ARTIFICIAL INTELLIGENCE

original reports abstract

Clinical Inflection Point Detection on the Basis of EHR Data to Identify Clinical Trial–Ready Patients With Cancer

Kenneth L. Kehl, MD, MPH[1]; Stefan Groha, PhD[1]; Eva M. Lepisto, MA, MSc[1]; Haitham Elmarakeby, PhD[1]; James Lindsay, PhD[1]; Alexander Gusev, PhD[1]; Eliezer M. Van Allen, MD[1]; Michael J. Hassett, MD, MPH[1]; and Deborah Schrag, MD, MPH[1]

PURPOSE To inform precision oncology, methods are needed to use electronic health records (EHRs) to identify patients with cancer who are experiencing clinical inflection points, consistent with worsening prognosis or a high propensity to change treatment, at specific time points. Such patients might benefit from real-time screening for clinical trials.

METHODS Using serial unstructured imaging reports for patients with solid tumors or lymphoma participating in a single-institution precision medicine study, we trained a deep neural network natural language processing (NLP) model to dynamically predict patients' prognoses and propensity to start new palliative-intent systemic therapy within 30 days. Model performance was evaluated using Harrell's c-index (for prognosis) and the area under the receiver operating characteristic curve (AUC; for new treatment and new clinical trial enrollment). Associations between model outputs and manual annotations of cancer progression were also evaluated using the AUC.

RESULTS A deep NLP model was trained and evaluated using 302,688 imaging reports for 16,780 patients. In a held-out test set of 34,770 reports for 1,952 additional patients, the model predicted survival with a c-index of 0.76 and initiation of new treatment with an AUC of 0.77. Model-generated prognostic scores were associated with annotation of cancer progression on the basis of manual EHR review (n = 1,488 reports for 110 patients with lung or colorectal cancer) with an AUC of 0.78, and predictions of new treatment were associated with annotation of cancer progression on the basis of manual EHR review with an AUC of 0.84.

CONCLUSION Training a deep NLP model to identify clinical inflection points among patients with cancer is feasible. This approach could identify patients who may benefit from real-time targeted clinical trial screening interventions at health system scale.

JCO Clin Cancer Inform 5:622-630. © 2021 by American Society of Clinical Oncology

INTRODUCTION

There is increasing interest in leveraging electronic health records (EHRs) at scale to generate real-world evidence to optimize treatment for patients with cancer.[1] In particular, scalable methods for ascertaining outcomes from the EHR could inform tailored care delivery.[2] This could be particularly critical for driving precision oncology[3] at health system scale. A key precision oncology task is to test novel therapies for many individual genomic alterations, each of which may be uncommon. To date, prospective studies of precision oncology strategies have yielded relatively low rates of enrollment in clinical trials, despite molecular matches between trials and tumor characteristics.[4,5] One barrier to enrollment in genomically matched trials is that patients may not be ready for a clinical trial at any given time. Clinical trials are generally available and relevant to patients when they are experiencing clinical inflection points—moments of progressive disease, worsening prognosis, or a high propensity to change therapy.

There is no structured EHR data field encoding these components of clinical inflection points. This presents a barrier to deployment of clinical decision support tools to deliver real-time information about clinical trials to patients and providers. One approach to clinical inflection point detection could be to develop a manually ascertained metric of cancer progression. Trained human curators or clinicians could then review medical records or clinical images and generate structured outcomes according to this cancer progression metric. These annotations could be used to train natural language processing (NLP) models to recapitulate and automate this process.[6-8] However, simple supervised machine learning methods for training these models require thousands of manually curated gold standard records, the collection of which may be challenging or even prohibitive. This could

ASSOCIATED CONTENT

Data Supplement

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

Accepted on April 21, 2021 and published at ascopubs.org/journal/cci on June 7, 2021: DOI https://doi.org/10.1200/CCI.20.00184

CONTEXT

Key Objective

Historically, a small fraction of adults with cancer has been enrolled in therapeutic clinical trials. Tools are in development to match patients to clinical trials on the basis of their clinical and tumor molecular profiles. However, a major challenge in deploying these tools has been that patients are generally only ready for clinical trials when they experience clinical inflection points corresponding to worsening disease or a high propensity to change therapy. We developed a machine learning natural language processing model to identify such inflection points in real time on the basis of serial imaging reports for each patient.

Knowledge Generated

Our model was able to discriminate between better and worse prognoses, and high and low propensity to change treatment, within a large pan-cancer single-institution precision medicine study.

Relevance

Deployment of our model to identify clinical inflection points at health system scale for patients with cancer could inform targeted delivery of information about relevant clinical trials to patients and clinicians.

similarly limit the training of NLP models to ascertain cancer progression per the RECIST,[8,9] since such training requires labeled data, and RECIST is not routinely applied to label patients' records outside of therapeutic clinical trials. Furthermore, even if large quantities of labeled data were available, cancer progression, per se, may be insufficient for identifying clinical trial–ready patients. For example, clear cancer progression may sometimes represent just slight clinical worsening, insufficient for individual providers to recommend a treatment change.

