Abstract

IMPORTANCE Predicting short- and long-term survival of patients with cancer may improve their care. Prior predictive models either use data with limited availability or predict the outcome of only 1 type of cancer.

OBJECTIVE To investigate whether natural language processing can predict survival of patients with general cancer from a patient’s initial oncologist consultation document.

DESIGN, SETTING, AND PARTICIPANTS This retrospective prognostic study used data from 47 625 of 59 800 patients who started cancer care at any of the 6 BC Cancer sites located in the province of British Columbia between April 1, 2011, and December 31, 2016. Mortality data were updated until April 6, 2022, and data were analyzed from update until September 30, 2022. All patients with a medical or radiation oncologist consultation document generated within 180 days of diagnosis were included; patients seen for multiple cancers were excluded.

EXPOSURES Initial oncologist consultation documents were analyzed using traditional and neural language models.

MAIN OUTCOMES AND MEASURES The primary outcome was the performance of the predictive models, including balanced accuracy and receiver operating characteristics area under the curve (AUC). The secondary outcome was investigating what words the models used.

RESULTS Of the 47 625 patients in the sample, 25 428 (53.4%) were female and 22 197 (46.6%) were male, with a mean (SD) age of 64.9 (13.7) years. A total of 41 447 patients (87.0%) survived 6 months, 31 143 (65.4%) survived 36 months, and 27 880 (58.5%) survived 60 months, calculated from their initial oncologist consultation. The best models achieved a balanced accuracy of 0.856 (AUC, 0.928) for predicting 6-month survival, 0.842 (AUC, 0.918) for 36-month survival, and 0.837 (AUC, 0.918) for 60-month survival, on a holdout test set. Differences in what words were important for predicting 6- vs 60-month survival were found.

CONCLUSIONS AND RELEVANCE These findings suggest that models performed comparably with or better than previous models predicting cancer survival and that they may be able to predict survival using readily available data without focusing on 1 cancer type.

Key Points

Question Can natural language processing be used to predict the survival of patients with cancer from their initial oncologist consultation document?

Findings In this prognostic study of 47 625 patients with cancer, 6-, 36-, and 60-month survival were predicted using both traditional and neural language models with performance similar to or better than that found in prior work.

Meaning These findings suggest it is feasible to predict survival of patients with cancer using a common cancer document without additional data and without training separate models for specific types of cancer.

Supplemental content

Author affiliations and article information are listed at the end of this article.
Introduction

Cancer is a leading cause of death globally, with survival depending on variables including cancer site, tumor type, age, sex, and comorbidities. Accurately predicting an individual patient’s survival could be used to improve cancer care. For example, it might suggest earlier referral to palliative care resources or consideration of more aggressive therapies upfront. Traditionally, survival rates are calculated retrospectively and categorized by only a few factors, primarily by cancer site and histology. Despite familiarity with these odds, oncologists can be inaccurate when predicting an individual patient’s survival prospectively; they have trouble factoring in personal factors such as age.

Predictive models trained by machine learning may allow more personalized predictions by using many features of a patient’s particular characteristics and disease, and have been shown to outperform the prediction of treating oncologists. Some models developed to date utilize structured data, that is, data that are processed into specific features such as the presence of genetic markers, demographics, or specific aspects of clinical history. This may limit the widespread use of such models, as data availability varies among cancer treatment centers and between patients. It also limits what data can be used for a model, as not all clinical data are easily coded or categorized for extraction and analysis. The use of unstructured data, such as text within medical documents, may address these drawbacks. Almost all patients receiving treatment for cancer have an initial consultation document from their oncologist. Such documents generally have many details relevant to survival, for example, tobacco use or marital status, even if the clinic does not routinely store such data in structured data sets.

