Using Neural Language Models to Predict the Psychosocial Needs of Cancer Patients

by

John-Jose Andres Nunez

B.Sc., The University of British Columbia, 2013

M.D., The University of British Columbia, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Computer Science)

The University of British Columbia

(Vancouver)

April 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Using Neural Language Models to Predict the Psychosocial Needs of Cancer Patients

submitted by **John-Jose Andres Nunez** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Raymond Ng, Professor, Computer Science, UBC
Supervisor

Alan T. Bates, Clinical Associate Professor, Psychiatry, UBC
Supervisory Committee Member

Abstract

Cancer is associated not only with mortality, but also with impacts on physical, mental, and social health. When unmet, these resulting psychosocial needs are associated with worsened quality-of-life and survival. Cancer centres employ psychiatrists, counsellors, and other allied health clinicians to help address these needs. However, these needs often go unmet even when these resources exist. It can be difficult for treating oncologists to detect these needs and refer patients to these resources.

In this work, we investigated the use of neural natural language processing (NLP) models to predict these psychosocial needs using initial oncologist consultation documents. We compared a non-neural model, bag-of-words (BOW), with three neural models: convolutional neural networks (CNN), long-short term memory (LSTM), and bidirectional encoder representation from transformers (BERT). We used these models to predict self-reported emotional and informational needs around the time these documents were generated. We also used these models to predict whether the patient will have clinician-addressed needs – specifically, seeing a psychiatrist or counsellor within the five years following document generation. We compared the prediction of these psychosocial needs to predicting a non-psychosocial outcome, survival.

We found these models can predict whether patients will see a psychiatrist with balanced accuracy and receiver-operator-area-under-curve (AUC) above 0.70. This is a similar performance to comparable prior work predicting mental health outcomes. We also predicted seeing a counsellor with AUC above 0.70, but predicting self-reported psychosocial needs seemed to be a more difficult task, with these metrics usually below 0.70. We predicted the non-psychosocial outcome, sur-

vival, with higher performance. For this task, balanced accuracy was above 0.80 and AUC above 0.90. Predictions using subsets of our study population suggest that predicting these psychosocial outcomes is easier in females, and with cancer patients diagnosed with a Stage II illness. We found that CNN and LSTM models performed the best, and investigated how BERT's document size limit may hinder its performance on these tasks. This work is the first of its kind using NLP for this application, and builds a foundation to improve how these techniques may one day help cancer patients.

Lay Summary

Cancer is a leading cause of death and can harm a patient's mental health and social situation. Cancer centres employ psychiatrists, counsellors, and others to help patients deal with these problems. However, not all patients will get enough help, because it can be hard for cancer doctors to determine which of their patients will need this support.

In this project, we used a type of artificial intelligence (AI) that can read oncologists' initial reports and predict which patients will need to see a counsellor or psychiatrist, and which may need emotional or educational support. Our study used a new type of AI that uses neural networks, which are inspired by the human brain. We were able to predict which patients will need to see a psychiatrist or counsellor reasonably well. This work creates a foundation for future improvements so we can one day use these techniques to help cancer patients.

Preface

This thesis is original, unpublished, independent work by the author, John-Jose Nunez. It is covered by the University of British Columbia - BC Cancer Research Ethics Board certificate H17-03308.

Table of Contents

List of Tables

List of Figures

Glossary

AJCC American Joint Committee on Cancer

AI artificial intelligence

AUC area-under-curve

BERT Bidirectional Encoder Representations from Transformers

BOW bag-of-words

CNN convolutional neural network

EMR electronic medical record

LSTM long short-term memory

ML machine learning

NLP natural language processing

NPV negative predictive value

PPV positive predictive value

PSSCAN-R Psychosocial Screen for Cancer - Revised

RNN recurrent neural network

TD-IDF term frequency — inverse document frequency

Acknowledgements

I would like to thank the many supervisors and mentors who have helped me to get where I am today. Prof. Gregor Kiczales for awakening my love of research and computer science, and Profs. Harley Kurata and Michael Doebeli for supporting my interesting combining the life and computational sciences. Drs. Raymond Lam and all those involved in the UBC Psychiatry Research Track, which made it possible to pursue this work during residency. My supervisors Dr. Alan Bates and Professor Raymond Ng, for believing in me, sticking with me, and for all their support, guidance, and wisdom over these past years.

I was able to use this dataset thanks to the hard work of many throughout BC Cancer and PHSA including Arthur Hastings, Bonnie Leung, and Dr. Cheryl Ho. In particular, I would like to thank Colleen Wong and the Data Requests team for the many, many rounds of help.

I would also like to thank my friends. Victor, for setting me upon the path of being a total nerd at a young age; Kuba, for continuing this trend, and suggesting I do this degree in the first place. Paige, Darren, Liv, Daniel, Mike, Sebastian, Magnus, and others for the much needed distraction, meals, and support.

My family has played an invaluable role; Daniel and Natalie for keeping me grounded via chats and hot tubs, my parents, and my nonna and Auntie Jojo, for their love, support, and food.

I could not have done this work without Becky, whose daily support, constant inspiration, and empathy as a former grad student has been truly essential.

Lastly, I would like to acknowledge the contribution of the nearly sixty thousands patients whose data made this work possible. Each row of my dataset represents a person and their unique life; it is an immense privilege to use their data.

Dedication

To my parents:

To my father, whose lifelong dedication to my success was so strong that, halfway through this project, he ensured we had a first-hand experience of the psychosocial impacts of cancer on both patients and their families.

To my mother, whose limitless, constant, unending belief in me gave me the courage to pursue something as silly as a master's thesis in computer science during residency.

Chapter 1

Introduction

Cancer is not only a leading cause of death worldwide [16], but a disease associated with substantial impacts on physical, mental, and social health [65]. Patients are at an increased risk of developing mental illnesses following diagnosis [45], and the around one-third of cancer patients who already have a mental health condition before cancer diagnosis are at particular risk for worsened distress. Cancer can impact a patient's socioeconomic status, such as preventing them from working, and can strain their relationships including with their caregivers [15, 24, 63]. Given these challenges, cancer patients may need support in multiple areas, including psychological, informational, social, and physical symptom domains [61].

Factors such as social isolation, depression, anxiety and these unmet *psychosocial cancer needs* put patients at risk for not only worse quality-of-life, but also survival [48, 53, 54, 56]. Possible causal links between these psychosocial needs and survival include their impact on a patient's ability to follow through with cancer treatment, and their associated increase in substance use. To address these psychosocial needs, cancer centres employ psychiatrists, counsellors and other allied health staff specializing in psychosocial care for people with cancer. [17].

Despite the development of these psychosocial care fields as part of cancer care, cancer patients continue to have unmet psychosocial needs [4, 34, 67]. Achieving equity-oriented healthcare in cancer will require better support of these needs and those with concurrent mental illness [31]. While lack of sufficient psychosocial resources contributes to these unmet needs, there is also evidence that a lack of

detection may play a role, especially in high-resourced settings [60].

This is likely multifactorial; patients can be reluctant to seek supportive care, or not know of available resources [68]. It can also be hard for the treating oncologists to identify these psychosocial needs, as their primary goal is the treatment of the patient's cancer through chemotherapy, radiation, or surgery [70]. Prior work has found that treating oncologists could only identify around one-third of severely distressed patients [49, 70]. As well, their recommendations for addressing psychosocial needs, such as referring patients to counselling, did not strongly correlate with these needs. Multiple reasons for this difficulty have been postulated, such as the patient's denial of needs, the provider's time constraints, and the provider's use of close-ended questions related to the specific disease. Cultural and socioeconomic differences between oncologists and their patients can also make psychosocial concerns more difficult to discuss.

Given that unmet psychosocial needs are connected with worsened survival and quality of life in cancer patients, and that these needs can be difficult to detect, **we sought to investigate whether we could detect such needs using modern natural language processing (NLP) methods**. Specifically, we sought to predict these needs using NLP with initial oncologist consultation documents, as almost all patients will have such a document generated at the start of their cancer care. Allowing the detection of patients' psychosocial needs would allow better targeting of psychosocial resources to these patients, and so may improve quality-of-life and possibly even survival. For example, a predictive model could be incorporated into an electronic medical record (EMR) system, and suggest treating oncologists refer a patient to psychiatry when appropriate.

NLP is the branch of artificial intelligence concerned with allowing computer models to understand written human language. In this work, we investigated the use of methods using neural networks. These are techniques to build an artificial intelligence based upon how the human brain's neurons are interconnected. Neural models can learn complex relationships between data; in our case, complex relationships between words. We compare the traditional non-neural NLP method bag-of-words (BOW) with two commonly used NLP neural methods, convolutional neural network (CNN) and long short-term memory (LSTM) models. We also use a newer technique based on transformers, Bidirectional Encoder Representations

from Transformers (BERT), as well as a variation, Longformer.

We describe these models in detail in Chapter 3. BOW models simply use the number of times a word occurs in a document, so cannot learn relationships such as the ordering of words. BOW understands "He has cancer but does not have dementia" the exact same as "He has dementia but does not have cancer" as both phrases have the same words. On the other hand, neural models can build an understanding of language that takes into account complexities such as word order, or the presence of other words elsewhere in a document. CNN models use "convolutions", or smaller clusters of words, to build a more connected understanding of an entire document. LSTMs are a type of recurrent neural network (RNN) that understands a document one-word at a time, but can remember information about words that came beforehand. Transformers such as BERT allow a interconnected understanding of language after being trained on an immense amount of general language data. These models can then be fine-tuned on a specific task such as our predictions. We also use a variation of BERT called Longformer which bypasses BERT's limitation in only being able to intake a limited number of words. These models all differ in their ability to understand how words in a document are related to each other, and so may have different performance when used for predictions.

This work seeks to use and compare these models to predict a patient's psychosocial needs. As predictive targets, we include self-reported emotional and informational needs. We also include clinician-addressed needs, by having targets based on whether patients will go on to see a psychiatrist or counsellor. We seek to make these predictions using only medical consultation documents that oncologists generate at the start of a cancer patient's care, with no other data. These documents typically include details such as the patient's present illness, medications, and personal, social, and family history. While these documents are readily available at the start of cancer care, they often do not specifically record many psychosocial details. This makes their potential use for predicting psychosocial outcomes unclear, and the focus of this work.

As we discuss in the next Chapter, we were unable to find prior work using predictive models to predict the psychosocial outcomes of cancer patients. As such, this work addresses this gap in the literature by investigating the following primary hypotheses:

1. Neural NLP models can be used with a patient's initial oncologist consultation document to predict **clinician-addressed psychosocial cancer needs**, such as whether a patient will see a counsellor or psychiatrist. Specifically, they can predict these outcomes with balanced accuracy and area-under-curve (AUC) numerically at least 0.70, a performance threshold often met in other psychosocial predictions.

2. Neural NLP models can be used with a patient's initial oncologist consultation document to predict a patient's **self-reported emotional and informational cancer needs** around the time of this document being generated, also at the performance threshold of balanced accuracy and AUC being above 0.70.

3. Neural NLP models can be used with a patient's initial oncologist consultation document to predict a non-psychosocial outcome, **surviving at least five years** after this document was generated. This performance will be numerically higher than the performance of predicting the psychosocial outcomes, but will serve as a useful comparator, and balanced accuracy and AUC will be at least 0.80, a performance typically achieved by prior work predicting survival.

In this work, we seek to understand our predictions further by investigating additional questions. Given that the psychosocial needs of cancer patients can vary based on their cancer stage at diagnosis, and their biological sex, we aim to investigate whether these attributes impacted the performance of our models. Similarly, while the new BERT models have often achieved state-of-the-art performance on NLP applications, they are limited by how large of a document they can be used with, so we explore whether this limit impacts our models. These aims result in the following secondary hypotheses:

1. When neural NLP models are used with a patient's initial oncologist consultation document to predict psychosocial needs, the performance is numerically similar between **female and male patient subgroups**.

2. When neural NLP models are used with a patient's initial oncologist consultation document to predict psychosocial needs, the performance is numerically

similar between **subgroups based on a patient's cancer stage at diagnosis**.

3. The **number of tokens a model is able to use** will impact its ability to predict whether a patient will see a psychiatrist using a patient's initial oncologist consultation document. Specifically, the performance of BERT will be numerically worse than its variation, Longformer, which is able to use more words.

We discuss relevant related work in Chapter 2. We then describe our methods and the data used in our project in Chapter 3. This is followed by Chapter 4, where we show the results of our investigations alongside some interpretations of our model. In Chapter 5, we discuss these results, their limitations, and possible future work, before our conclusion in Chapter 6.

Chapter 2

Related Work

In this chapter, we start by reviewing prior work seeking to better understand the psychosocial needs of cancer patients in Section 2.1. Given the lack of such work using computational methods, we spend the rest of the chapter discussing relevant prior work using NLP for document classification tasks, first discussing such work in psychiatry (Section 2.2), and then in cancer and other fields of medicine (Section 2.3). We focus our review on work using similarly sized and complex medical documents, excluding work using documents generated in medical imaging or pathology, as these documents are usually shorter and simpler than the full medical consultation documents used in this work.

2.1 Clinical and Computational Methods to Predict the Psychosocial Needs of Cancer patients

We were largely unable to find relevant prior work seeking to use computational methods to predict the psychosocial needs of cancer patients, using NLP or machine learning using structured data. In their 2021 work, Watson et al. [75] use a statistical model called "Autoregressive Integrated Moving Average" to forecast cancer symptom complexity, including psychosocial symptoms, using structured data (season, age, sex, tumour group). However, their application was predicting the number of patients a clinic would see with high symptom complexity, as opposed to predicting a patient's symptom burden individually.