Clinical text exists in the context of other EHR data elements, which may be highly relevant to identifying clinical inflection points that correlate with worsening cancer. In particular, structured overall survival and treatment data may be available to researchers working with retrospective EHR data sets. If a clinical inflection point is defined as a time when patients have a poor or worsening prognosis and/or a high propensity to initiate a new treatment, a machine learning model could be trained to identify such moments using these existing structured data elements as labels. This model could then be applied in real time to clinical text for inflection point detection among future patients who have not yet made a treatment change. Such patients, and their oncologists, could receive targeted information about clinical trials or other cancer care delivery interventions for which they may be eligible at a specific time. The objective of this analysis was to train and evaluate such a model.

METHODS

Overview

A neural network NLP model was trained to dynamically predict a patient's prognosis and propensity to start new treatment, using the text from reports generated by radiologists reviewing imaging studies, including plain films, computed tomography, magnetic resonance imaging, and nuclear medicine scans. This algorithm bases each prediction on both a patient's prior radiology reports and new information introduced with each report to identify times at which a patient has poor or worsening prognosis and/or high propensity to initiate a new treatment.

Cohort

Because the primary motivation of this analysis was to improve rates of enrollment to genomically matched oncology trials, the data for the study were derived from the EHRs of patients with any malignant solid tumor or lymphoma, who had genomic profiling performed through the Dana-Farber Cancer Institute (DFCI) PROFILE[10,11] precision medicine effort from 2012 to 2019. 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. For model training, patients were divided randomly, at the patient level, into training (80%), tuning or validation (10%), and test (10%) subsets.[12,13]

Dynamic Prognostic Model Training

A neural network model was trained and applied on a per-patient basis. Each patient's data set included all imaging reports performed at DFCI and affiliated sites, sorted in chronologic order. Details of preprocessing, algorithm architecture, and training are provided in the Data Supplement. Training was performed using Pytorch.[14] Model code can be obtained from Github[15]; the underlying data constitute protected health information and are not publicly available. The model generated two output vectors per patient representing (1) a vector of predicted instantaneous log hazard of mortality per year following each report and (2) a vector of predicted log odds of initiating new palliative-intent anticancer systemic treatment within 30 days of each report.

Dynamic Prognostic Model Inference and Evaluation

This approach was evaluated by measuring the performance of the two novel model outputs: (1) log hazard of mortality per year and (2) log odds of new treatment at 30 days, calculated using the current model, and (3) probability of manually annotated progression, calculated among patients with lung cancer using a previously published algorithm,[6] at predicting three outcomes: (1) overall survival, (2) new treatment at 30 days, and (3) new clinical trial enrollment at 30 days. Evaluation was performed for imaging reports before June 30, 2018, to ensure follow-up for outcome ascertainment. Inference was performed by leaving dropout on, which renders output nondeterministic, generating predictions for each patient 10 times. The mean prediction for each of the two outputs was then calculated together with its standard deviation to yield a measure of model uncertainty for each report.[16]

Initial model performance was evaluated in the validation set by calculating Harrell's c-index[17] separately for each output in predicting overall survival and by calculating the area under the receiver operating characteristic curve (AUC)[18] for each log odds of initiating new treatment within 30 days output in predicting outcomes (2)-(3). Hyperparameters, including learning rate, dropout rate, network depth, and network width, were then manually adjusted on the basis of these metrics. These statistics and their 95% CIs were calculated using a modified bootstrapping approach, in which one report for each patient was sampled randomly and the statistic wsa calculated; this procedure was then repeated 500 times. The reported statistic was calculated as the mean of these samples; the lower bound of a percentile CI was calculated as the 2.5th percentile of the samples, and the upper bound was calculated as the 97.5th percentile of the samples.

After all training was complete, model performance was evaluated for the test set in the same manner. Performance was evaluated both among all patients in the test set and separately among patients with lung, breast, prostate, and colorectal cancer. To facilitate visual evaluation of performance, the distribution of predicted log hazards and predicted new treatments was also calculated and plotted for patients who died within 6 months of an imaging report and for those who did not, and for those who started new treatment within 30 days of a report versus those who did not.

Evaluation of Clinical Inflection Point Detection: Measurement of the Association Between Model Outputs and Manually Curated Cancer Progression

A subset of cohort of patients who had lung or colorectal cancer was identified, and their imaging reports were manually annotated in an REDCap database[19] according to the PRISSMM framework.[7] The PRISSMM abstraction process has been described previously.[6] Briefly, for each imaging report following a patient's pathologic diagnosis, abstractors reviewed the report for a description of active cancer. If cancer was present, abstractors recorded whether it was improving or responding, progressing or worsening, both (mixed), neither (stable) nor indeterminate.

