address these limitations, there are emerging best practices regarding assembling appropriate data sets; how these data are handled between training, testing, and validation; and methods to provide results that are directly understandable to humans (explainability).

In an article in *JAMA Oncology,* we see an example of the power of deep natural language processing to predict clinically relevant oncologic end points from radiologic reports that addresses some of these potential limitations of deep learning approaches. Kehl et al\(^2\) leveraged an existing set of electronic health record data collected for a precision medicine initiative at a single institution, Dana-Farber Cancer Institute, over the course of 5 years. The authors previously recognized the importance of a robust annotation framework of data collected in a clinical setting to inform precision medicine initiatives. They have described a novel structured framework for curation of clinical outcomes among patients with solid tumors using medical records data, PRISSMM, that ensures provenance and annotation from all data sources within the health system across pathology, radiology, signs and symptoms, molecular markers, and medical oncologist assessments toward clinical outcome assessment. They highlight that there are high upfront costs in this type of manual curation and that the implementation of algorithms to annotated records showed significant reduction in the time needed to process large numbers of records.

In their study, Kehl et al\(^2\) restricted the training and testing of their algorithm to radiologic reports obtained from patients with a diagnosis of primary lung cancer. Reports were included regardless of histologic characteristics, stage, history of treatment, or modality of imaging included in the electronic health record. They took appropriate steps to separate the data into training, validation, and testing sets, and did so by randomly assigning patients and not individual reports into each group—an important step to ensure that information about a patient is not shared between phases of neural network development.

Despite reporting significant area under the receiver-operating characteristic curves for classifying a report along important clinical end points (existence of cancer, improvement, progression, and specific areas of concern for metastasis), the authors also incorporated local-interpretable model-agnostic explanation, which is a method that explores small changes to input data and the output of the model, giving some level of interpretability of the model. The local-interpretable model-agnostic explanation algorithm highlights key words within specific reports that the model used to determine classifications. By highlighting elements of the data, the local-interpretable model-agnostic explanation provides the end user an opportunity to begin to evaluate how the algorithm made a specific determination. This additional layer of data provides a check against the black-box nature of the algorithm and should be standard in solutions for clinical decision support.

One limitation the authors did not mitigate in their approach is the generalizability of the model to other patients with lung cancers treated in settings outside Dana-Farber Cancer Institute. Although they used self-reported race to inform the discussion around how their model could be used in other settings, geographic and socioeconomic factors not included in their evaluation might show that the retrospective data included in their study are not representative of the mix of patients seen at other centers. Furthermore, despite Dana-