The use of machine learning with documents, known as natural language processing (NLP), has increasingly been applied throughout medicine. Many of the applications both in medicine generally and cancer specifically have utilized smaller, specific documents such as radiology or pathology reports. Some studies such as that by Liu et al have used patient encounter documents to predict the onset of 3 noncancer illnesses. They found that models using the unstructured text data in these documents outperformed using only structured data, and that adding structured data like demographics and laboratory data increased performance by only a marginal amount. Within cancer, recent work predicted survival in patients with lung cancer by extracting structured data from unstructured document data and used nonneural NLP on progress notes to update an individual’s prognosis.

Neural networks are a type of machine learning modeled after the interconnectedness of neurons. When utilized in NLP, neural models can develop a more complex understanding of language, such as the presence of words with respect to each other, even when not adjacent. We were unable to find prior work using neural NLP methods to predict the survival of patients with general cancer using oncologist consultation documents, nor were such works identified in recent reviews of neural network applications in cancer, in general medical applications, or in a recent review of machine learning techniques used in cancer survival prediction.

Our work sought to develop and evaluate neural NLP models that predict the survival of patients with general cancer using their initial oncologist consultation, without the use of structured data. By using this common document without structured data, we hoped to build models that would not be constrained by requiring the collection and processing of specific data. Similarly, we utilized our models with the patient population with general cancer seen across a provincial cancer control system, as opposed to focusing on predicting survival in patients with 1 cancer location treated in a single center. This allowed us to investigate the performance of more generalizable models. We also did not utilize any feature extraction techniques for our neural methods, instead allowing the neural networks to directly utilize the text after only low-level text preparation. We further sought to contribute to the field by comparing 2 common neural networks used in NLP with a recently developed neural network using transformers. We hypothesized that our models would be able to achieve predictive performance at least in line with prior work, predicting with an accuracy,
balanced accuracy (BAC), and receiver operating characteristics area under the curve (AUC) all above 0.800 when tested upon a never-seen holdout set, and that the neural models would outperform the traditional nonneural bag-of-words (BoW) algorithm, with the models using transformers the best performing.

Methods
The University of British Columbia BC Cancer Research Ethics Board approved this prognostic study, including exemption from requiring informed consent from study participants due to this being infeasible. We report this study following the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline.

Data Source
Our study cohort was selected from the 59,800 patients who started cancer care at BC Cancer between April 1, 2011, and December 30, 2016. These patients sought care for malignant disease or for precancerous or nonmalignant pathological findings that required specialist cancer care. BC Cancer provides all radiation therapy in British Columbia, and more than 85% of medical oncologists in the province have affiliation with the organization. BC Cancer cares for patients with cancer at 6 centers located in geographically diverse settings throughout the province and oversees systemic therapy at most of the 53 smaller Community Oncology Network sites. Data were provided to us by BC Cancer, which obtained mortality data from BC Vital Statistics administrative data.

Data Selection and Preparation
We excluded patients who had more than 1 cancer diagnosis. We included all patients who had at least 1 valid document recorded as being a medical or radiation oncologist consultation generated within 180 days of diagnosis irrespective of disease and selected the document closest to the date of diagnosis. At BC Cancer, some patients may not see an oncologist for a few months after diagnosis as they may first receive surgery elsewhere, and come to the organization for chemotherapy, radiation therapy, or other specialized treatments.

We applied some preprocessing to documents before they were used by our models, as outlined in eMethods in Supplement 1. When used with BoW models, words had their endings removed to become tokens. We calculated survival as the number of months between the selected document being generated, and either the patient’s recorded death date or April 6, 2022, when mortality data were last extracted from administrative data. We then produced binary labels for whether a patient survived 6, 36, or 60 months or not.