For the current analysis, a binary variable was derived from these manual annotations, representing whether the radiology report contained a description of progressing or worsening disease, versus any other category. Next, a predicted log hazard of mortality per year and a predicted

TABLE 1. Patient Characteristics

Characteristics	Patients, No. (%)	Reports, No. (%)
All	20,916 (100)	377,221 (100)
Cancer primary site		
Lung	3,238 (15)	68,927 (18)
Colorectal	2,095 (10)	31,615 (8)
Breast	2,054 (10)	44,697 (12)
Pancreatic	758 (4)	11,568 (3)
Urothelial	664 (3)	11,919 (3)
Prostate	602 (3)	9,327 (2)
Others	11,505 (55)	199,168 (53)
Age, years		
< 40	2,173 (10)	32,945 (9)
40-49	2,451 (12)	46,231 (12)
50-59	5,196 (25)	95,683 (25)
60-69	6,271 (30)	114,586 (30)
70-79	3,768 (18)	69,199 (18)
80+	1,057 (5)	18,577 (5)
Sex		
Male	9,273 (44)	150,948 (40)
Female	11,643 (56)	226,273 (60)
Race		
Asian	609 (3)	11,506 (3)
Black or African American	661 (3)	13,658 (4)
Native American	22 (0)	343 (0)
Pacific Islander	7 (0)	99 (0)
White	18,887 (90)	339,498 (90)
More than one race	51 (0)	590 (0)
Others or unknown	679 (3)	11,527 (3)
Year of first genomic sequencing report		
2012	592 (3)	15,117 (4)
2013	2,628 (13)	58,090 (15)
2014	2,779 (13)	58,441 (15)
2015	4,609 (22)	88,147 (23)
2016	4,837 (23)	81,858 (22)
2017	3,071 (15)	44,511 (12)
2018	1,936 (9)	24,222 (6)
2019	464 (2)	6,835 (2)

TABLE 2. Utility of Model Outputs for Predicting Survival, New Treatment, and Clinical Trial Enrollment (Held-Out Test Set)

			Outcome: Overall Survival, C-index (95% CI)		Outcome: New Treatment Within 30 Days, AUC (95% CI)		Outcome: New Clinical Trial Within 30 Days, AUC (95% CI)	
Cohort	No. of Patients	No. of Reports	Predictor: Log Hazard of Mortality per Year	Predictor: Log Odds of New Treatment in the Next 30 Days	Predictor: Log Hazard of Mortality per Year	Predictor: Log Odds of New Treatment in the Next 30 Days	Predictor: Log Hazard of Mortality per Year	Predictor: Log Odds of New Treatment in the Next 30 Days
All patients	1,952	34,770	0.76 (0.74 to 0.77)	0.75 (0.74 to 0.76)	0.64 (0.61 to 0.68)	0.77 (0.74 to 0.81)	0.61 (0.55 to 0.68)	0.78 (0.70 to 0.86)
Cancer type								
Lung	320	6,211	0.76 (0.73 to 0.78)	0.77 (0.74 to 0.80)	0.73 (0.64 to 0.81)	0.81 (0.71 to 0.89)	0.73 (0.52 to 0.97)	0.83 (0.60 to 0.99)
Colorectal	175	2,288	0.73 (0.69 to 0.77)	0.75 (0.71 to 0.79)	0.53 (0.43 to 0.65)	0.71 (0.40 to 0.86)	0.60 (0.39 to 0.97)	0.90 (0.66 to 1.0)
Breast	190	4,343	0.79 (0.76 to 0.82)	0.78 (0.74 to 0.81)	0.61 (0.54 to 0.69)	0.75 (0.63 to 0.85)	0.57 (0.48 to 0.69)	0.72 (0.45 to 0.89)
Pancreatic	52	950	0.74 (0.68 to 0.79)	0.72 (0.64 to 0.79)	0.61 (0.47 to 0.82)	0.71 (0.47 to 0.93)	0.56 (0.51 to 0.63)	0.92 (0.78 to 0.98)
Prostate	62	905	0.78 (0.68 to 0.86)	0.81 (0.73 to 0.88)	0.68 (0.48 to 0.85)	0.81 (0.52 to 1.0)	0.68 (0.48 to 0.87)	0.78 (0.52 to 0.98)
Urothelial	68	1,115	0.76 (0.68 to 0.82)	0.78 (0.70 to 0.85)	0.64 (0.47 to 0.81)	0.79 (0.57 to 0.96)	0.65 (0.49 to 0.94)	0.80 (0.54 to 1.0)
Other solid tumors or lymphomas	1,077	18,879	0.75 (0.74 to 0.77)	0.74 (0.72 to 0.75)	0.65 (0.60 to 0.70)	0.78 (0.72 to 0.83)	0.63 (0.53 to 0.72)	0.78 (0.67 to 0.88)
Under-represented subgroups								
Non-White patients	161	2,018	0.78 (0.74 to 0.81)	0.75 (0.71 to 0.79)	0.67 (0.55 to 0.77)	0.76 (0.65 to 0.87)	0.67 (0.50 to 0.98)	0.78 (0.51 to 0.98)
Black patients	56	991	0.79 (0.73 to 0.84)	0.77 (0.72 to 0.83)	0.71 (0.45 to 0.95)	0.76 (0.54 to 0.93)	0.61 (0.51 to 0.87)	0.66 (0.51 to 0.95)

NOTE. Test set, restricted to imaging reports at least 6 months before the cohort censoring date of December 31, 2018.