Traditionally, physicians have sought to understand the psychosocial cancer needs of patients either as part of the clinical interview, or through the use of questionnaires. The Canadian Problem Checklist was first developed by Ashbury et al. [8] to understand the needs of cancer patients, and has been used and further developed in later work [19, 62, 66]. It asks patients about emotional, informational, practical, social/family, spiritual, and physical needs. At BC Cancer, this checklist is incorporated into a larger questionnaire used to screen the psychosocial needs of cancer patients, the Psychosocial Screen for Cancer - Revised (PSSCAN-R) [42, 43].

2.2 Document Classification in Psychiatry

We reviewed past psychiatry NLP literature for work predicting outcomes similar to our psychosocial need targets. Abbe et al. [1] conducted a systematic review of text mining applications in psychiatry up to late 2013. Of the 38 studies they include in their analysis, 13 used unstructured documents from EMR as in this project. Most of these studies did not predict psychosocial outcomes, instead looking at drug safety outcomes, genetic pathways, ontology development, or data extraction. The most relevant is perhaps the work by Gara et al. [29], which looked at using a language model to confirm diagnoses of schizophrenia, but used transcripts of psychiatric interviews, not the resulting physician's consultation documents. Unsurprisingly given the time period, no work in this review used methods based upon neural networks, with BOW and other n-gram techniques often used instead.

Some more recent work has employed modern NLP techniques, such as neural network based approaches. Much of this work was conduced using a dataset used in Track 2 of the 2016 N-GRID NLP challenge [27]. This dataset consists of 1,000 de-identified initial psychiatric evaluation records, which they noted was the first corpus of mental health records available to the scientific community. Participants competed to predict the lifetime severity of a patient's mental illness, in an ordinal scale from zero to three. Karystianis et al. [36] employed a neural network with three dense layers, creating features from BOW, bag-of-string, and the presence of manually identified terms. However, they found that a "rule-based" approach (explicit instructions around specific words or phrases) performed bet-

ter than this method. Clark et al. [18] also used a three-layer densely connected neural network with various, mostly engineered features. Tran and Kavuluru [73] used only the history of presenting illness section of the documents, using a CNN previously adapted for medical text [58], with pretrained word embeddings trained on PubMed abstracts, and also tried an LSTM. Rios and Kavuluru [59] constructed a model using a CNN combined with conventional feature engineering. Dai and Jonnagaddala [20] solely used CNN models, trying different model hyperparameters, and using GoogleNews word embeddings. However, some groups from this competition did not use neural techniques, and still achieved competitive results [55]; the competition winner used an ensemble of linear methods with engineered features and a neural network [23, 27].

Some work predicting psychiatric outcomes using clinical documents did not implement neural methods, possibly due to using datasets with relatively few documents. Wu et al. [77] used 500 discharge summaries to classify whether patients had major depression or not, using a linear method based on the presence of specific words. Fernandes et al. [26] used linear and rule-based techniques to extract whether discharge summaries, in a database of 500 documents, contained patients who had suicidal ideation. Work by Ford et al. [28] did use a large database, including data from 4.6 million patients, and used engineered text features with a linear classifier to classify whether patients had bipolar disorder.

More recently, there have been some studies using and developing neural NLP techniques in mental health applications. A pre-print by Ji et al. [33] outlines the development of a pretrained BERT model for mental healthcare. Dai et al. [21] used CNN, hierarchical attention networks, and BERT to determine the diagnosis of patients in a dataset of 500 discharge summaries. Numerous studies have applied neural techniques to detect or classify mental illness indicated by social media texts, though these documents are different from ours in size, scope, and complexity [13, 81].

We were unable to find prior work in psychiatry using NLP with clinical documents to predict outcomes that take place in the future. We were also unable to find other work using neural techniques such as CNN, LSTM, BERT with documents similar to those in our study besides the few noted above using datasets with 1,000 or fewer subjects. We could not find previous studies using non-psychiatric

documents to predict psychosocial or psychiatric outcomes.

2.3 Document Classification in Cancer and other Fields of Medicine

We can find other examples of relevant work using neural methods with medical documents by looking at examples in cancer and elsewhere in medicine. Wu et al. [78] conducted a review of clinical NLP studies using "deep learning", which they define as using neural methods. Of the 212 studies they include, 88 employ text or document classification. However, many of these studies used smaller pieces of text than full clinical documents, or used clinical documents from radiology, pathology, or histology reports, which are smaller and less similar to the documents included in our dataset. Of the 88 studies which used various types of documents, only a few examples were used for tasks similar to the targets in our work, such as those predicting symptom severity (seven), life expectancy (one), and mortality (five).

Some examples of relevant work in cancer include a study by Yuan et al. [79], which estimated the survival of lung cancer patients using both structured data and unstructured text from an EMR. However, they did not use neural methods, instead using NLP techniques to extract features from text which they then fed into a linear model. Prior studies have used neural methods to extract information from text, such as the work by Banerjee et al. [9] which used a densely connected neural network with two hidden layers to detect when breast cancer patients were documented to have disease recurrence. This was akin to uses of neural networks elsewhere in medicine to automatically extract whether patients have a disease, such as work by Rajput et al. [57] using CNN and LSTM models, or work by Liang et al. [40] using LSTM models. In their more recent work, Wang et al. [74] used a CNN to predict breast cancer recurrence based on structured data in addition to features extracted from both pathology and progress notes. Unlike our work, they used documents generated until right before recurrence, as opposed to only a document at the start of cancer treatment. Beeksma et al. [11] used both structured and unstructured data from an EMR to predict life expectancy in a general population of patients using an LSTM, but extracted features from the text which was

collapsed into a monthly vector representing both structured and unstructured data, as opposed to using the neural method directly on text.

One of the most relevant prior studies may be work by Liu et al. [44]. Unlike most of the previously mentioned studies which used less than ten thousand subjects, this work used both the unstructured and structured data of around one million patients to predict the future development of three chronic illnesses - strokes, kidney failure, and heart failure. They used CNN and LSTM models directly with clinical documents, and also tried a hierarchical model, flattening each clinical encounter document into a single vector, which they then used with the neural models. They find that unstructured data allows better prediction than only using structured data, with their best performing model using both, but having only a small advantage over using only unstructured text. As in this work, they used integrated gradients to interpret their neural models.

Recently, BERT and transformer models have begun to be introduced to applications in medicine. BERT models pretrained on clinical and biomedical datasets exist, with some evidence they may lead to higher performance on relevant tasks than using models trained on general English corpora [5, 30]. Hu et al. [32] used a transformer model to estimate survival of cancer and seriously ill hospitalized patients, using structured data including disease comorbidities and genetic markers. Lin et al. [41] used BERT models to predict ICU-mortality using the texts of radiology reports, alongside other clinical and imaging features.

Based on our review, this work will address multiple gaps in the existing literature. It will be the first study attempting to predict psychosocial needs from medical documents using NLP. It may also be the first predicting psychosocial or psychiatric outcomes from non-mental health documents. In both mental health and medicine generally, it will add to only a few prior examples of using neural NLP methods on full medical documents to predict future outcomes, or outcomes that are not text extraction (e.g. what diagnosis a patient was given).

Chapter 3

Methods and Data

In Chapter 3, we describe the methodology of this work, as well as the data that we used. We report the results of our data processing here, instead of in our results chapter, Chapter 4, to keep that chapter focused on the results of using our NLP models.

In Section 3.1, we describe our data, including both the unstructured text documents our NLP models use, and the structured data used to generate our prediction targets. In Section 3.2 we explain our targets, and detail the class breakdown of them in our dataset. We describe subgroups of our study population in Section 3.3, and the specific NLP models in Section 3.4. We then detail how we trained and evaluated the models in Section 3.5, and then how we sought to interpret the models in Section 3.6. We end this chapter remarking on code availability in Section 3.7.

3.1 Dataset

Our dataset consisted of both unstructured data - clinical documents generated for cancer patients by care providers at BC Cancer - as well as structured data, which we used to generate targets and select patients. In this section, we describe the data, including how we obtained it, its characteristics, how we chose the clinical document used to train and evaluate our NLP models, and how we generated target labels.

3.1.1 Inclusion and Exclusion Criteria

Our dataset included data from 59,800 BC Cancer patients who completed the PSSCAN-R questionnaire between April 1, 2011 and December 31, 2016. This questionnaire was completed by most BC Cancer patients, generally at the start of their care. This time frame allowed full analysis of five year survival data.

We excluded patients if they were recorded as having more than one cancer treated at BC Cancer, as it would be difficult to determine which cancer affected the various outcomes we investigated. This involved dropping patients with duplicate entries in the provided cancer details and PSSCAN-R datasets.

3.1.2 Obtaining the Data

The study was approved by the UBC BC Cancer Agency Research Ethics Board with REB number H17-03309. A copy of this certificate is include in Appendix Section **??**. BC Provincial Health Services Authority Information Management and Information Technology Services and BC Cancer Data Requests collaborated to extract and provide the data.

They provided the consultation documents as Microsoft Word `.doc` files. They also provided structured data used for different parts of this project. The provided cancer details spreadsheet contains patient details such as sex, age, partial postal code, diagnosis date, death date, last appointment, last contact, last follow-up, cancer stage, metastatic status, and cancer site. Names and identifying health numbers were replaced by a study ID. A document details file provided information on each included document, such as its type, which health discipline generated it (e.g. Medical Oncology, Psychiatry), the date it was generated, and the patient it corresponds to. We also received the PSSCAN-R data, which included the patient-reported cancer needs.

3.1.3 Patient Characteristics

Table 3.1 shows the characteristics of the 53,157 subjects included in our dataset after applying inclusion and exclusion criteria, but before we split the subjects into training, validation, and testing datasets. As outlined below, some patients were excluded from certain targets if were missing data required to generate the target.

Table 3.1: Characteristics of the subjects in our dataset, including their age, sex, staging at diagnosis, and survival. Survival is calculated until death or the end of the period observed, which varies between patients but is at least approximately five years. We include months survived since the subjects were diagnosed with cancer, as well as the months survived since the generation of the initial oncologist consultation document used by our models to make predictions.

	n	%
Total	53157	100
Female	27871	52.4
Stage I	7097	13.4
Stage II	9502	17.9
Stage III	6477	12.2
Stage IV / Metastatic	6317	11.9
Unknown Stage	23764	44.7
	Mean	Standard Deviation
Age at Diagnosis	64.7	13.8
Observed Months Survived since Diagnosis	36.7	31.1
Observed Months Survived since Document	31.6	26.6

3.1.4 Document Selection

We sought to train, develop and evaluate our models by using physician consultation documents completed by the treating oncologist at the start of a patient's care at BC Cancer. To do this, we looked at documents created within five days of a patient completing their PSSCAN-R questionnaire, which is provided to patients at the start of their BC Cancer involvement. This amount of time on either side accounts for some of the variation that occurs with administering the PSSCAN-R, and for finalizing physician documents. Documents had to be marked as consultation documents as opposed to other types of documents such as progress notes. To capture only documents made by physicians involved with the direct treatment of a patient's cancer, as opposed to those who address other needs, we only included documents from medical and surgical specialties directly involved in treatment.

Specific details of the documents and how this selection was implemented can be found in Appendix Section A.1.

If multiple documents met this criteria, we included the document closest to when the PSSCAN-R questionnaire was recorded.

3.1.5 Document Preprocessing

We converted the provided raw `.doc` files into text files using the `win32com.client` Python library to automate Microsoft Word's `.doc` to `.txt` conversion.

We then cleaned the files by using regular expressions to remove non-alphanumeric characters, convert punctuation to periods, and to replace all spacing with single spaces.

Clinical documents at BC Cancer have automatically inserted text at the beginning and ends of the documents corresponding to information such as date, patient identifying information, dictating physician, facility, and the care providers to copy the document to. As this information is likely not useful for predictions, and would vary between facilities, we sought to remove this text.

To do this, we employed filters using manually created regular expressions to remove this text. We needed multiple filters as there are different document formats to account for. These regular expressions can be found in the `remove_beginning` and `remove_ending` functions in the `process_text.py`, in the Github repositories described in Section 3.7.

We also replaced punctuation with periods and removed characters that were not punctuation, parentheses, quotations, or alphanumeric characters prior to tokenization,

For our BOW models, we made all text lowercase, and then tokenized the text using the NLTK English SnowballStemmer [14]. For LSTM models and CNN models, we used the default tokenizer from PyTorch's Torchtext library [51]. For BERT and `Longformer` models, we used the default tokenizers from their HuggingFace library implementation [76], after making text lowercase and removing numeric characters.

Table 3.2: Class distribution for our targets representing whether a patient has seen or not seen a psychosocial cancer discipline, within the five years following when the document used by our models was generated.

Discipline	Not Seen (%)	Seen (%)
Psychiatry	43488 (98.2)	799 (1.8)
Counselling	34483 (77.9)	9804 (22.1)

3.2 Targets

In this work, we examined a few different targets representing psychosocial needs. We include clinician-addressed psychosocial needs based on seeing a counsellor or psychiatrist. We also included needs self-reported by patients when completing a questionnaire at the start of their cancer care. The targets and their generation are described in this section; please see Section A.1 for additional details.

3.2.1 Clinician-addressed Psychosocial Needs

We identified targets representing clinician-addressed psychosocial needs, whether patients go on to see a counsellor or psychiatrist. Counsellors at BC Cancer see patients for emotional needs, but also other psychosocial needs such as help with securing housing, and practical needs such as assistance getting to and from appointments. Patients can self-refer themselves to counsellors, or can be referred by clinicians. Patients see psychiatrists for a variety of mental health concerns, including management of preexisting mental illnesses, as well as for new psychiatric symptoms. Patients must be referred to psychiatry by other clinicians, such as counsellors, or members of the treatment team.