Natural Language Models
Language models are used in NLP to assign probabilities to the order of words, such as determining whether one sequence of words is more likely than another. We can extend this understanding to other tasks, such as predicting the binary survival outcomes in this work. Herein, we compare 4 language models: the traditional nonneural method, BoW, and 3 models using neural networks, consisting of convolutional neural networks (CNN), long short-term memory (LSTM), and a more recent transformer model, bidirectional encoder representations from transformers (BERT). We show simplified diagrams of these models in the Figure. Neural networks allow models to have more complicated understandings of language, such as how nearby or even distant words can change each other’s meaning. We provide further description of these models, including hyperparameters, in the eMethods in Supplement 1. We implemented our models in the Python 3 programming language. We describe further details, including libraries used and techniques for dealing with imbalance between survivors and nonsurvivors, in the eMethods in Supplement 1. Our code and trained models are publicly available through a GitHub repository.
Statistical Analysis

The primary outcome was model performance when predicting patients’ survival of 6, 36, or 60 months. To avoid overfitting, when a model performed well on its training data but not on new data,\textsuperscript{31} we randomly separated our data into training (70%), development (10%) and testing (20%) sets. We trained our neural models on the training data up to 100 times (epochs), stopping when performance evaluated on the development set did not increase for 5 epochs. We used the training and development sets to tune our model hyperparameters, then evaluated the model on the test set to ensure no tuning or development could overfit the test set. We assessed model performance by reporting the accuracy, BAC, AUC, F1 score, sensitivity, specificity, and other metrics of this test run, which we define in eTable 1 in Supplement 1.

To better understand our BoW models, we measured a word’s importance by the absolute value of coefficient weights of the L2-regularized logistic regression model. For our neural models, we used the Captum interpretability library for PyTorch\textsuperscript{32} to implement integrated gradients.\textsuperscript{33} This attribution method allows us to visualize what words in a document positively or negatively contribute to a prediction. While interpretable, the visualization is specific to words in an individual document; a word’s or phrase’s shown importance may be different in the context of a different document. To protect privacy, we anonymized the shown text by changing dates, names, and other identifying components, ensuring this did not change the interpretation.

Figure. Simplified Diagrams of the Language Models Used in This Work

A, The bag-of-words model counts word occurrences in a document, which is then used by a traditional machine learning algorithm. B, The convolutional neural network model understands a document in small adjacent clusters of words called convolutions (one is shown with black lines). The model can then learn to predict from combinations of these convolutions. C, The long short-term memory model updates the prediction by reading the document one word at a time. It has a memory cell that allows it to remember some prior context (dotted lines). D, The bidirectional encoder representations from transformers model can understand how each word is connected to all other words in the document but can only read small portions of text. One word’s possible connections are indicated by a black line.
Results

Patient and Document Selection
Of the 59,800 patients from BC Cancer, we excluded 2784 who were recorded as starting cancer care multiple times. Of these remaining 57,016 patients, 9391 did not have a consultation from a medical or radiation oncologist within 180 days following their cancer diagnosis. This left 47,625 patients fulfilling our inclusion and exclusion criteria; all had updated mortality data. This cohort consisted of 25,428 female patients (53.4%) and 22,197 male patients (46.6%) with a mean (SD) age of 64.9 (13.7) years (Table 1). We observed patients surviving a mean (SD) of 61.7 (40.3) months after diagnosis, and 59.9 (39.9) months after their initial oncologist consultation; these numbers are limited by the observation period, which was a minimum of 5 years. For our prediction targets, 41,447 (87.0%) survived 6 months, 31,143 (65.4%) survived 36 months, and 27,880 (58.5%) survived 60 months, calculated from the initial oncologist consultation.

Model Performance
In Table 2, we show how well our different NLP models predict whether patients will survive 5 years after their initial oncologist consultation when evaluated on a holdout test set. We see numerically similar performance between BoW, CNN, and LSTM, with BAC above 0.800 and AUC above 0.900. BERT has a lower performance across most metrics.

Performance by Survival Length
Table 3 compares the performance for our BoW model with one of our best neural models, CNN, when predicting the different survival lengths. On a holdout test set, BoW performed best for predicting 6-month survival, achieving a BAC of 0.856 (AUC 0.928); CNN had best performance for predicting 36-month survival, with a BAC of 0.842 (AUC, 0.918) and for 60-month survival, with a BAC of 0.837 (AUC, 0.918). We see similar performance between both models. Additional metrics, and performance of all models when predicting all durations, are described in eTable 2 in Supplement 1.