Abbreviations: AUC, area under the receiver operating characteristic curve; C-index, Harrell's concordance index.

log odds of initiating new treatment within 30 days were assigned to each report by applying the trained neural network model to the report text. Finally, associations between each output and manually annotated cancer progression were calculated using the AUC in the same bootstrapped manner described above.

Model Interpretability

To generate simple human-interpretable approximations of how our neural network generated predictions globally across the data set, we fit linear regression models with LASSO regularization to predict the outputs of the neural network at each time point within the validation set. This approach was chosen because available local interpretability methods[20,21] are not readily applicable to a model architecture that treats each patient as a sequence of observations or reports. In these explanatory models, input text was converted into vectors consisting of sequences 1-2 words, which were encoded using term frequency-inverse document frequency vectorization.

RESULTS

Patients

Our cohort inclusion criteria identified 20,916 patients with 434 distinct types of cancer; these patients had a total of 377,221 imaging reports for analysis. There was a median of 12 reports per patient (interquartile range, 4-26). Additional cohort characteristics are provided in Table 1.

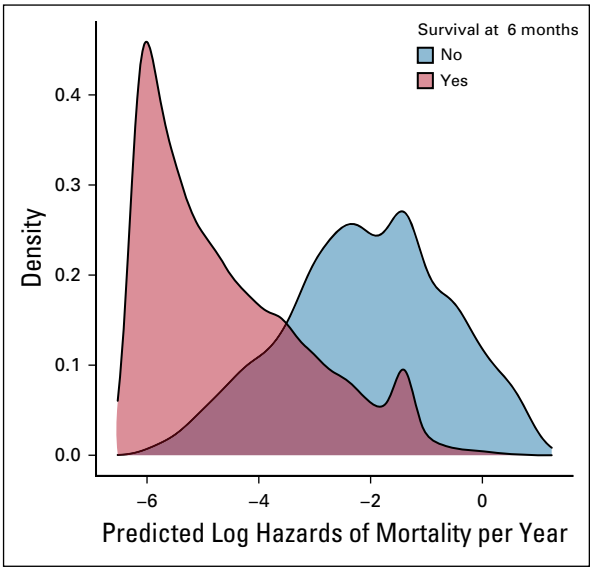

FIG 1. Distribution of predicted log hazard of mortality by actual subsequent 6-month mortality (n = 34,770 imaging reports for 1,952 patients). Histogram of model output 1 (predicted log hazard of mortality per year), by actual 6-month survival, following each imaging report, demonstrating that patients with worse predicted prognoses at any given time point have a lower rate of 6-month survival following that time point. Test set, restricted to imaging reports at least 6 months before the cohort censoring date of December 31, 2018.

Model Performance

Prognostication. Within the test set, the estimated log hazards were generated by the model predicted overall survival with a c-index of 0.76 (95% CI, 0.74 to 0.77). Performance among patients with specific types of cancer is provided in Table 2. In a sensitivity analysis restricted to imaging reports generated after tumor genomic profiling, which represented a cohort eligibility criterion, the c-index was 0.82 (95% CI, 0.81 to 0.82). In another sensitivity analysis in which the model was re-trained using only reports before the end of 2017 and evaluated on test set reports from 2018, to evaluate the ability of this technique to yield models useful for future patients, the results were similar (Data Supplement). The distribution of predictions by actual 6-month survival following each report is illustrated in Figure 1. The second model output variable, corresponding to predicted log odds of new treatment within 30 days, was also itself prognostic, with a c-index of 0.75 (95% CI, 0.74 to 0.76; Table 2). For comparison, among a subset of patients with lung cancer (n = 30,377 reports for 2,276 patients), the output of a prior algorithm trained to recapitulate manual annotations of cancer progression[6] yielded a c-index of 0.67 (95% CI, 0.65 to 0.68) for survival prediction.