We identified these targets based on whether a patient ever goes on to see a counsellor or psychiatrist at BC Cancer, within the first five years after the document used for the predictions was generated. We show the class distribution of these targets in Table 3.2

Please check all of the following items that have been of concern or a problem for you in the past week including today.*

6. **Emotional:** ○ Fears/Worries ○ Sadness ○ Frustration/Anger ○ Changes in appearance ○ Intimacy/Sexuality	7. **Informational:** ○ Understanding my illness/treatment ○ Talking with the health care team ○ Making treatment decisions ○ Knowing about available resources
8. **Practical:** ○ Work/School ○ Finances ○ Getting to & from appointments ○ Accommodation	9. **Spiritual:** ○ Meaning/Purpose of life ○ Faith
10. **Social/Family:** ○ Feeling a burden to others ○ Worry about family/friends ○ Feeling alone	11. **Physical:** ○ Concentration/Memory ○ Sleep ○ Weight

Figure 3.1: Portion of the Canadian Problem Checklist used used by patients to self-report their psychosocial cancer needs.

3.2.2 Self-reported Psychosocial Needs

To generate targets representing the self-reported psychosocial needs of cancer patients, we used scores from the Canadian Problem Checklist portion of the PSSCAN-R, a questionnaire patients at BC Cancer fill out at the start of their treatment [43]. Figure 3.1 shows the relevant portion of the questionnaire which patients fill out. This work looked at predicting whether patients will report a threshold number of *Emotional* and *Informational* needs.

The primary cancer needs we investigated were the *Emotional* needs, as it corresponds directly to various aspects of emotional and psychological health. To determine how performance would change when predicting a different need, we also looked at *Informational* needs, picking this one due to its focus on patient educational needs.

Within these two categories of needs, we sought to predict whether patients had a minimum numbers of needs, as an attempt to quantify severity. We investigated the spectrum of having at least one need to having all four or five emotional needs. Table 3.3 reports how many patients did or did not meet each threshold of having

Table 3.3: Class distribution for our labels representing whether a patient meets a threshold of a minimum number of cancer needs.

Type of Need	Minimum Needed	Does not Meet Threshold (%)	Meets Threshold (%)
Emotional	1	23116 (42.3)	31510 (57.7)
Emotional	2	39090 (71.6)	15536 (28.4)
Emotional	3	46728 (85.5)	7898 (14.5)
Emotional	4	51797 (94.8)	2829 (5.2)
Emotional	5	53672 (98.3)	954 (1.7)
Informational	1	24455 (44.8)	30171 (55.2)
Informational	2	38235 (70.0)	16391 (30.0)
Informational	3	44184 (80.9)	10442 (19.1)
Informational	4	47625 (87.2)	7001 (12.8)

a minimum number of reported cancer needs.

To reduce the number of possible comparisons when evaluating different models, we chose to look at our different models using targets corresponding to having four or more emotional needs, and having four informational needs. We choose these cutoffs to isolate patients with a high number of needs in each category. Around 5% of patients met this threshold for emotional needs, and around 12% for informational needs.

3.2.3 Survival

To compare the prediction of psychosocial cancer needs with a non-psychosocial outcome, we also generated targets based on survival. We calculated a patient's survival starting from the date the document used by the model was generated, at the start of their care at BC Cancer. While cancer literature typically examines survival from diagnosis, we were interested in predicting survival from the document creation. The time between diagnosis and this document being generated varies, such as when patients first have a surgery outside of BC Cancer, and only come to BC Cancer months later for chemotherapy.

We calculated survival by finding the number of months from the document date until a patient passed away, if they did. If not recorded as dying, we instead

Table 3.4: Class distribution for our labels representing whether a subject survived a certain number of months.

Months	Did not Survive (%)	Survived (%)	Total Subjects
6	5472 (13.8)	34188 (86.2)	39660
36	14767 (43.8)	18953 (56.2)	33720
60	17140 (67.4)	8283 (32.6)	25423

calculated their survival being until the last date we know they were still alive, by looking at their `last_contact_date`, `last_attended_appointment` or `PSSCAN_screening_date`. For some subjects and targets, it was unclear whether a patient survived a certain number of months, as they may have had neither a recorded death date or another date indicating they were still alive by the survival cutoff. These patients were excluded from the training and evaluation for these targets. We show the survival class balance and number of subjects included in Table 3.4.

3.3 Specific Subgroups

3.3.1 Stage and Metastatic Status

A patient's psychosocial needs are expected to change based on how advanced their disease is, which can affect a patient's function and risk of mortality [10]. The patients in our dataset have staging data based on the American Joint Committee on Cancer (AJCC)'s staging manual, where cancers are staged from Stage I to IV [7]. Stages are specific to the type of cancer, but generally correspond to being only in one spot (Stage I), in one area but larger (Stage II), spreading locally such as to nearby lymph nodes (Stage III) or spreading distantly (Stage IV, metastatic).

Final staging is made by considering a patient's clinical staging, based on symptoms and imaging scans, and also on their pathological staging, which is based on examining tissue removed from a patient. For this project, we used the final summary staging. Some patients are missing staging data so were excluded. Specific details of this processing can be found in Section A.1

We sought to investigate whether a patient's cancer stage at diagnosis impacts

the performance of our models. To do this, we trained and evaluated our models separately for patients diagnosed with each cancer stage. We also compared our models when using data from patients with metastatic illness at diagnosis, where their cancer has spread to distant parts of the body, to those without metastatic illness. Metastatic illness corresponds to a poorer prognosis and symptom burden.

We evaluated our outcomes using BOW and CNN models with these cancer stage subgroups. We show the number of subjects in each stage in Table 3.1.

3.3.2 Male or Female

We also investigated whether model performance varies by biological sex. We again predicted our different targets using CNNs with male and female subjects. In our dataset, all patients were recorded as either male or female. For the entire dataset, 52.4% of subjects are recorded as being female.

3.4 Natural Language Models

In NLP, *language models* assign probabilities to words based on how they relate to each other [35]. Most directly, they can be used to determine the probability of one sequence of words occurring compared to another. However, by using this understanding of words within a document, language models can be used to classify a document, by extending such models to predict a classification task.

In this work, we compare and evaluate different types of language models, extending them for document classification to predict our binary targets. We test simple BOW models, and the common neural models CNN and LSTM. We also deploy a recent neural model which uses transformers, BERT, and a long-document adaptation, Longformer. This comparison allows us to investigate whether these more advanced models perform better for our task. Generally, more complicated or recent methods in machine learning (ML) and NLP may not always be better suited for a task and the data being used.

3.4.1 Bag-of-Words

BOW is a traditional non-neural language model, and one of the simplest ways to understand text. A collection of text is used to produce a list of words; for

example, the most frequently occurring n words. A piece of text in this collection is then understood as a vector with length n, with each element corresponding to how many times a word was found in the text. These vectors can then be used for classification tasks by being fed directly to machine learning algorithms [14]. Linear classifiers and decision-tree methods are commonly used.

Our use of BOW was meant to be a non-neural comparison against our neural language models. We implemented it using the commonly used L2-regularized logistic regression. We used term frequency — inverse document frequency (TD-IDF) weighting, an adjustment for a word's importance that takes into account both its *term frequency*, how often a word occurs in a document, and the *inverse document frequency*, the inverse of how many documents a word appears in [46]. We used a C value of 0.2, corresponding to the inverse of the lambda regularization constant, and a vector length of 5000 words (features), choosing these hyperparameters empirically as shown in Appendix Section A.1.

3.4.2 Convolutional Neural Networks

CNNs are neural networks originally used for signal and image processing [82], centred upon the use of *convolutions*. Convolutions are groups of features - in image processing adjacent pixels, in NLP adjacent words - considered by the model together. This allows a reduction of the feature space, while still allowing the model to consider how features are grouped together.

We based our model on the CNN model originally developed for sentence classification by Kim [37], and then further adapted by Rios and Kavuluru [58] for document classification. We based our hyper-parameters on these works, as well as the further work by Rios and Kavuluru [59] which used psychiatric documents.

In our model, tokenized words are represented as 300-length, randomly instantiated vectors. Convolutions of 3,4, and 5 tokens (window length) are then used to output single real number values to 500 output channels each. Each output channel is a vector whose length is equal to the number of convolutions, which is based on the number of convolutions needed to cover a document by moving over one word at a time.

These output channels are then connected to a single max-pooled feature vec-

tor, which is then fed to a final softmax layer which predicts the binary outcomes. To prevent over-fitting, and as an alternative to other methods of regularization, the model employs dropout. Elements of the max-pooled layer are randomly set to zero before being fed to the softmax layer, with probabilities drawn from a Bernoulli distribution with $p=0.5$.

3.4.3 Long Short-Term Memory Models

Recurrent neural networks are a type of neural network designed to handle sequences of data, especially sequences that can vary in length such as documents [82]. They utilize a *hidden state*, which is updated with each part of sequential data, in addition to the prior hidden state, learning weights to transform both inputs.

LSTM is an extension of the RNN model which also includes a *memory* cell. A particular hidden state may or may not affect the memory cell as determined by *input and output gates*, which are also learned by the model. An *output gate* then controls what will be used for the next hidden state, either using the current hidden state or the memory cell, allowing the next hidden state to be influenced by recent and further away parts of the sequence.

Our implementation was based upon the model described by Adhikari et al. [3], which uses multiple types of regularization with an LSTM. We again used random word embeddings with a dimension of 300. Our hidden unit has a dimension of 512. As in this previous work, we employed dropout as in our CNN model with p=0.5. We also incorporated embedding dropout, where the model randomly drops entire word embeddings, with a dropout rate of 0.1. Lastly, as in their work, we also used weight dropping, which drops some of the values in the weight matrices, with a dropout rate of 0.2.

3.4.4 Bidirectional Encoder Representations from Transformers

Devlin et al. [22] proposed BERT models to allow deep bidirectional understanding of text, creating a pretrained model that can be fine tuned to accomplish a wide variety of tasks. BERT is based on *transformers*, which, unlike the sequential reading in CNN or RNN models, allow all of the words or tokens in a piece of text to

be considered at once. Transformers accomplish this by using an *encoder* to transform the entire piece of text into a sequence of vectors as numerous as the number of tokens. Fully connected layers then connect these vectors, and can be used to accomplish tasks. The models were then trained on a large corpora to accomplish two tasks: predicting masked words, and whether one sentence follows another. This results in a pretrained model with a sophisticated understanding of language. Different layers can then be connected on top of this base model to accomplish different tasks after some fine-tuning.

In this work, we used the pretrained BERT model `bert-base-uncased`, which was pretrained on general English text. We tried using some BERT models pretrained on more specific clinical data created by Alsentzer et al. [5], but these did not seem to improve performance. We fine-tuned the model with a binary classification head, and used the maximum number of tokens supported by BERT, 512.

3.4.5 Longformer

BERT is limited in only being able to use up to 512 tokens, as the model's memory requirement scales quadriatically, $\mathcal{O}(n^2)$, with the number of tokens. While this may be enough to cover many pieces of text such as sentences or tweets, almost all of our documents are larger than this, as we show in Figure 3.2. Only 7.07% of the documents in our psychosocial outcome training dataset had 512 or fewer tokens when tokenized by BERT, with the mean and median number of tokens being 988.49 and 945 respectively.

Beltagy et al. [12] proposed the Longformer model as an extension of BERT that can handle larger documents, up to 4092 tokens. This model's memory requirements scale linearly instead of quadriatically. Instead of full attention between all tokens, Longformer uses three more selective attention mechanisms. A *sliding window* has a token attend to a set number of tokens on either side of it. A *dilated sliding window* allows attention between a token and others farther away. *Global attention* is also used, but only for a set number of tokens, which are allowed to attend to all other tokens and vice versa.

As for BERT, we used a model pretrained on a general English corpus:

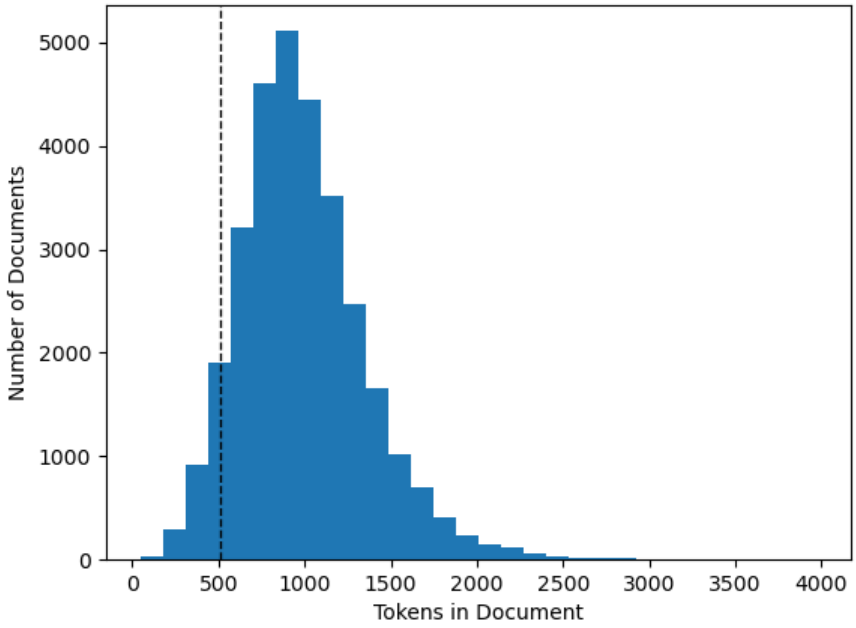


Figure 3.2: Histogram of the number of tokens in the documents used to train our models to predict whether a patient will see a psychiatrist in the five years after document generation, after being tokenized by the bidirectional encoder representations from transformers (BERT) Tokenizer. The dotted line shows the limit of BERT, at 512 tokens.

`longformer-base-4096`. We used the default attention window size, 512. To utilize this model fully, we used it with up to the maximum 4096 tokens per document as supported by this model.