Interpretation
When observing the importance of tokens (words with endings removed), we see some similar tokens have top 10 importance for BoW models used to predict both 6- and 60-month survival (Table 4), but also some differences. For example, the token palliative and palliation, is the most important feature in both. Different cancer locations or types are important for the 2 durations; breast and prostate are positive predictors for 6-month survival, while liver, glioblastoma, and lung are negative predictors for 60-month survival. The token NO is a top 10 positive predictor for 60-month survival, although we see no other TNM classification tokens in top 10 features for either survival duration.

eFigures 1 and 2 in Supplement 1 show the importance of words in a patient's document for our CNN models, which correctly predict that this patient survives 6 but not 60 months. We again see similarities and differences between the models developed to predict surviving both lengths. For example, the patient's age had negative importance when predicting whether they would survive 60 months, but not 6 months. Positive aspects of the surgery and pathology, such as having clear margins and no lymphatic or vascular involvement, predicted positively in both models, but more so for the 6-month survival prediction. Medical comorbidities such as congestive heart failure and Barrett esophagus were negative predictors for 60-month survival, but less so in predicting 6-month survival.
Discussion

In this study, we trained and evaluated NLP models using both traditional and neural models to predict whether a patient with cancer will survive 6, 36, and 60 months using only their initial oncologist consultation document and no other data. The performance of our models, when evaluated on a never-before-seen internal holdout set, achieved accuracy, BAC, and AUC over 0.800, with our best models achieving AUC over 0.900. This performance was similar to or better than prior work seeking to predict survival of patients with cancer, despite using data that were more generalizable and readily available. Prior studies have predicted cancer survival for specific tumor sites such as breast or lung, or with the use of structured data such as processed clinical and genetic characteristics.

Our results suggest it is possible to predict the survival of patients with cancer without having to construct structured data sets, or limiting the predictions to specific types or locations of cancer. The availability of structured data may vary. Given the widespread availability of initial oncologist consultation documents, this opens up the possibility of more easily training and using such models across cancer types at different cancer centers.

Table 2. Model Performance for Predicting 60-Month Survival After Patients' Initial Oncologist Consultation Document Was Generated

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>BAC</th>
<th>AUC</th>
<th>F1</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.823</td>
<td>0.835</td>
<td>0.915</td>
<td>0.856</td>
<td>0.798</td>
<td>0.871</td>
</tr>
<tr>
<td>CNN</td>
<td>0.828</td>
<td>0.837</td>
<td>0.918</td>
<td>0.862</td>
<td>0.809</td>
<td>0.866</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.826</td>
<td>0.834</td>
<td>0.914</td>
<td>0.860</td>
<td>0.810</td>
<td>0.859</td>
</tr>
<tr>
<td>BERT</td>
<td>0.804</td>
<td>0.813</td>
<td>0.894</td>
<td>0.842</td>
<td>0.787</td>
<td>0.838</td>
</tr>
</tbody>
</table>

Abbreviations: AUC, receiver operating characteristics area under the curve; BAC, balanced accuracy; BERT, bidirectional encoder representations from transformers; BoW, bag-of-words; CNN, convolutional neural network; LSTM, long short-term memory.

Table 3. Model Performance for Predicting the Given Numbers of Months After Patients' Initial Oncologist Consultation Document Was Generated