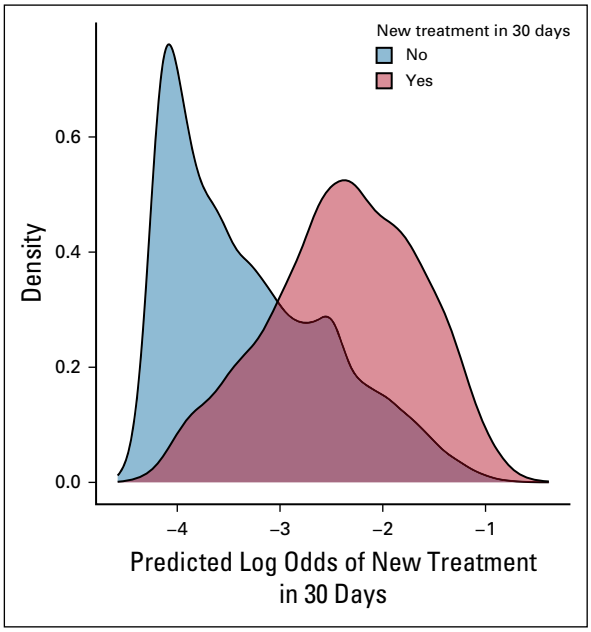

FIG 2. Distribution of predicted log odds of new treatment in 30 days versus actual subsequent treatment in 30 days (n = 34,770 imaging reports for 1,952 patients). Histogram of model output 2 (predicted log odds of new treatment within 30 days), by actual rate of new treatment within 30 days, following each imaging report, demonstrating that patients with higher predicted odds of initiating new treatment after any given time point had a higher actual rate of initiating new treatment. Test set, restricted to imaging reports at least 6 months before the cohort censoring date of December 31, 2018.

TABLE 3. Utility of Model Output for Identifying Manually Annotated Cancer Progression (Held-Out Test Set)

Cohort	No. of Patients	No. of Reports	Outcome: Manually Annotated Cancer Progression, AUC (95% CI)	
			Predictor: Log Hazard of Mortality per Year	Predictor: Log Odds of New Treatment in the Next 30 Days
Non–small-cell lung cancer	69	937	0.79 (0.67 to 0.88)	0.86 (0.75 to 0.95)
Colorectal cancer	41	551	0.76 (0.60 to 0.91)	0.80 (0.65 to 0.93)
Non–small-cell lung cancer plus colorectal cancer	110	1,488	0.78 (0.69 to 0.86)	0.84 (0.75 to 0.90)

NOTE. Test set.
Abbreviation: AUC, area under the receiver operating characteristic curve.

Predicting new treatment and clinical trial enrollment. In the full test set, the estimated log odds of new treatment within 30 days predicted actual new treatment within 30 days with an AUC of 0.77 (95% CI, 0.74 to 0.81). In a sensitivity analysis restricted to imaging reports generated after tumor genomic profiling, the AUC was 0.81 (95% CI, 0.77 to 0.85). The results were similar in another sensitivity analysis in which the model was re-trained using only reports through 2017 and evaluated on test set reports from 2018 (Data Supplement). The distribution of predictions by actual initiation of new palliative intent systemic therapy within 30 days is illustrated in Figure 2. Performance among patients with specific types of cancer is provided in Table 2. The alternate model output, corresponding to the predicted log hazard of mortality per year, was also itself associated with new treatment within 30 days (AUC 0.64; 95% CI, 0.61 to 0.68; Table 2). For comparison, the output of our prior algorithm trained to recapitulate manual annotations of cancer progression among patients with lung cancer[6] (n = 30,377 reports for 2,276 patients) yielded an AUC of 0.71 (0.67 to 0.74) for predicting new treatment. Performance of each model for predicting new clinical trial enrollment was similar to performance for predicting any new treatment, as expected, since new clinical trials were a subset of new treatments (Table 2).

Predicting manual annotations of cancer progression. Among 1,488 imaging reports for 110 test set patients with colorectal cancer or non–small-cell lung cancer whose records were manually reviewed, the predicted log hazard of mortality generated by the trained model yielded an AUC of 0.78 (95% CI, 0.69 to 0.86), and the predicted log odds of new treatment yielded an AUC of 0.84 (95% CI, 0.75 to 0.90), for identifying manually annotated cancer progression (Table 3). The distribution of predicted log hazard of mortality by actual progression annotations is illustrated in Figure 3, and the distribution of predicted log odds of new treatment by actual progression annotations is illustrated in Figure 4.

Model Performance in Under-Represented Subgroups

Within our cohort, 2018 patients (9.6%) were non-White and 661 (3.2%) were Black. Model performance in these subgroups was similar to that in the full cohort (Table 2).

Model Interpretability

Individual words associated with the most positive and most negative coefficients in linear models trained to use imaging report text to predict deep NLP model output are provided in Table 4. Similar words were associated with higher predicted log hazard of mortality and log odds of new treatment, including metastatic and increase (likely representing radiologists' descriptions of increasing tumor burden).