3.5 Training and Evaluation

3.5.1 Training, Validation, and Test Sets

When training machine learning models for prediction, it is important to consider that models may learn to fit training data well, but not be able to predict new data successfully, a phenomenon called *overfitting* [64]. The potential for overfitting

worsens as models become more complex, such as when using the neural methods employed in this project.

To avoid overfitting, we separated both hold-out validation and test sets. We included 70% of subjects in the training set, used to train our models. We made a hold-out validation set with 10% of subjects, and a hold-out test set with 20% of subjects. We trained models using the training set, and evaluated them on the validation set while working on this project, to understand and guide development. Results in this work represent the best performance on the validation set. Evaluating these models on the test set will provide the most accurate understanding of how the models would perform on novel data. However, for the purpose of this thesis, we have not yet evaluated on the test set, and will be reporting these results in forthcoming publications, after some additional work such as further tuning.

To distribute patients randomly between the sets, we assigned sets based on the last digit of their study ID. The validation set included IDs ending in the number two, while the test set has those ending in five or eight. The remaining subjects were included in the training set. We confirmed with BC Cancer Data Access that these IDs are randomly generated and not biased.

3.5.2 Implementing and Training our Models

We conducted this project on a virtual installation of Windows Server 2012 R2, with an eight processor Intel Xeon 8160 CPU, and 16 GB of RAM. We had access to a shared GPU through a NVIDIA GRID V100D-16Q, with 16 GB of VRAM allocated to our virtualisation. We ran BOW models on the CPU, and all other models on the virtual GPU.

For all models, we used a learning rate of 0.00001, except for Longformer where we used 0.0001. For our BOW implementation, we deployed and fit our model using scikit-learn [52], using their `LogisticRegression` model and its default functions to fit and predict.

For our neural models, we allowed up to 100 epochs, stopping training after five epochs of no increase in balanced accuracy when tested on the validation set. We used binary cross-entropy loss for all of these models, along with an Adam optimizer [38].

For our CNN and LSTM models, we implemented them as PyTorch `nn` modules, manually writing code to implement forward and back-propagation through batches and epochs. We re-used some code from hedwig Github repository[1], though modifications were required due to recent PyTorch and other updates

For BERT and Longformers, we used their respective classes from HuggingFace's transformer library [76], the `BertModel` and `LongformerModel` classes respectively. We utilized PyTorch Lightning [25] to handle some of the boilerplate code for training and evaluating.

3.5.3 Dealing with Class Imbalance

Many of the targets we examined are imbalanced, as shown in Tables 3.2, 3.3, and 3.4. We explored two options to deal with this class imbalance, *undersampling* and *loss weighting*.

For our undersampling technique, we randomly selected the training dataset to have at most twice as much of one class as another. For our loss weighting, we adjusted our binary cross entropy loss, multiplying the loss by a factor so that the product of the factor and a class's proportion would be the same between both classes.

As shown in Table A.2, weighted loss usually outperformed undersampling, so we generally used this throughout this work, except for when comparing the impact of token length on prediction. This helped cut down on the time required to run the Longformer models, which could take weeks when using the full training dataset and weighted loss.

3.5.4 Evaluation Metrics

We chose a variety of evaluation metrics to allow comparison of our results with prior work in both computer science and medicine, including metrics that adjust for class imbalance. We define these metrics, besides AUC in Table 3.5. Throughout this work, we display these metrics for all results, except for positive predictive value (PPV), negative predictive value (NPV), and specificity, which we recorded separately. AUC represents the area under the receiver-operator curve, which is a

[1] https://github.com/castorini/hedwig

Table 3.5: Evaluation metrics used in this work, expressed in terms of the four components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). We also list equivalent definitions of Balanced Accuracy and F1 based on other metrics.

Metric	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall/Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Specificity	$\frac{TN}{TN+FP}$
Positive Predictive Value	$\frac{TP}{TP+FP}$
Negative Predictive Value	$\frac{TN}{TN+FN}$
Balanced Accuracy	$\frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2}$
F1	$\frac{2TP}{2TP+FP+FN}$
Balanced Accuracy	$\frac{Sensitivity+Specificity}{2}$
F1	$2 * \frac{Precision*Recall}{Precision+Recall}$

plot of sensitivity against one minus specificity, thereby providing a metric taking into account both of these components.

3.6 Interpreting our Models

3.6.1 Interpreting Bag-of-word Models

To understand our BOW models, we examined *feature importance* using scikit-learn's implementation. This corresponds to the absolute value coefficient size of the L2-regularized logistic regression models. We extracted the top ten tokens from

our BOW models by this weight to better understand the models, as a representation of what tokens are most important for either positive or negative predictions.

3.6.2 Interpreting Neural Models

Given that neural models have often been described as "black boxes" that are difficult to interpret, we sought to better understand our models using the Captum interpretability library for PyTorch [39]. Specifically, we used the library to implement *integrated gradients* [69].

This is an attribution method that allows visualization of input features, in our case text, to understand what features are contributing to a prediction. Integrated gradients accomplish this by generating a linear interpolation between an empty input text and the actual text, and then calculating an averaged gradient representing how much each token impacts the prediction throughout the interpolation. This allows a straightforward understanding of how a token in a document is impacting its prediction when used by a trained model. However, this calculated importance is specific to a token's surrounding text, so a single token may have different importance depending on its context and the document it is in.

3.7 Code Availability

Due to the size of this project and it being part of larger endeavours to use and analyze this dataset, the code for this project is found in three separate Github repositories: initial subject selection and document processing[2], preparation of the datasets directly used in this project[3], and the actual code used for training and evaluating our models[4].

[2] https://github.com/jjnunez11/scar
[3] https://github.com/jjnunez11/scar_nlp_data
[4] https://github.com/jjnunez11/scar_nlp

Chapter 4

Results

In Chapter 4 we report the results of training our models to predict psychosocial needs and then evaluating them on our hold-out validation set, deferring evaluation on our test set for now. After describing these results, we show the results of interpreting the models to understand how they are arriving at their predictions.

We start in Section 4.1 with a summary of how the models compared against each other when predicting the various targets. We then describe the results of investigating our primary hypotheses. We report the results of predicting our clinician-addressed psychosocial need targets, whether subjects will see a psychiatrist or counsellor, in Section 4.2. The self-reported psychosocial targets, emotional or informational needs, then follow in Section 4.3. We then compare this to predicting a non-psychosocial target, survival, in Section 4.4.

We then describe the results pertaining to our secondary hypotheses. We look at the performance of our models when predicting outcomes within subgroups based on biological sex and by cancer stage, in Section 4.5. As BERT models are limited in the number of a document's tokens (words) they can use, we then investigate the impact of token limits in Section 4.6. We do this by using Longformer models, which can use more tokens, alongside BERT, with different token inclusion limits.

Finally, in Section 4.7, we interpret our models, examining co-efficient weighting for our BOW models, and using Integrated Gradients to further understand our neural models.

Table 4.1: Summary of the best performing model on three performance metrics for the outcomes examined in this work. Here we compare Bag-of-Words (BOW), Convolutional Neural Networks (CNN), Long-term Short-term Memory (LSTM) and bidirectional encoder representations from transformers (BERT) models, across balanced accuracy, receiver operator-curve area-under-curve (AUC), and F1 metrics. Seeing a counsellor or psychiatrist, and survival, are based on the five years following when the document used by the models was generated. Two models indicates a tie at two decimal places.

Target	Balanced Accuracy	AUC	F1
Seeing a psychiatrist	BOW	LSTM	CNN
Seeing a counsellor	CNN/LSTM	CNN/LSTM	CNN/LSTM
Four or more emotional needs	BOW/BERT	BERT	BoW
Four informational needs	CNN	CNN	CNN
Surviving five years	CNN	CNN	CNN

4.1 Comparing Models

When comparing BOW, LSTM, CNN and BERT models across our outcomes, we see that CNN and LSTM models were often the best performing, though BOW and BERT were sometimes most performant. We show a summary of these results in Table 4.1, and further details for the specific metrics in the following sections, in Tables 4.2, 4.3, 4.4, 4.5, and 4.6. We also found that a variation of BERT which can use more tokens has higher performance than standard BERT, as shown in Section 4.6. We note that, across models, performance was higher when predicting our non-psychosocial outcome, survival (highest balanced accuracy 86%), than any of our psychosocial outcomes; the highest balanced accuracies were in the mid eighties for survival, but low sixties to low seventies for the psychosocial outcomes.

Table 4.2: Model performance when predicting whether a patient will see a psychiatrist within the first five years after their cancer diagnosis. We compare bag-of-words (BOW), convolutional neural network (CNN), long-short term memory (LSTM) and bidirectional encoder representations from transformers (BERT) models. We report various performance metrics including receiver-operator-curve area-under-curve (AUC).

Model	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	0.80	**0.73**	0.73	0.11	0.06	**0.66**
CNN	**0.87**	0.68	0.75	**0.12**	**0.07**	0.49
LSTM	0.80	0.72	**0.77**	0.10	0.06	0.63
BERT	0.73	0.62	0.63	0.07	0.03	0.51

4.2 Predicting Clinician-addressed Psychosocial Needs

4.2.1 Seeing a Psychiatrist

We find that LSTM and BOW models generally had the highest performance when predicting whether a subject will see a psychiatrist within five years, a task with more class imbalance than our other outcomes. As shown in Table 4.2, performance was generally higher than when predicting the self-reported emotional or informational needs, with our LSTM model achieving a balanced accuracy of 0.72 and AUC of 0.77.

4.2.2 Seeing a Counsellor

As another outcome related to a cancer patient's clinician-addressed psychosocial needs, we predicted whether patients will see a counsellor within the five years after generation of the document used by our models. Comparing the four models in Table 4.3, we find CNN and LSTM models performed best, and that performance for predicting this outcome was higher in some metrics, such as AUC and balanced accuracy, than when predicting self-reported emotional or informational needs. The three neural models all achieved AUC over 0.70, though the balanced accuracies fell short of this threshold.

Table 4.3: Model performance when predicting whether a patient will see a counsellor within the first five years after their cancer diagnosis. We compare bag-of-words (BOW), convolutional neural network (CNN), long-short term memory (LSTM) and bidirectional encoder representations from transformers (BERT) models. We report various performance metrics including area-under-curve (AUC).

Model	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	0.66	0.66	0.66	0.48	0.37	0.67
CNN	**0.70**	**0.68**	**0.75**	**0.49**	**0.40**	0.65
LSTM	0.68	**0.68**	**0.75**	**0.49**	0.39	**0.68**
BERT	0.65	0.65	0.70	0.46	0.35	0.66

4.3 Predicting Self-reported Psychosocial Needs

4.3.1 Emotional Needs

We predicted whether patients self-reported a minimum threshold of emotional cancer needs when filling out a questionnaire at the start of their cancer care. We evaluated BOW and CNN models on all of the possible cutoffs, as the latter was generally best performing, and the former is a non-neural method. In Table A.3 we see that having a cutoff at neither extreme tends to lead to the best performance. The CNN models generally outperformed BOW across the evaluated metrics.

Only around 10% of patients have four or more emotional needs. We evaluated this cutoff across four models, as shown in Table 4.4. For this number of needs, BOW and BERT models generally performed the best, though CNN models are somewhat similar. Performance was numerically lower than when predicting clinician-addressed needs, with only our BERT model having an AUC above 0.70, and no models achieving a balanced accuracy above this level.

4.3.2 Informational Needs

We also looked at the prediction of informational needs, as a psychosocial need more related to needing educational support than psychological help. We com-

Table 4.4: Model performance when predicting whether a patient self-reports four or five emotional needs at the start of their cancer care. We compare bag-of-words (BOW), convolutional neural network (CNN), long-short term memory (LSTM) and bidirectional encoder representations from transformers (BERT) models. We report various performance metrics including receiver-operator-curve area-under-curve (AUC).

Model	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	**0.69**	**0.66**	0.66	**0.18**	**0.10**	0.63
CNN	**0.69**	0.63	0.67	0.16	0.09	0.57
LSTM	0.47	0.58	0.63	0.12	0.07	**0.70**
BERT	0.64	**0.66**	**0.71**	0.17	0.09	0.68

Table 4.5: Model performance when predicting whether a patient self-reports four informational needs at the start of their cancer care. We compare bag-of-words (BOW), convolutional neural network (CNN), long-short term memory (LSTM) and bidirectional encoder representations from transformers (BERT) models. We report various performance metrics including receiver-operator-curve area-under-curve (AUC).

Model	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	**0.61**	0.60	0.60	0.27	**0.18**	0.59
CNN	0.59	**0.61**	**0.65**	**0.28**	**0.18**	0.64
LSTM	0.44	0.57	0.61	0.25	0.15	**0.75**
BERT	0.60	0.59	0.62	0.27	0.17	0.59

pared the models using a cutoff of having all four informational needs asked, which around 12% of our subjects reported. As shown in Table 4.5, we see that CNN models performed best, though generally the prediction of these needs seemed harder than the emotional needs. None of the models achieved either balanced accuracies or AUC above 0.65.

Again using BOW and CNN models to examine performance on different need cutoffs, we see our metrics are similar between the cutoffs, with a slight edge for CNN on some metrics (Appendix Table A.4).