<table>
<thead>
<tr>
<th>Model*</th>
<th>Survival, mo</th>
<th>Accuracy</th>
<th>BAC</th>
<th>AUC</th>
<th>F1</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>6</td>
<td>0.849</td>
<td>0.856</td>
<td>0.928</td>
<td>0.907</td>
<td>0.847</td>
<td>0.866</td>
</tr>
<tr>
<td>CNN</td>
<td>6</td>
<td>0.843</td>
<td>0.853</td>
<td>0.926</td>
<td>0.903</td>
<td>0.839</td>
<td>0.867</td>
</tr>
<tr>
<td>BoW</td>
<td>36</td>
<td>0.837</td>
<td>0.837</td>
<td>0.915</td>
<td>0.872</td>
<td>0.836</td>
<td>0.838</td>
</tr>
<tr>
<td>CNN</td>
<td>36</td>
<td>0.844</td>
<td>0.842</td>
<td>0.918</td>
<td>0.878</td>
<td>0.848</td>
<td>0.836</td>
</tr>
<tr>
<td>BoW</td>
<td>60</td>
<td>0.823</td>
<td>0.835</td>
<td>0.915</td>
<td>0.856</td>
<td>0.798</td>
<td>0.871</td>
</tr>
<tr>
<td>CNN</td>
<td>60</td>
<td>0.828</td>
<td>0.837</td>
<td>0.918</td>
<td>0.862</td>
<td>0.809</td>
<td>0.866</td>
</tr>
</tbody>
</table>

Abbreviations: AUC, receiver operating characteristics area under the curve; BAC, balanced accuracy; BoW, bag-of-words; CNN, convolutional neural networks. * Comparison of the traditional model, BoW, with our best performing neural model, CNN.

Table 4. Top 10 Tokens Used by Bag-of-Words Models for 6- and 60-Month Survival Prediction*

<table>
<thead>
<tr>
<th>Feature importance rank</th>
<th>Survival prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 mo</td>
</tr>
<tr>
<td></td>
<td>Token</td>
</tr>
<tr>
<td>1</td>
<td>Palliat</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
</tr>
<tr>
<td>3</td>
<td>Risk</td>
</tr>
<tr>
<td>4</td>
<td>Hospit</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Unfortun</td>
</tr>
<tr>
<td>7</td>
<td>Prostat</td>
</tr>
<tr>
<td>8</td>
<td>Breast</td>
</tr>
<tr>
<td>9</td>
<td>Metastat</td>
</tr>
<tr>
<td>10</td>
<td>Stage</td>
</tr>
</tbody>
</table>

* Feature importance was calculated using the absolute value of coefficient weights in these L2-regularized logistic regression models. Tokens are words in which the word endings have been removed for processing. * Note that N0 refers to a patient without disease in lymph nodes in TNM cancer staging.
To help our models be generalizable, we did minimal text processing. We did not extract features from the text to be used by our models, instead providing text directly. The documents in our data set were generated by oncologists practicing at 6 centers located in geographically varied regions without set templates or formatting requirements. Future work will be needed to externally validate our models on documents generated in other jurisdictions.

The performance of our CNN, LSTM, and BoW models was numerically similar. BERT’s inferior performance, despite its more deeply connected network, may be due to its limitation of only being able to use the first 512 tokens. The lack of clear performance gain with neural models compared with the traditional BoW model is surprising, but in line with some prior work. This may suggest survival prediction is largely dependent on the presence of words, compared with understanding how words relate to each other in a document.

We interpreted our neural models using integrated gradients and our BoW models using coefficient values. Both methods have their limitations. Feature correlation impacts L2-regularized coefficients, while integrated gradients show word importance based on their specific context. However, our results generally supported that our models used generalizable, interpretable portions of text to make their predictions, consistent with known mortality risk and protective factors. For example, palliative care is often accessed in the last months of a patient’s life, so it is unsurprising that an oncologist mentioning palliative care would support that a patient may not survive 6 months.

There are multiple avenues for further development to build on our results. While some related work found adding structured data led to only a small, likely not clinically significant improvement in performance, it may still be worthwhile investigating the addition of structured data. Future work could improve on our results by using different models or configurations, including those both more and less complex. Hardware advancement will allow more complex versions of our models, and there are new transformer models such as Longformer and BigBird that can use more text. Neural models can especially benefit from training on very large amounts of data. Future work could use our methodology to train models using widely available oncologist consultation documents from multiple jurisdictions or fine-tune our models by further training them with a relatively small number of a different jurisdiction’s documents. Future research could also adapt our models to update survival prediction after the initial consultation, such as by using progress notes. Further performance improvements and external validation may allow clinicians to use this methodology for improving care, such as better facilitating palliative resources with short-term survival predictions or using long-term predictions to consider more intense initial treatment.