Model Uncertainty

The distribution of model outputs vs the standard deviation of the outputs is illustrated in the Data Supplement. The standard deviation, which might be interpreted as a metric

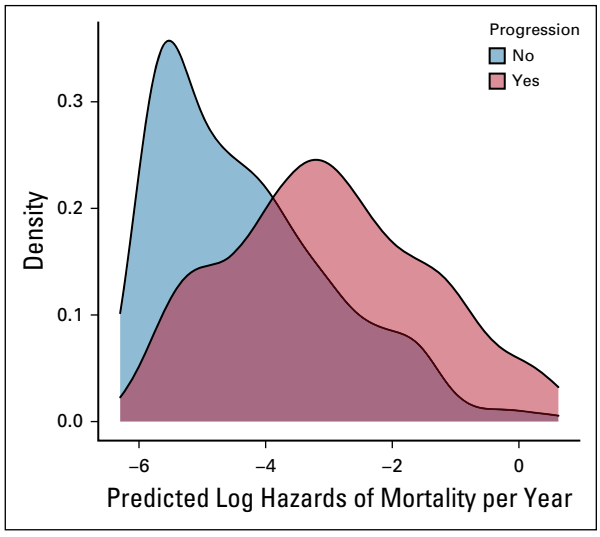

FIG 3. Distribution of predicted log hazards of mortality versus manually annotated cancer progression among patients with non–small-cell lung cancer or colorectal cancer (n = 1,488 imaging reports for 110 patients). Histogram of model output 1 (predicted log hazards of mortality per year), by manual annotation of the presence or absence of cancer progression on each imaging report, demonstrating that patients with higher predicted log hazards of mortality at any given time point were more likely to have cancer progression manually annotated at that time point. Test set, restricted to imaging reports at least 6 months before the cohort censoring date of December 31, 2018.

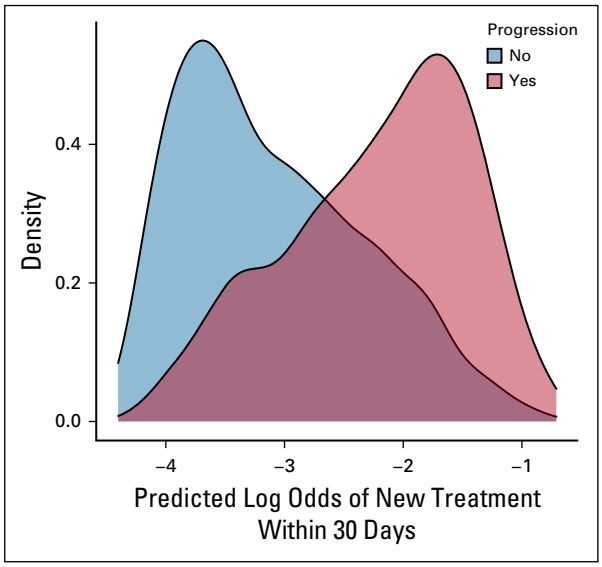

FIG 4. Distribution of predicted log odds of new treatment within 30 days versus manually annotated cancer progression among patients with non–small-cell lung cancer or colorectal cancer (n = 1,488 imaging reports for 110 patients). Histogram of model output 2 (predicted log odds of new palliative-intent systemic therapy within 30 days), by manual annotation of the presence or absence of cancer progression on each imaging report, demonstrating that patients with a higher predicted log odds of new treatment at any given time point were more likely to have cancer progression manually annotated at that time point. Test set, restricted to imaging reports at least 6 months before the cohort censoring date of December 31, 2018.

TABLE 4. Words Associated With Model Output

Association with Model Output	Model Output: Log Hazard of Mortality per Year	Model Output: Log Odds of New Palliative-Intent Systemic Treatment in the Next 30 Days
Most associated	Metastatic	Metastatic
^	Ascites	Metastases
I	Metastases	Increased
I	Pleural	Metastasis
I	Increased	Increase
I	Signed by	Ascites
…		
I	Nodule	Signatures
I	Cyst	Exam
I	Resection	Resection
I	No pleural	No new
I	No new	Status
V	No	No
Least associated	cc	cc

NOTE. Each row represents a coefficient in a linear regression model with LASSO regularization trained to predict the outputs of the clinical inflection point detection within the validation set.

Abbreviation: LASSO, Least Absolute Shrinkage and Selection Operator.

of model uncertainty at any given time point, was highest near the middle of the distribution of each model output.

DISCUSSION

A deep learning model trained to use radiology report text to identify clinical inflection points was reliably able to predict prognosis and changes in systemic therapy among patients with cancer. This approach could be used to identify patients experiencing such inflection points and then carefully match them to clinical trials for which they might be eligible. Algorithm outputs were correlated with human abstraction of cancer progression from EHRs, but they were equally or more predictive of prognosis and treatment changes than such curation.

Although this method could inform cancer care delivery strategies in general, a key use case will be to inform targeted delivery of clinical trials options to patients at specific moments in time. This could overcome a critical barrier to enrollment in precision oncology clinical trials in practice: Patients experience periods of improvement and worsening over the course of their disease trajectories. Even if it is clear that patients' tumor characteristics might render them eligible for a specific clinical trial, patients may actually be candidates for trials only during times when their disease is worsening. From a practical perspective, it is no longer feasible for community-based physicians to track all potential investigational trials for their patients. Even oncologists practicing at major cancer centers, whose practice focuses on a particular cancer type, struggle to identify the right trial for the right patient at the right time. This analysis demonstrates the potential for harnessing deep learning methods to facilitate streamlined procedures for clinical trial matching. A model-detected inflection point could serve as a trigger for oncologists and research staff to review patients for clinical trials.