Table 4.6: Model performance when predicting whether a patient will survive for five years after the document used for training models was generated. We show results of Bag-of-Words (BOW), convolutional neural networks (CNN), long-short term Memory (LSTM) and bidirectional encoder representations from transformers (BERT) models. We report various performance metrics including receiver-operator-curve area-under-curve (AUC)

Model	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	0.84	0.83	0.83	0.77	**0.72**	0.82
CNN	**0.85**	**0.86**	**0.92**	**0.79**	0.71	**0.89**
LSTM	0.84	0.85	0.91	0.78	0.71	0.86
BERT	0.84	0.84	0.91	0.77	**0.72**	0.84

4.4 Predicting Survival, a Non-psychosocial Outcome

In order to compare the performance of predicting psychosocial outcomes with a non-psychosocial target, we trained our models to predict survival, an important outcome investigated in prior work. As five-year survival is the common duration-of-interest, we compare the four models when predicting it in Table 4.6. Of note, our target represents whether a subject survived five years from the document used by the model being generated, instead of from their cancer diagnosis date as usually used in the literature. We found that CNN models offer the highest performance, and that the metrics are higher for this outcome than the psychosocial ones. For this target, the balanced accuracies reached the mid-eighties and AUC the low-nineties.

As different lengths of survival can be of interest depending on the clinical situation, we again used BOW and CNN models to compare the prediction of different survival lengths: six months, thirty-six months, and sixty months. We show these results in Appendix Table A.5. The CNN models have slightly higher performance when predicting the longer survival duration, though metrics are similar for the BOW models.

4.5 Predicting with Subgroups

4.5.1 Fewer Subjects

As we are interested in examining model performance for subgroups of our dataset, such groupings based on biological sex or cancer stage, we first examined how performance changes when using fewer subjects (Appendix Table A.6). Predicting whether a patient will see a psychiatrist within 5 years, we saw an increase for most metrics as the number of subjects is increased, all the way to the maximum number available in our dataset, 30953.

4.5.2 Metastatic Status and Cancer Stage

Given that the stage of a patient's cancer can affect their prognosis and cancer needs, we sought to investigate how our models would perform when training and evaluating on only subjects diagnosed with a certain stage of cancer. In Table 4.7, we show the results of training our models to predict seeing a psychiatrist within five years, with subgroups based on AJCC staging. As not all subjects in our dataset had intact staging data, we show the numbers of subject in each stage, and also include a group of subjects with non-metastatic disease, Stage I, II, and III.

We see a mixed set of results; both models seem to have predicted better for patients with Stage II disease, and the prediction seems more difficult for patients with Stage IV illness. Notably, the metrics were generally higher when predicting with 5535 Stage II patients, than with 6000 randomly selected patients. Of note, the Stage IV prediction had only a small number of patients who saw psychiatry; only 54/3636 in the training set, and 6/552 in the validation set.

Given the class imbalance of predicting whether a patient will see psychiatry, we also examined predicting whether patients will have four or more emotional needs, to see whether performance would again be higher when only including patients diagnosed with Stage II illness. In Table 4.8, we again see improved performance when predicting with only Stage II patients.

Table 4.7: Model performance when predicting whether a patient will see psychiatry in the five years following document generation at the start of their cancer care. Here we compare the results of Bag-of-Words (BOW) and Convolutional Neural Networks (CNN) by subgroups of patients based on their cancer stage at diagnosis. Not all subjects in our dataset had staging data, and we show the number of subjects used to train each model (n). AUC: receiver-operator-curve area-under-curve.

Model	Stage	n	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	I	4078	0.88	0.66	0.66	0.07	0.04	0.43
BOW	II	5535	0.72	0.76	0.76	0.10	0.05	0.80
BOW	III	3809	0.77	0.69	0.69	0.09	0.05	0.60
BOW	I+II+III	13422	0.79	0.76	0.76	0.10	0.06	0.72
BOW	IV	3638	0.92	0.55	0.55	0.05	0.03	0.17
CNN	I	4078	0.78	**0.89**	**0.92**	0.09	0.05	**1.00**
CNN	II	5535	0.57	0.75	0.85	0.08	0.04	0.93
CNN	III	3809	0.91	0.66	0.66	**0.14**	**0.08**	0.40
CNN	I+II+III	13422	0.71	0.67	0.77	0.07	0.04	0.62
CNN	IV	3638	**0.99**	0.50	0.46	0.0	0.0	0.0

4.5.3 Biological Sex

Given the differences between biological sexes in mental health and associated care patterns, we sought to determine whether prediction performance would be different between the two groups. As shown in Table 4.9, performance appears to be higher when predicting seeing a psychiatrist for a subset of patients identified as having female sex. There is a difference in the rate at which the sexes see psychiatry, with almost twice as many females (2.28%) seeing the discipline than males (1.28%). All subjects in our dataset were documented as having either male or female sex.

4.6 Impact of Token Limits for Transformer Models

BERT models are limited to using 512 tokens, and our documents are often much larger than this. We used Longformer models to determine if a similar transformer

Table 4.8: Model performance when predicting whether a patient self-reports four or five emotional needs at the start of their cancer care. Here we compare the results of Bag-of-Words (BOW) and Convolutional Neural Networks (CNN) by subgroups of patients based on their cancer stage at diagnosis. Not all subjects in our dataset had staging data, and we show the number of subjects used to train each model (n). AUC: receiver-operator-curve area-under-curve.

Model	Stage	n	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	I	4078	0.77	0.56	0.56	0.10	0.06	0.32
BOW	II	5535	0.63	**0.71**	0.71	**0.18**	**0.10**	0.81
BOW	III	3809	0.64	0.57	0.57	0.12	0.07	0.48
BOW	I+II+III	13422	0.7	0.63	0.63	0.15	0.09	0.56
BOW	IV	3638	0.65	0.54	0.54	0.1	0.06	0.42
CNN	I	4078	0.62	0.54	0.57	0.09	0.05	0.44
CNN	II	5535	0.58	0.70	**0.73**	0.16	0.09	**0.84**
CNN	III	3809	**0.88**	0.55	0.55	0.13	**0.10**	0.19
CNN	I+II+III	13422	0.69	0.58	0.63	0.12	0.07	0.46
CNN	IV	3638	0.74	0.57	0.55	0.12	0.07	0.38

Table 4.9: Model performance when predicting whether a patient will see psychiatry in the five years following document generation at the start of their cancer care. Here we compare the results of Bag-of-Words (BOW) and Convolutional Neural Networks (CNN) on those who identified as males vs female. All subjects in our dataset were recorded as having one of these two sexes. AUC: receiver-operator-curve area-under-curve.

Sex	Model	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
Female	BOW	0.81	**0.69**	0.69	0.12	0.07	**0.57**
Female	CNN	**0.92**	0.64	**0.74**	**0.16**	**0.10**	0.36
Male	BOW	0.85	0.67	0.67	0.08	0.04	0.48
Male	CNN	0.73	0.61	0.61	0.05	0.02	0.48

Table 4.10: Model performance when predicting whether a patient will see psychiatry in the five years following document generation at the start of their cancer care. Here we compare the impact of token lengths, when using bidirectional encoder representations from transformers (BERT), and a variation that can use more tokens, Longformer. AUC: receiver-operator-curve area-under-curve.

Model	Max Tokens	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BERT	256	0.17	0.51	0.51	0.04	0.02	**0.87**
BERT	512	0.64	0.63	0.63	0.06	0.03	0.61
Longformer	256	**0.80**	0.62	0.66	**0.07**	**0.04**	0.43
Longformer	512	0.77	0.61	0.66	**0.07**	**0.04**	0.45
Longformer	1024	0.67	0.63	0.63	0.06	0.03	0.60
Longformer	2048	0.75	**0.64**	**0.71**	**0.07**	**0.04**	0.52
Longformer	4096	0.72	**0.64**	0.69	**0.07**	**0.04**	0.56

model would be able to achieve higher performance if able to use more of the document. In Table 4.10 we see that Balanced Accuracy and AUC increased when BERT can use 512 tokens instead of 256 tokens, and as Longformer is able to use up to 2048 tokens. Longformer is able to achieve an AUC 0.08 higher than BERT when using 2048 tokens. Using a limit of 4096 tokens did not improve performance.

To determine whether this benefit was isolated to predicting whether patients would see a psychiatrist, we also investigated the impact of the number of tokens used when predicting five year survival, as we show in Table 4.11. Using more tokens increased the performance by only a small amount. The performance of our Longformer model only increased by 0.01 when increasing the number of tokens from 256 to 2048. It then fell slightly when using a maximum of 4096 tokens.

4.7 Interpreting our Models

In this section, we show the results of interpreting our models. As we discuss the layout and ordering of the sections of medical consultation documents, in Table 4.12 we show the headings of the document used to interpret our neural models in Subsection 4.7.2. These headings are typical, though "Gynecologic History"

Table 4.11: Model performance when predicting whether a patient will see psychiatry in the five years following document generation at the start of their cancer care, when models can use a varying maximum number of tokens. We compare using bidirectional encoder representations from transformers (BERT), and a variation that can use more tokens, Longformer. AUC: receiver-operator-curve area-under-curve.

Model	Max Tokens	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BERT	256	0.84	0.82	0.90	0.76	0.76	0.76
BERT	512	0.84	0.83	0.90	0.76	0.72	0.81
Longformer	256	0.82	0.83	0.91	0.75	0.68	0.83
Longformer	512	0.83	0.83	0.91	0.76	0.70	0.83
Longformer	1024	0.85	**0.84**	**0.92**	0.77	0.75	0.79
Longformer	2048	**0.86**	**0.84**	**0.92**	**0.78**	**0.78**	0.78
Longformer	4096	0.84	**0.84**	0.91	0.77	0.70	**0.86**

would be omitted or replaced by other specific histories depending on cancer type, and "Impression and Plan" are commonly separated into two separate headings.

4.7.1 Bag-of-Words

To better understand the performance of our BOW models, we examined the *feature importance* of the trained models. As we employed L2-regularized logistic regression models, this feature importance is based on the absolute value of the coefficients, which can correspond to importance for either the positive or negative prediction. We show the top ten tokens based on this metric in Table 4.13. These tokens are words that have been *stemmed* to remove endings. Due to the possibility of correlated variables affecting coefficient weights, especially given that we did not use any L1-regularization, we must use caution when interpreting these rankings. However, we see some expected results, such as the stems *depress* (depression, depressive), *counsel* (counselling, counsellor), and *anxieti* (anxiety, anxieties) having large weights for many of the psychosocial targets. Types of cancer and *palliat* (palliative, palliation) are important for the survival target.

Table 4.12: Headings in the document we used to interpret our neural models, as a representation of typical headings in oncology consultation documents. In other documents, "Impression and Plan" may be split into two separate sections, and "Gynecologic History" would be replaced depending on the patient's type of cancer.

Section
Reason for Referral
History of Present Illness
Gynecologic History
Past Medical History
Past Surgical History
Medications
Allergies
Social History
Family History
Physical Examination
Impression and Plan

4.7.2 Neural Models

To better understand our neural models, we show the results of using integrated gradients to interpret our models when applied to a piece of text using the Captum library [39]. To balance privacy considerations with wanting to show the results on real data, we show only relevant excerpts of applying integrated gradients to an entire document which had identifying information anonymized; this anonymization did not change these results.

The document used was from a patient who did see a psychiatrist within five years, and who also survived five years. To keep our results concise, we show the interpretation of CNN and BERT models, a pair that differs in their use of transformers and ability to use the whole text. We compare trained models used to predict a psychosocial outcome, seeing a psychiatrist, and contrast this with surviving five years, a non-psychosocial outcome.

Table 4.13: Top ten features by feature importance of our bag-of-word (BOW) model when used to predict the stated outcomes. Survival, psychiatry, and counselling are all outcomes in the five years following the document used by the models being generated at the start of the subject's cancer care. Our BOW model is using L2-regularized logistic regression, so feature importance is especially impacted by correlation between features.

	Emotional Needs	Informational Needs	Seeing Psychiatry	Seeing Counselling	Survival
1	depress	2011	depress	depress	palliat
2	2011	2011.	anxieti	counsel	2015
3	counsel	radiat	radiat	princ	breast
4	anxieti	skin	prostat	georg	risk
5	no	no	daughter	prostat	lung
6	pound	tamoxifen	also	anxieti	2016
7	pain	basal	mri	any	lymphoma
8	colon	also	treatment	as	no
9	retir	2015	x	no	servic
10	anxious	md	counsel	retir	.

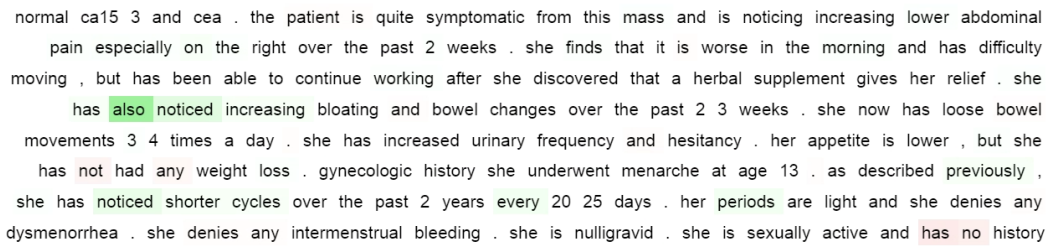
Interpreting Models Predicting Seeing a Psychiatrist

For our example text, our CNN model correctly predicted that the patient would go on to see psychiatry, but the BERT model did not. We show some relevant excerpts in Figure 4.1. In this figure, we show how the models assign importance to words in a section of the document describing symptoms. In Figure 4.1 (a), we see our CNN model finds "also noticed" and another "noticed" to be predictive, but not an initial "noticing". This pattern may correspond to predicting a patient more likely to see psychiatry when endorsing many symptoms, as we will discuss further in Chapter 5. On the other hand, our BERT model finds only light importance for a few words, except for a stronger negative importance for "bleeding". Our CNN model attributes importance to more words in a section of the document describing social and family history, as in Fig 4.1 (c). It found the mention of grandparents having cancer to be important for the prediction, as well as "breast", and a "currently" that is preceding mention of being in a relationship. There is some lighter negative

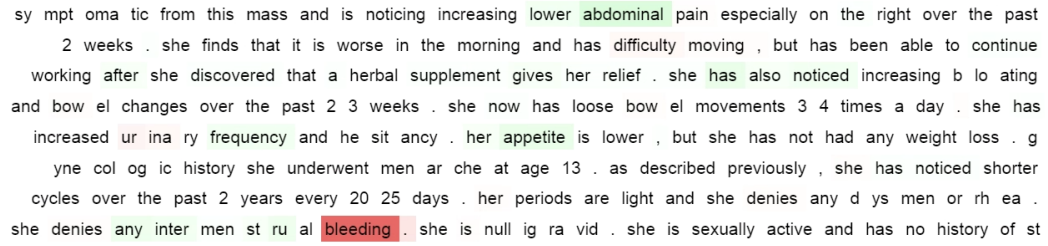
importance assigned to "nonsmoker". Due to BERT's limit of 512 tokens, it is not able to use this section, or any of the document following gynecologic history, including missing medications, family and social history, and the impression and plan.