Limitations
This study has some limitations. We took steps to ensure the validity of our models, including the use of a holdout test set, and eliminating location-specific text that was automatically added to documents, but we did not externally validate our models. Our work calculated survival from the first oncologist consultation document, as opposed to from diagnosis, as is common. This was done because the time from diagnosis to consultation can vary from a few weeks to a few months at BC Cancer if patients first received treatment such as surgery from outside clinicians, or initial workup from family physicians. We also note our data set consisted of patients first seen between 2011 and 2016, so we could establish 5-year survival. Given the rapid advancement of cancer treatment, accuracy may be limited for patients starting treatment presently. Finally, we did not compare our models’ performance with that of oncologists, as has been done in previous work.
Conclusions

In this prognostic study of 47,625 patients with cancer, we trained and evaluated traditional and neural models to predict whether patients will survive 6, 36, or 60 months based on their initial oncologist consultation. We evaluated our model on an internal holdout set of data, and found our best models achieved accuracy and BAC above 0.800 and AUC above 0.900. This performance was comparable with or superior to that of prior work, which has predicted survival only for specific types of cancer or used data that are more difficult to obtain.

Our findings suggest that a clinically useful survival prediction model for patients with general cancer may be possible without needing specific models for different cancer types and by using data that are readily accessible without complex data processing or data mining. These results still require external validation and have multiple avenues for further improvement; regardless, they suggest that this methodology may one day assist in care by helping patients with cancer and their treatment team by providing an individualized expectation of survival.

ARTICLE INFORMATION
Accepted for Publication: January 12, 2023.
Published: February 27, 2023. doi:10.1001/jamanetworkopen.2023.0813

Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2023 Nunez JJ et al. JAMA Network Open.

Corresponding Author: John-Jose Nunez, MD, MSc, BC Cancer and Department of Psychiatry, University of British Columbia, 2215 Wesbrook Mall, Vancouver BC, V6T 1Z3 Canada (johnjose.nunez@bccancer.bc.ca).

Author Affiliations: BC Cancer, Vancouver, British Columbia, Canada (Nunez, Leung, Ho, Bates); Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada (Nunez, Ng); Department of Psychiatry, University of British Columbia, Vancouver, British Columbia, Canada (Nunez, Bates).

Author Contributions: Drs Nunez and Bates had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Nunez, Bates, Ng.
Acquisition, analysis, or interpretation of data: All authors.

Drafting of the manuscript: Nunez, Bates, Ng.

Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Nunez, Ng.

Obtained funding: Bates.

Administrative, technical, or material support: Leung, Ho, Bates.

Supervision: Bates.

Conflict of Interest Disclosures: Dr Nunez reported receiving unrestricted grant funding from Pfizer Canada through the Pfizer Innovation Fund during the conduct of the study. Ms Leung reported receiving personal fees from AstraZeneca outside the submitted work. Dr Ho reported personal fees from AbbVie, Amgen Inc, Bayer AG, Bristol-Myers Squibb Company, Eisai Co Ltd, Janssen Pharmaceuticals, Jazz Pharmaceuticals PLC, Merck & Co Inc, Novartis AG, and Takeda Pharmaceutical Company Limited and grant funding from AstraZeneca, EMD Serono, and F. Hoffmann-La Roche AG outside the submitted work. Dr Bates reported receiving unrestricted grant funding from Pfizer Inc to BC Cancer allocated to the Psychiatry Department during the conduct of the study and receiving advisory board honoraria from Eisai Co Ltd, outside the submitted work. No other disclosures were reported.

Funding/Support: This work was supported in part by an unrestricted research grant from the BC Cancer Foundation with the funds originating from the Pfizer Innovation Fund and in part by a UBC Institute of Mental Health Marshall Fellowship (Dr Nunez).

Role of the Funder/Sponsor: The sponsors 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.

Data Sharing Statement: See Supplement 2.
REFERENCES