Strengths of this analysis include its derivation from a large number of patients with multiple different types of cancer. As demonstrated in evaluations of performance by cancer type and on the basis of explanatory linear models, the algorithm focused particularly on terms consistent with worsening cancer across cancer types, rather than simply learning, for example, that patients with pancreatic cancer generally have worse prognoses than those with colorectal cancer. It also generated estimates of prediction uncertainty, which could be relevant to any cancer care delivery intervention. By simultaneously predicting prognosis and changes in treatment, it could further facilitate filtering of interventions on the basis of prognostic criteria. However, we would not propose using the prognostic estimates generated by our model to exclude patients from trials. On the contrary, these estimates appear to capture concepts associated with worsening disease (Table 4), such that poor

prognoses might best constitute a potential criterion for clinical trial readiness in implementation studies.

Limitations include a single-institution cohort of patients participating in a precision medicine study for which tumor genomic profiling was an eligibility criterion, given the intent to apply this approach to inform matching to clinical trials of targeted therapy. In such a context, complex dynamics involving selection bias and temporal selection bias[22]—cohort entry specifically because of clinical worsening—may apply. In this analysis, however, model performance was similar after restricting to imaging reports collected after genomic testing. Additionally, the need to apply algorithms fairly is always a concern when implementing machine learning in clinical practice,[23] although we found that performance was similar in underrepresented subgroups. We applied a surrogate linear regression model to identify words and short phrases associated with the output of our deep learning models. Further work would be needed to apply more granular model explanation frameworks[20,21] to the temporal deep neural network architecture over text documents in this study to understand the contribution of nonlinear interactions among phrases and across time to individual predictions. Finally, our model may detect not just changes in cancer status but also decisions that have already been made to change treatment (phrases such as restaging before initiation of new therapy or descriptions of intravenous access placement), to the extent that such decisions might be reflected in the text of imaging reports. Nevertheless, such reports might still reflect clinical inflection points, identification of which could remain useful for prompt targeted delivery of clinical trial information.

In conclusion, a deep NLP model can be trained to identify clinical inflection points, at which patients with cancer may benefit from access to information about clinical trial options, using unstructured EHR data. This could be a generalizable approach to clinical inflection point detection across institutions. Next steps at our center will include piloting incorporation of this algorithm into an institutional clinical trials matching tool.[24]

AFFILIATION

[1]Division of Population Sciences, the Knowledge Systems Group, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA

CORRESPONDING AUTHOR

Kenneth L. Kehl, MD, MPH, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215; Twitter: @kenlkehl; e-mail: kenneth_kehl@dfci.harvard.edu.

DISCLAIMER

Funding organizations had no role in the design or conduct of the study: collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; of the decision to submit the manuscript for publication. The content is solely the responsibility of the authors.

SUPPORT

Supported by Wong Family Foundation, Simeon J. Fortin Foundation, Doris Duke Charitable Foundation, National Cancer Institute (K99CA245899-01) (K.L.K.).

AUTHOR CONTRIBUTIONS

Conception and design: Kenneth L. Kehl, Deborah Schrag
Financial support: Kenneth L. Kehl, Deborah Schrag
Provision of study materials or patients: Eva M. Lepisto, Deborah Schrag
Collection and assembly of data: Kenneth L. Kehl, Eva M. Lepisto, Alexander Gusev
Data analysis and interpretation: Kenneth L. Kehl, Stefan Groha, Eva M. Lepisto, Haitham Elmarakeby, James Lindsay, Eliezer M. Van Allen, Michael J. Hassett, Deborah Schrag
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).

Kenneth L. Kehl
Honoraria: Roche, IBM
Consulting or Advisory Role: Aetion

Eva M. Lepisto
Employment: Kiniksa
Stock and Other Ownership Interests: Kiniksa

Alexander Gusev
Patents, Royalties, Other Intellectual Property: Pending patent on an immunotherapy biomarker

Eliezer M. Van Allen
Stock and Other Ownership Interests: Syapse, Tango Therapeutics, Genome Medical, Microsoft, Ervaxx, Monte Rosa Therapeutics, Manifold Bio
Consulting or Advisory Role: Syapse, Roche, Third Rock Ventures, Takeda, Novartis, Genome Medical, Invitae, Illumina, Tango Therapeutics, Ervaxx, Janssen, Monte Rosa Therapeutics, Manifold Bio
Speakers' Bureau: Illumina
Research Funding: Bristol Myers Squibb, Novartis
Patents, Royalties, Other Intellectual Property: Patent on discovery of retained intron as source of cancer neoantigens, Patent on discovery of chromatin regulators as biomarkers of response to cancer immunotherapy, and Patent on clinical interpretation algorithms using cancer molecular data
Travel, Accommodations, Expenses: Roche/Genentech

Michael J. Hassett
Research Funding: IBM

Deborah Schrag
Stock and Other Ownership Interests: Merck
Honoraria: Pfizer
Consulting or Advisory Role: JAMA—Journal of the American Medical Association
Research Funding: AACR, GRAIL
Patents, Royalties, Other Intellectual Property: PRISSMM model is trademarked, and curation tools are available to academic medical centers and government under Creative Commons license
Travel, Accommodations, Expenses: IMEDEX, Precision Medicine World Conference
Other Relationship: JAMA—Journal of the American Medical Association

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors would like to acknowledge the DFCI Oncology Data Retrieval System (OncDRS)[25] for the aggregation, management, and delivery of the clinical and operational research data used in this project.