Interpreting Models Predicting Survival

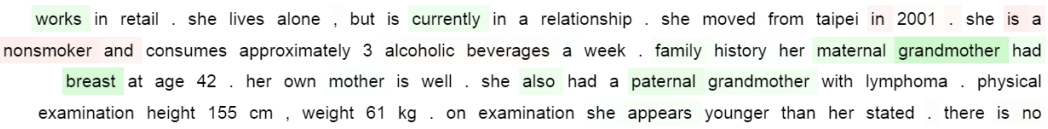
To compare how the models are working when predicting a non-psychosocial outcome, we show the importance of words when predicting five-year survival in Figure 4.2, again with our CNN and BERT models. We show an excerpt that seemed to have most of the important words in both models, from the beginning of the document where the physician describes the patient's cancer and investigations to date. Both models correctly predict that the patient will survive five years. In the CNN model, we see "ultrasound", "biopsy was negative" and "marrow" as important words or phrases that were important in predicting the patient's survival, while "multiple" preceding "peripheral solid nodules" was important in predicting not surviving. There are some similarities with the BERT interpretation, with "ultrasound" again important in predicting survival, but "repeat", "revealing", "marrow", and "peripheral" all shaded to indicate importance in predicting the patient not surviving.



(a) Excerpt of a clinical document describing symptoms with CNN word importance



(b) Excerpt of a clinical document describing symptoms with BERT word importance



(c) Excerpt of a clinical document describing family and social history with the CNN word importance

Figure 4.1: Interpretation of our convolutional neural network (CNN) and bidirectional encoder representations from transformers (BERT) models, showing the importance of words when predicting whether a patient will see a psychiatrist, according to interpretation with integrated gradients. We show excerpts from a clinical document that has been anonymized for presentation here, with any potentially identifying words, names, dates or numbers changed. We show only some relevant excerpts, showing a portion describing symptoms in (a) and (b). We also show a segment describing social and family history in (c) which is only used by the CNN model due to BERT's token limits. The darker the green highlighting, the more predictive of seeing a psychiatrist, while red similarly corresponds to not seeing a psychiatrist. A red-green colour-blind viewable version is available in Appendix Figure A.1

pelvic ultrasound performed on march 18 , 2014 , revealing complex left adnexal lesion with a cystic component measuring 8 . 3 cm and a solid irregular marrow component measuring 3 . 1 cm . the uterus and right ovary appeared normal . she went on to see dr . muir in late april and an endometrial biopsy was negative for malignancy . she had repeat ultrasound on april 29 , 2014 , which revealed that the left adnexal lesion was enlarging , now measuring 10 . 2 x 10 . 1 x 9 . 6 cm with multiple peripheral solid nodules with the largest

(a) Excerpt of a clinical document describing a patient's cancer with CNN word importance

canada , she had a repeat pe l vic ultrasound performed on april 14 , 2012 , revealing complex right ad ne xa l les ion with a cy stic component measuring 8 . 5 cm and a solid irregular marrow component measuring 3 . 3 cm . the ut erus and left o vary appeared normal . she went on to see dr . cohen in late april and an end ome tri al bio psy was negative for mali gnan cy . she had repeat ultrasound on may 22 , 2013 , which revealed that the right ad ne xa l les ion was en lar ging , now measuring 10 . 1 x 10 . 0 x 9 . 3 cm with multiple peripheral

(b) Excerpt of a clinical document describing a patient's cancer with BERT word importance

Figure 4.2: Interpretation of our convolutional neural network (CNN) and bidirectional encoder representations from transformers (BERT) models, showing the importance of words when predicting whether a patient will survive for five years, according to interpretation with integrated gradients. We show excerpts from a clinical document that has been anonymized for presentation here, with any potentially identifying words, names, dates or numbers changed. We show an excerpt describing a patient's cancer. The darker the green highlighting, the more predictive of surviving five years, while red similarly corresponds to not surviving. A red-green colour-blind viewable version is available in Appendix Figure A.2

Chapter 5

Discussion

In this Chapter we discuss the implications of this work. We start by returning to our primary hypotheses in Section 5.1, discussing the general results of predicting psychosocial needs and survival. We consider our results relative to our secondary hypotheses in Section 5.2, interpreting the performance of our models when using subgroups based on sex and cancer staging, and when we vary the number of tokens our models can use. We then discuss broader implications of this work, first from a clinical perspective in Section 5.3, and then from a computational perspective in Section 5.4. We end this chapter with Section 5.5, where we discuss limitations of this work and how they may be addressed, and then Section 5.6, how future work may build upon this project.

5.1 Primary Hypotheses

1. Neural NLP models can be used with a patient's initial oncologist consultation document to predict **clinician-addressed psychosocial cancer needs**, such as whether a patient will see a counsellor or psychiatrist. Specifically, they can predict these outcomes with balanced accuracy and AUC numerically at least 0.70, a performance threshold often met in other psychosocial predictions.

Our results partially support our first primary hypothesis. Two of our models, LSTM and BOW, were able to predict whether subjects will see a psychiatrist

within five years of document generation with balanced accuracy and AUC over 0.70, and our CNN model's AUC also surpassed this threshold. While this level of performance may not reach the 0.90 thresholds often thought to be needed for clinical applications, it is comparable to the performance others have found in other psychiatric or psychosocial work predicting outcomes in the future, such as using clinical data to predict antidepressant response [50]. As such, we find these results consistent with a successful first attempt at this predictive task, laying a foundation for further work to improve performance.

Interestingly, the performance of our models predicting if patients will see a counsellor in this time frame was numerically worse, and short of the 0.70 threshold for all models for balanced accuracy, though AUC did surpass this level for the three neural models. This was an unexpected result, as this target is less class imbalanced than whether patients will see psychiatry. There may be different reasons why this predictive task seemed to be more difficult. Patients see counsellors at BC Cancer for a wide variety of psychosocial reasons, including psychological support, but also for more functional help such as securing housing or obtaining assistance with transportation to receive care. On the other hand, psychiatrists see patients for a narrower scope, centred upon treating psychiatric illness. It may be more difficult for models to learn a wider range of reasons someone would be referred. Another consideration is that patients must be referred to psychiatry by a clinician. Clinicians may therefore document details pertaining to a psychiatry referral more reliably than when considering a counselling referral, which patients can self-refer to.

Prior work has often discussed the need for interpretability when applying artificial intelligence (AI) techniques to clinical domains [2]. While interpretation methods have limitations, we attempted to "look within the black box" in this work. Interpreting our BOW model by looking at coefficient weights revealed that some expected words (technically, their stems) had high importance. For instance, `depress` and `anxiety` were the top two words for predicting whether a patient will see a psychiatrist, and `depress` and `counsel` for seeing a counsellor. Other words with large absolute value co-coefficients for this task include `princ` and `georg`, which may correspond to different levels of counselling availability at the Prince George BC Cancer site, or the different rates of psychosocial needs in

this more rural population. Stems corresponding to types of cancer and treatment also seemed important for whether a patient will see psychiatry.

Interpreting the neural models with integrated gradients was less straightforward, but still interesting. In one demonstrative document, the model seemed to find that words describing a patient's family history of cancer was predictive of seeing psychiatry. This may suggest that a family history of cancer may make a patient more likely to need support from a psychiatrist due to reactivation of intergenerational trauma from facing this illness. Perhaps a patient lost a loved family member from a difficult journey with cancer at a young age, and now worries they will similarly have a challenging experience. A less interesting explanation could be that the oncologists who who do not spend time taking a comprehensive family history are less likely to refer a patient to psychiatry due to asking a narrow set of questions, a possible explanation of lack of referral to psychosocial supports documented in prior work [49, 70]. Providers who do ask about family history may also be more in tune with holistic cancer needs.

Our interpretation of the neural models may suggest that our models are able to pick up on subtle nuances in the language used by the treating oncologists. For example, in our demonstrative document, when predicting which patients will see psychiatry, our model assigned importance to "also noticed" and a later "noticed", but not an initial "notices". These three uses of the verb "to notice" preceded symptoms the patient was experiencing. The model may have picked up that when an oncologist uses the verb "to notice" multiple times, it corresponds to a patient that is endorsing many symptoms. Such patients may be experiencing anxiety, which can increase awareness of bodily sensations, and so may be more likely to see psychiatry. This could be an example of how neural models may be able to pick up on language features whose importance for prediction may not be particularly obvious.

2. Neural NLP models can be used with a patient's initial oncologist consultation document to predict a patient's **self-reported emotional and informational cancer needs** around the time of this document being generated, also at the performance threshold of balanced accuracy and AUC being above 0.70.

Unlike the results of predicting clinician-addressed needs, our results for predicting self-reported psychosocial cancer needs did not support our hypothesis. For both emotional and informational needs, the models' performance did not surpass the 0.70 threshold for either balanced accuracy or AUC, besides our BERT model whose AUC was 0.71 when predicting a patient reporting at least four emotional needs.

One could expect that this task would have been easier than the clinician-addressed needs given the temporal relationship; our models predicted whether a patient would see a psychiatrist or counsellor within five years of the document being generated, while these needs were self-reported by the patient within five days of the document being generated. However, it may simply be that these needs are more varied and less connected to the content documented by the treating oncologist. These needs are entirely self-reported, while the clinician-addressed needs are always (psychiatry) or sometimes (counselling) referred to by clinicians. As such, the treating oncologist may have less reason to include details around these needs. Our worse results for predicting informational needs may suggest these needs may vary greatly patient-to-patient with respect to the information recorded in these documents. This is somewhat surprising; one could expect that informational needs have a relationship with a patient's educational level and current employment, which is often recorded in these documents.

Interpretation of these models may shed some light. For example, when looking at absolute coefficient weights, we see some expected words corresponding to emotional needs (`depress`, `counsel`, `anxieti`, `pain`). However, the top two tokens for informational needs were years (`2011`, `2011.`), which seem to be a possible result of overfitting, or a complex set of variable correlation, as opposed to obviously helpful tokens. However, some of the words with high importance were consistent with informational needs, such as the breast cancer treatment `tamoxifen` suggesting that breast cancer patients may commonly have more or less informational needs.

3. Neural NLP models can be used with a patient's initial oncologist consultation document to predict a non-psychosocial outcome, **surviving at least five years** after this document was generated. This performance will be numeri-

cally higher than the performance of predicting the psychosocial outcomes, but will serve as a useful comparator, and balanced accuracy and AUC will be at least 0.80, a performance typically achieved by prior work predicting survival.

Compared to our psychosocial need predictions, our models were able to predict five years survival with numerically higher performance, and clearly surpassed our hypothesis threshold of 0.80. Our three neural models achieved balanced accuracy in the middle 0.80's, and AUC surpassing 0.90. This is inline with the performance of prior computational predictive models used for survival prediction [11]. We should also interpret these results in the context of known survival figures for patients solely based upon their cancer site and staging at diagnosis, which is commonly used to give patients a prognosis clinically. Future work could compare our results with the performance of these simpler prognoses.

We believe these results support that our methodology was successfully deployed, and that the lower results of our psychosocial outcome predictions reflect the increased difficulty of this task when using these types of documents, as opposed to a flaw in our methodology.

Interpreting our models, again using absolute coefficient weights for BOW models and integrated gradients for our neural models, was consistent with expectations. The top token for BOW, `palliat`, may correspond to a patient being referred for palliative care, which would often suggest an oncologist expects limited survival. Other important tokens correspond to types of cancer with generally good (`breast`, `lymphoma`) and bad (`lung`) prognoses. Future work could investigate whether these coefficients are positive or negative, to ensure these correspond with expectations.

When observing which words had predictive importance in a demonstrative document when used by our CNN and BERT models, we find that the oncologist's mention of ultrasound was supportive of survival, which may correspond to this technique being used for more survivable cancers such as cervical cancer. Similarly, a phrase like `an endometrial biopsy was negative for malignancy` was found to be important for predicting survival, which would correspond to a patient cancer that has not spread as far, a good prognostic factor.

Words like `repeat` and `multiple` were important for predicting a subject not surviving; we would expect these terms to indicate situations associated with poor survival prognosis, such as a patient needing to repeat a scan due to concerns about cancer recurrence, or having multiple disease sites.

5.2 Secondary Hypotheses

1. When neural NLP models are used with a patient's initial oncologist consultation document to predict psychosocial needs, the performance is numerically similar between **female and male patient subgroups**.

For both self-reported and clinician addressed needs, our models had numerically better performance when training and evaluating on a subgroup of patients with female biological sex than a subgroup with only males. These results are consistent with well documented prior literature that males are less likely to seek psychosocial support or endorse such needs. For example, while the rates of some psychiatric illnesses are smaller in males, the around double rate of females seeing a psychiatrist in our dataset was not expected. This does not correspond to known prevalence of common psychiatric diseases such as anxiety, depression, or schizophrenia in the general populations which have relatively small differences [6], or of depression, anxiety, or adjustment disorder in cancer populations, where there has been found to be no sex difference [47]. This suggests that there are situations where male cancer patients may benefit from a counsellor or psychiatrist but either refuse, or are not referred; this may make it harder for our models to predict these outcomes for male patients correctly.

2. When neural NLP models are used with a patient's initial oncologist consultation document to predict psychosocial needs, the performance is numerically similar between **subgroups based on a patient's cancer stage at diagnosis**.

The results of predicting psychosocial needs with subgroups based upon the cancer stage at diagnosis may be one of our most interesting results. We see numerically higher predictive performance for patients with Stage II cancer than with other cancers when predicting both emotional needs and whether a patient will see a psychiatrist.

When observing the results for predicting whether a patient will see psychiatry, we see the task is particularly difficult for patients diagnosed with Stage IV illness, whose cancer has spread to distant sites. This may simply correspond to the limited survival of these patients; it can take a month or more for a referred patient to be seen by psychiatry, so clinicians may not feel a referral is warranted, even if the model correctly identifies that a patient would benefit from this support. We also note our observed performance may be impacted by the relatively small number of patients who saw psychiatry, especially in our validation set.

Otherwise, it seems that for both clinician-addressed and self-reported needs, our models are able to predict better when training and evaluating on only patients who were diagnosed with Stage II illness. This is an interesting result, as it does not suggest that it is simply a question of disease severity, as performance is worse for both patients with less severe illness (Stage I), and more severe illness (Stages III, IV). A possible causal explanation could be that patients with Stage II illness have higher amounts of uncertainty than those with other stages, and uncertainty is often a source of anxiety. Those with Stage I illness generally will have an optimistic prognosis, while those with Stage III and especially Stage IV illness may have clearer poor prognoses. Patients with Stage II illness may be having more emotional needs stemming from worry, as their prognoses are often relatively good, but with a considerable chance of worsening disease and death. However, these results could be impacted by other correlations, especially considering that we do not have intact survival data for all patients.

3. The **number of tokens a model is able to use** will impact its ability to predict whether a patient will see a psychiatrist using a patient's initial oncologist consultation document. Specifically, the performance of BERT will be numerically worse than its variation, Longformer, which is able to use more words.

Models using transformers such as BERT have been able to achieve state-of-the-art performance on many applications of NLP, but in our work, seem to be outperformed by the older neural models, CNN and LSTM, especially when predicting our psychosocial outcomes. BERT's limitation of only being able to use 512 tokens may explain some of this difference, especially given that our documents have a

median of 945 tokens. As such, they are often limited to the first few sections of our oncologist consultation documents, even though our neural model interpretations suggest that portions in the middle of the document may be particularly useful.

The improved results of the Longformer model, which is able to use more tokens, supports this may indeed be BERT's limitation when predicting our psychosocial outcomes. We see improved performance when Longformer is able to use up to 2048 tokens. Few documents have more tokens than this, so it is unsurprising that using up to 4096 tokens worsens performance, as for many documents it would lead to more padding tokens.

However, Longformer's performance still did not appear as good as using the other neural models. Longformer is able to use more tokens due to limits in its ability to understand all words in relation to each other. Unlike BERT, where all words are connected with each other, Longformer only connects to a few words somewhat close and somewhat farther from itself. As such, CNN and LSTM models may be able to interpret the documents more holistically, and gain performance by observing patterns of words being present further away from each other.

When predicting survival, Longformer's ability to use more of a document's tokens seem to confer less benefit. The almost negligible difference of 0.01 in both balanced accuracy and AUC between Longformer using 256 versus 4096 suggests that the first 256 tokens may largely be sufficient to predict survival. This is consistent with our interpretations, where we found that important tokens (words) were usually found at the beginning of the consultation document when the oncologist describes the cancer and the patient's history to date.

5.3 Clinical Implications

The goal of this work was to determine the feasibility of using neural language models with a patient's initial oncologist consultation document to better understand their psychosocial needs, to be able to better detect psychosocial needs so that they can be addressed. We believe that this work supports that this application may be feasible, especially predicting clinician-addressed needs such as whether a patient will go on to see a psychiatrist or counsellor.

The performance of our models predicting psychosocial needs is not as high as

compared to when predicting survival. This is unsurprising, given that we would expect oncologists to describe many aspects of a patient's cancer related to prognosis, such as its type, stage, spread, and treatment so far. On the other hand, the treating oncologists will generally document fewer details directly related to psychosocial health. If all patients were to receive a comprehensive psychosocial assessment at the start of their care, we may expect that this document would allow much better predictions of our psychosocial outcomes. However, such an assessment is not routinely done. This work instead focuses on using a document that is widely available, the initial treating oncologist consultation. Our results suggest that, with improvement, this methodology may one day produce a clinically useful prediction, especially when predicting clinician-addressed needs.

It is unclear what performance would be needed for clinical application, and this may depend on the task at hand. We focused in this work on balanced accuracy and AUC, as they are metrics often used in clinical applications. They correspond to general predictive power, and balancing the importance of predicting both negative and positive labels. However, the particular use case of predicting psychosocial needs may have impact on what metrics would be required.

For example, a possible application could be to deploy these models to read a treating oncologist's initial consultation, and suggest that the oncologist asks patients predicted to have a clinician-addressed need whether they would like to see a counsellor or psychiatrist. In such a scenario, the oncologist may not mind a relatively high false positive rate, so long as it does not suggest asking this question for all patients. However, they would likely want a relatively low false negative rate as the point of this application would be to prevent patients from "falling through the cracks". As such, in this scenario, sensitivity (recall) and NPV may be more important than specificity, precision, and PPV.

For identifying the self-reported psychosocial needs, it is less clear than this combination of methodology and data may one day lead to a useful clinical application. We do note that the performance was better when using BERT, and when predicting with Stage II patients. This task may require more complex language models, such as transformers able to have more connected understanding of documents with more tokens. Further investigation may be warranted to understand why this task seems to be more difficult.

The possible applications of predicting self-reported psychosocial needs may be more focused on detection at a population level, as if a clinician wanted to know a patient's self-reported cancer needs in the moment, they could simply inquire, or ask them to fill a questionnaire. If these techniques could be shown to apply to different documents, such as the progress notes patients receive throughout their stay, our models could be used to show how self-reported needs might vary through a patient's cancer care. Lower performance may be tolerable in this situation if the prediction seems accurate in aggregate, though further work will be needed to explore such applications.

Of note, while we included survival prediction as a comparison to validate our methodology with a non-psychosocial outcome, this predictive task also has potential clinical application in helping address psychosocial needs. For example, patients nearing the end of life may benefit from care at a hospice, extra supports around estate planning, and from particular therapies such as meaning centred psychotherapy [72]. However, it can be unclear when patients are nearing their last six months of life. Predicting that a patient is unlikely to have six-month survival may allow these end-of-life resources to be better provided for these patients. This could improve not only quality-of-life, but also survival, as has been shown with the provision of early palliative care [71].

The results pertaining to our secondary hypotheses also have clinical implications. The worse performance when predicting psychosocial outcomes with male subjects substantiates the difficulty of supporting these patients with their psychosocial needs. Given that males may more often stand to benefit from psychosocial support but not access it, this may make it difficult for our models to learn the correct prediction. It could make sense to train models on only female patients, and apply them to patients of all sexes.

The results of predicting our outcomes using subgroups based on cancer stage at diagnosis may also have implications for the clinical utility of this work. If this result is indeed valid, it may mean that applications of these techniques are more accurate for those diagnosed with certain cancer stages than others. This would be important to know if deployed clinically; for example, perhaps the tool would not be used for those with Stage IV illness. However, these results also may speak to the current deficits in supportive cancer care. Or model's poor performance may

result from a lack of access of these resources for certain patients. For instance, our model may have predicted false positives for patients that would have benefited from a psychiatry referral, but were not referred due to psychiatric support not being available promptly enough, or the oncologist thinking it would be futile.

5.4 Implications for Using Neural Language Models in this Domain

As our work may be the first exploring the application of neural NLP techniques to predict psychosocial outcomes from cancer or other medical documents, this work may provide a baseline for other work to improve upon our results. Our findings that CNN and LSTM models often performed the best suggests these models should be continue to be explored.

Our work also suggests that transformers may need further adaption to be used in this domain, potentially owing to the length of full consultation documents. Many of our documents had many more tokens than BERT's maximum of 512. While using Longformer models seems to partially improve these results, further development of long-document transformers may be warranted, perhaps with different attention mechanisms, especially considering the better performance of our CNN and LSTM models even when Longformer is able to use 2048 tokens. The recently released BigBird model may possibly improve these results, though they largely seem to be using some of the adaptations found in Longformer [80]. Alternatively, given the potential relevance of certain portions of the documents, BERT using a middle or end 512 tokens may also be able to perform better, as may techniques using hierarchical attention.

We should also note the relatively strong performance of our non-neural BOW models. This may suggest that the more advanced neural models may still have further room for improvement. Alternatively, these results may also suggest that further exploration of additional n-gram and other non-neural NLP models is warranted.

5.5 Limitations

5.5.1 Validity of our Targets

We should consider the validity of the targets used in this work. The self-reported psychosocial needs are as valid as the construct validity of the questionnaire they are from, though the thresholds we used have not been used before. There may be more robust or applicable ways to construct targets from these PSSCAN-R results, such as excluding certain items.

The clinician-addressed need targets should be understood in the context of the problem their prediction is hoping to address. As we laid out earlier, cancer patients continue to have unmet psychosocial needs such as not always being able to see a counsellor or psychiatrist. Our models learned to predict whether patients have this need from data in a system where these unmet needs are present. As such, it may be worthwhile qualitatively exploring false positives, where patients are predicted to see a psychiatrist or counsellor but do not. If there seems to be situations where patients may have benefited from these supports even though they did not receive them, some amount of the false positives that our models predict may actually be helpful to address unmet needs.

Our survival targets' construct validity is limited by the accuracy of the mortality data provided to us and our technique to label it. In this work we excluded patients who do not have death dates recorded or dates clearly documenting they survived until a threshold such as five years. This meant we excluded a number of patients, which may introduce some bias into our data and results.

5.5.2 Internal Validity of our Results

A limitation of this work is that we solely considered the numerical differences of our results. We did not conduct statistical analysis to know if results are significantly different from each other, or from the thresholds established in our hypotheses. As our models can train differently every time, we cannot interpret our results with respect to their statistical significance. Future work could repeat our results numerous times to determine their variances, which would then allow the appropriate statistical tests to be conducted.

5.5.3 External Validity of our Results

A major limitation of this work is that our results are based upon evaluation on a hold-out validation set. This validation set was used during our development, and to select the number of training epochs. This may lead to some overfitting, as results may simply be a configuration of a model that fit the validation set the best, but would not fit new data as well. We plan to further validate our results on a hold-out test set in the future.

Further external validity could be established by evaluating our trained models on an entirely new dataset, such as patients who received BC Cancer care during a different time period, or even more robustly, data from a different cancer centre.

5.5.4 Methodological Limitations

This work was meant to be an initial investigation into the feasibility of predicting our targets using neural NLP models, not an exhaustive investigation into how these techniques could be optimized to accomplish such tasks.

Many avenues exist to further improve our work. We generally did not tune our neural models, instead using hyperparameters from prior work instead. We used a consistent learning rate across models, besides a larger learning rate for Longformer, and tuning this can often improve performance. Additionally, after some initial testing, we used randomized word vectors for CNN and LSTM models. Pretrained general English or clinical word embedding are available, and they may improve performance.

Similarly, BERT model pretrained on clinical and biomedical documents may also improve performance, even though we did not find this during our brief testing. Alternative strategies for dealing with class imbalance, and for text preprocessing, could also be attempted. We believe our results, especially when predicting survival, suggest that our configurations are grossly sufficient. Yet, future work could likely improve our performance further.

5.6 Future Work

Besides addressing some of the limitations discussed in Section 5.5, there are many ways to expand and build upon this work. We were interested in solely using

unstructured text to make these psychosocial predictions, but structured data could likely also be useful. Future work could investigate comparing and combining structured and unstructured (text) data.

We only used the initial oncologist consultation document for the predictions in this work. Multiple documents from the beginning of a patient's cancer care could be useful. As well, future work could extend this work to use documents over time, updating predictions. This would likely require other types of documents to be used, such as progress notes. A general prediction for any medical document could allow cancer patients to be assessed at every visit for current or future psychosocial needs. This may allow us to even better support our patient needs, especially longitudinally.

Chapter 6

Conclusion

In this work, we investigated whether neural language models can be used to predict the psychosocial needs of cancer patients from initial oncologist consultation documents. This is a novel application of NLP, as we were unable to find similar prior work in the literature. We investigated the prediction of both self-reported and clinician-addressed psychosocial cancer needs, and compared these to the prediction of a non-psychosocial cancer need, survival. We compared the non-neural NLP technique BOW with three neural methods, CNN, LSTM, and BERT. We conducted additional investigations using subgroups of our dataset to better understand our results and how these techniques may be applied to patients based on their biological sex or cancer stage at diagnosis. We also used techniques to interpret our trained models, to better understand how they came to their predictions.

We found that our models were able to predict clinician-addressed cancer needs with performance similar to some other applications of predictive models in psychiatry and psychosocial medicine. However, predicting self-reported emotional and informational cancer needs seemed to be more difficult. Our models were not able to reach our hypothesized performance threshold for these self-reported needs, even though the validity of our techniques was supported by our models being able to predict survival well. Additional analyses found differences in predictive performance between biological sexes, as well as with a patient's cancer stage at diagnosis. This may have implication for how to apply these techniques clinically. Given the relatively long length of our documents, we investigated the

impact of our transformer model, BERT, being able to use only the first portion of a document, finding that a long-document adaption of this model, Longformer, had higher performance.