Manually curated outcomes data were collected and managed using Research Electronic Data Capture (REDCap) electronic data capture tools hosted at Partners Healthcare. REDCap 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 packages; and (4) procedures for importing data from external sources.[19] Electronic health records were curated using the PRISSMM v2.0 data standard for defining clinical cancer end points from unstructured text. The PRISSMM 2.0 data standard, training modules, and curation directives are freely available to academic and governmental entities with a license. They will be made available to commercial entities at cost. Questions and requests for a PRISSMM license and data curation tools should be sent to PRISSMM@dfci.harvard.edu.

REFERENCES

1. Sherman RE, Anderson SA, Dal Pan GJ, et al: Real-world evidence—What is it and what can it tell us? N Engl J Med 375:2293-2297, 2016
2. Manz CR, Parikh RB, Small DS, et al: Effect of integrating machine learning mortality estimates with behavioral nudges to clinicians on serious illness conversations among patients with cancer: A stepped-wedge cluster randomized clinical trial. JAMA Oncol 6:e204759, 2020
3. Garraway LA, Verweij J, Ballman KV: Precision oncology: An overview. J Clin Oncol 31:1803-1805, 2013
4. Flaherty KT, Gray R, Chen A, et al: The molecular analysis for therapy choice (NCI-MATCH) trial: Lessons for genomic trial design. J Natl Cancer Inst 112:1021-1029, 2020
5. Meric-Bernstam F, Brusco L, Shaw K, et al: Feasibility of large-scale genomic testing to facilitate enrollment onto genomically matched clinical trials. J Clin Oncol 33:2753-2762, 2015
6. 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
7. Schrag D: GENIE: Real-world application. ASCO Annual Meeting, Chicago, IL, June 4, 2018
8. Arbour KC, Luu AT, Luo J, et al: Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. Cancer Discov 10.1158/2159-8290.CD-20-0419 [epub ahead of print on September 21, 2020]
9. 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
10. MacConaill LE, Garcia E, Shivdasani P, et al: Prospective enterprise-level molecular genotyping of a cohort of cancer patients. J Mol Diagn 16:660-672, 2014
11. Sholl LM, Do K, Shivdasani P, et al: Institutional implementation of clinical tumor profiling on an unselected cancer population. JCI Insight 1:e87062, 2016
12. Saeb S, Lonini L, Jayaraman A, et al: The need to approximate the use-case in clinical machine learning. Gigascience 6:1-9, 2017
13. James G, Witten D, Hastie T, et al: An Introduction to Statistical Learning: With Applications in R. New York, NY, Springer Publishing Company, Incorporated; 2014
14. Paszke A, Gross S, Massa F, et al: PyTorch: An imperative style, high-performance deep learning library. arXiv. 2019. http://arxiv.org/abs/1912.01703
15. Kehl KL: Imaging inflection points. Available at: github.com/prismmnlp/imaging_inflection_points. Accessed May 10, 2021.
16. Gal Y, Ghahramani Z: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv 48, 2015. http://www.jmlr.org/proceedings/papers/v48/gal16.pdf
17. Harrell FE, Lee KL, Mark DB: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361-387, 1996
18. Mandrekar JN: Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 5:1315-1316, 2010
19. Harris PA, Taylor R, Thielke R, et al: Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42:377-381, 2009
20. Ribeiro M, Singh S, Guestrin C: "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Stroudsburg, PA, Association for Computational Linguistics, 2016:97-101
21. Lundberg S, Lee S-I: A unified approach to interpreting model predictions. arXiv. 2017. http://arxiv.org/abs/1705.07874
22. Kehl KL, Schrag D, Hassett MJ, et al: Assessment of temporal selection bias in genomic testing in a cohort of patients with cancer. JAMA Netw Open 3:e206976, 2020
23. Rajkomar A, Hardt M, Howell MD, et al: Ensuring fairness in machine learning to advance health equity. Ann Intern Med 169:866-872, 2018
24. Lindsay J, Fitz CDV, Zwiesler Z, et al: MatchMiner: An open source computational platform for real-time matching of cancer patients to precision medicine clinical trials using genomic and clinical criteria. bioRxiv. 2017. https://doi.org/10.1101/199489
25. 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 Genomics 6:18-25, 2015