Our work investigated a novel application of neural NLP by seeking to help detect the psychosocial needs of cancer patients so that these needs can be better met. We provide a foundation for future work to develop this application further, so that we can one day use these techniques to improve and extend the lives of our cancer patients.

Bibliography

[1] A. Abbe, C. Grouin, P. Zweigenbaum, and B. Falissard. Text mining applications in psychiatry: A systematic literature review. 25(2):86–100. ISSN 1557-0657. doi:10.1002/mpr.1481. URL http://onlinelibrary.wiley.com.ezproxy.library.ubc.ca/doi/10.1002/mpr.1481/abstract. → page 7

[2] A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). 6:52138–52160. ISSN 2169-3536. doi:10.1109/ACCESS.2018.2870052. → page 45

[3] A. Adhikari, A. Ram, R. Tang, and J. Lin. Rethinking Complex Neural Network Architectures for Document Classification. In *Proceedings of the 2019 Conference of the North*, pages 4046–4051. Association for Computational Linguistics. doi:10.18653/v1/N19-1408. URL http://aclweb.org/anthology/N19-1408. → page 21

[4] I. M. Alananzeh, J. V. Levesque, C. Kwok, Y. Salamonson, and B. Everett. The Unmet Supportive Care Needs of Arab Australian and Arab Jordanian Cancer Survivors: An International Comparative Survey. 42(3):E51, May/June 2019. ISSN 0162-220X. doi:10.1097/NCC.0000000000000609. URL https://journals.lww.com/cancernursingonline/Abstract/2019/05000/The_Unmet_Supportive_Care_Needs_of_Arab_Australian.17.aspx. → page 1

[5] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics. doi:10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909. → pages 10, 22

[6] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, fifth edition edition. ISBN 978-0-89042-555-8 978-0-89042-557-2.

doi:10.1176/appi.books.9780890425596. URL https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596. → page 49

[7] M. B. Amin, F. L. Greene, S. B. Edge, C. C. Compton, J. E. Gershenwald, R. K. Brookland, L. Meyer, D. M. Gress, D. R. Byrd, and D. P. Winchester. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. 67(2):93–99. ISSN 1542-4863. doi:10.3322/caac.21388. URL https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21388. → page 18

[8] F. D. Ashbury, H. Findlay, B. Reynolds, and K. McKerracher. A Canadian Survey of Cancer Patients' Experiences: Are Their Needs Being Met? 16 (5):298–306. ISSN 0885-3924. doi:10.1016/S0885-3924(98)00102-X. URL https://www.sciencedirect.com/science/article/pii/S088539249800102X. → page 7

[9] I. Banerjee, S. Bozkurt, J. L. Caswell-Jin, A. W. Kurian, and D. L. Rubin. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. (3):1–12. doi:10.1200/CCI.19.00034. URL https://ascopubs.org/doi/full/10.1200/CCI.19.00034. → page 9

[10] F. K. Barg, P. F. Cronholm, J. B. Straton, S. Keddem, K. Knott, J. Grater, P. Houts, and S. C. Palmer. Unmet psychosocial needs of pennsylvanians with cancer: 1986–2005. 110(3):631–639. ISSN 1097-0142. doi:10.1002/cncr.22820. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.22820. → page 18

[11] M. Beeksma, S. Verberne, A. van den Bosch, E. Das, I. Hendrickx, and S. Groenewoud. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. 19:36. ISSN 1472-6947. doi:10.1186/s12911-019-0775-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6394008/. → pages 9, 48

[12] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer. URL http://arxiv.org/abs/2004.05150. → page 22

[13] J. A. Benítez-Andrades, J.-M. Alija-Pérez, M.-E. Vidal, R. Pastor-Vargas, and M. T. García-Ordás. Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)–Based Automatic Classification of Tweets About Eating Disorders: Algorithm

Development and Validation Study. 10(2):e34492. doi:10.2196/34492. URL https://medinform.jmir.org/2022/2/e34492. → page 8

[14] S. Bird, E. Klein, and E. Loper. Natural Language Processing with Python [Book]. URL https://www.oreilly.com/library/view/natural-language-processing/9780596803346/. → pages 14, 20

[15] C. G. Blanchard, T. L. Albrecht, and J. C. Ruckdeschel. The crisis of cancer: Psychological impact on family caregivers. 11(2):189–94; discussion 196, 201–2. ISSN 0890-9091. → page 1

[16] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram. The ever-increasing importance of cancer as a leading cause of premature death worldwide. 127(16):3029–3030. ISSN 1097-0142. doi:10.1002/cncr.33587. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.33587. → page 1

[17] P. Butow, A. Girgis, and P. Schofield. Psychosocial aspects of delivering cancer care: An update. 37(1):20–22. ISSN 0311-306X. → page 1

[18] C. Clark, B. Wellner, R. Davis, J. Aberdeen, and L. Hirschman. Automatic Classification of RDoC Positive Valence Severity with a Neural Network. 75 Suppl:S120–S128. ISSN 1532-0464. doi:10.1016/j.jbi.2017.07.005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5705444/. → page 8

[19] C. A. Cuthbert, D. J. Boyne, X. Yuan, B. R. Hemmelgarn, and W. Y. Cheung. Patient-reported symptom burden and supportive care needs at cancer diagnosis: A retrospective cohort study. 28(12):5889–5899. ISSN 1433-7339. doi:10.1007/s00520-020-05415-y. URL https://doi.org/10.1007/s00520-020-05415-y. → page 7

[20] H.-J. Dai and J. Jonnagaddala. Assessing the severity of positive valence symptoms in initial psychiatric evaluation records: Should we use convolutional neural networks? 13(10):e0204493. ISSN 1932-6203. doi:10.1371/journal.pone.0204493. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204493. → page 8

[21] H.-J. Dai, C.-H. Su, Y.-Q. Lee, Y.-C. Zhang, C.-K. Wang, C.-J. Kuo, and C.-S. Wu. Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients. 11:533949. ISSN 1664-0640. doi:10.3389/fpsyt.2020.533949. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7874001/. → page 8

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. URL http://arxiv.org/abs/1810.04805. → page 21

[23] S. Eglowski. CREATE: Clinical Record Analysis Technology Ensemble. URL https://digitalcommons.calpoly.edu/theses/1771. → page 8

[24] C. Erker, K. Yan, L. Zhang, K. Bingen, K. E. Flynn, and J. Panepinto. Impact of pediatric cancer on family relationships. 7(5):1680–1688. ISSN 2045-7634. doi:10.1002/cam4.1393. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cam4.1393. → page 1

[25] W. Falcon, J. Borovec, A. Wälchli, N. Eggert, J. Schock, J. Jordan, N. Skafte, Ir1dXD, V. Bereznyuk, E. Harris, T. Murrell, P. Yu, S. Præsius, T. Addair, J. Zhong, D. Lipin, S. Uchida, S. Bapat, H. Schröter, B. Dayma, A. Karnachev, A. Kulkarni, S. Komatsu, Martin.B, J.-B. SCHIRATTI, H. Mary, D. Byrne, C. Eyzaguirre, Cinjon, and A. Bakhtin. PyTorchLightning/pytorch-lightning: 0.7.6 release. URL https://zenodo.org/record/3828935. → page 25

[26] A. C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran. Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. 8(1):7426. ISSN 2045-2322. doi:10.1038/s41598-018-25773-2. URL http://www.nature.com/articles/s41598-018-25773-2. → page 8

[27] M. Filannino, A. Stubbs, and O. Uzuner. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID Shared Tasks Track 2. 75 Suppl:S62–S70. ISSN 1532-0464. doi:10.1016/j.jbi.2017.04.017. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5656549/. → pages 7, 8

[28] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell. Extracting information from the text of electronic medical records to improve case detection: A systematic review. 23(5):1007–1015. ISSN 1067-5027, 1527-974X. doi:10.1093/jamia/ocv180. URL https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocv180. → page 8

[29] M. A. Gara, W. A. Vega, I. Lesser, M. Escamilla, W. B. Lawson, D. R. Wilson, D. E. Fleck, and S. M. Strakowski. The Role of Complex Emotions in Inconsistent Diagnoses of Schizophrenia. 198(9):609–613. ISSN

0022-3018. doi:10.1097/NMD.0b013e3181e9dca9. URL https://journals.lww.com/00005053-201009000-00001. → page 7

[30] B. Hao, H. Zhu, and I. Paschalidis. Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.57. URL https://aclanthology.org/2020.coling-main.57. → page 10

[31] T. C. Horrill, A. J. Browne, and K. I. Stajduhar. Equity-Oriented Healthcare: What It Is and Why We Need It in Oncology. 29(1):186–192. ISSN 1718-7729. doi:10.3390/curroncol29010018. URL https://www.mdpi.com/1718-7729/29/1/18. → page 1

[32] S. Hu, E. Fridgeirsson, G. van Wingen, and M. Welling. Transformer-Based Deep Survival Analysis. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, pages 132–148. PMLR. URL https://proceedings.mlr.press/v146/hu21a.html. → page 10

[33] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. URL http://arxiv.org/abs/2110.15621. → page 8

[34] D. A. John, I. Kawachi, C. S. Lathan, and J. Z. Ayanian. Disparities in Perceived Unmet Need for Supportive Services Among Patients With Lung Cancer in the Cancer Care Outcomes Research and Surveillance Consortium. 120(20):3178–3191. ISSN 0008-543X. doi:10.1002/cncr.28801. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4414401/. → page 1

[35] D. Jurafsky and J. H. Martin. *Speech and Language Processing (Draft)*. → page 19

[36] G. Karystianis, A. J. Nevado, C.-H. Kim, A. Dehghan, J. A. Keane, and G. Nenadic. Automatic mining of symptom severity from psychiatric evaluation notes. 27(1):e1602. ISSN 1557-0657. doi:10.1002/mpr.1602. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/mpr.1602. → page 7

[37] Y. Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for

Computational Linguistics. doi:10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181. → page 20

[38] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. URL http://arxiv.org/abs/1412.6980. → page 24

[39] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. URL http://arxiv.org/abs/2009.07896. → pages 27, 39

[40] H. Liang, B. Y. Tsui, H. Ni, C. C. S. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, L. He, J. Zhu, P. Tian, H. Shao, L. Zheng, R. Hou, S. Hewett, G. Li, P. Liang, X. Zang, Z. Zhang, L. Pan, H. Cai, R. Ling, S. Li, Y. Cui, S. Tang, H. Ye, X. Huang, W. He, W. Liang, Q. Zhang, J. Jiang, W. Yu, J. Gao, W. Ou, Y. Deng, Q. Hou, B. Wang, C. Yao, Y. Liang, S. Zhang, Y. Duan, R. Zhang, S. Gibson, C. L. Zhang, O. Li, E. D. Zhang, G. Karin, N. Nguyen, X. Wu, C. Wen, J. Xu, W. Xu, B. Wang, W. Wang, J. Li, B. Pizzato, C. Bao, D. Xiang, W. He, S. He, Y. Zhou, W. Haw, M. Goldbaum, A. Tremoulet, C.-N. Hsu, H. Carter, L. Zhu, K. Zhang, and H. Xia. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. 25(3):433–438. ISSN 1546-170X. doi:10.1038/s41591-018-0335-9. URL http://www.nature.com/articles/s41591-018-0335-9. → page 9

[41] M. Lin, S. Wang, Y. Ding, L. Zhao, F. Wang, and Y. Peng. An empirical study of using radiology reports and images to improve ICU-mortality prediction. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 497–498. doi:10.1109/ICHI52183.2021.00088. → page 10

[42] W. Linden, A. Andrea Vodermaier, R. McKenzie, M. C. Barroetavena, D. Yi, and R. Doll. The Psychosocial Screen for Cancer (PSSCAN): Further validation and normative data. 7:16, . ISSN 1477-7525. doi:10.1186/1477-7525-7-16. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2651864/. → page 7

[43] W. Linden, D. Yi, M. C. Barroetavena, R. MacKenzie, and R. Doll. Development and validation of a psychosocial screening instrument for cancer. 3(1):54, . ISSN 1477-7525. doi:10.1186/1477-7525-3-54. URL https://doi.org/10.1186/1477-7525-3-54. → pages 7, 16

[44] J. Liu, Z. Zhang, and N. Razavian. Deep EHR: Chronic Disease Prediction Using Medical Notes. In *Machine Learning for Healthcare Conference*, pages 440–464. PMLR. URL https://proceedings.mlr.press/v85/liu18b.html. → page 10

[45] D. Lu, T. M. L. Andersson, K. Fall, C. M. Hultman, K. Czene, U. Valdimarsdóttir, and F. Fang. Clinical Diagnosis of Mental Disorders Immediately Before and After Cancer Diagnosis: A Nationwide Matched Cohort Study in Sweden. 2(9):1188–1196. ISSN 2374-2437. doi:10.1001/jamaoncol.2016.0483. URL https://doi.org/10.1001/jamaoncol.2016.0483. → page 1

[46] C. Manning, P. Raghavan, and H. Schuetze. *Introduction to Information Retrieval*. Cambridge University Press. → page 20

[47] A. J. Mitchell, M. Chan, H. Bhatti, M. Halton, L. Grassi, C. Johansen, and N. Meader. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: A meta-analysis of 94 interview-based studies. 12(2):160–174. ISSN 14702045. doi:10.1016/S1470-2045(11)70002-X. URL https://linkinghub.elsevier.com/retrieve/pii/S147020451170002X. → page 49

[48] M. G. Nayak, A. George, M. Vidyasagar, S. Mathew, S. Nayak, B. S. Nayak, Y. Shashidhara, and A. Kamath. Quality of Life among Cancer Patients. 23 (4):445–450. ISSN 0973-1075. doi:10.4103/IJPC.IJPC_82_17. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5661349/. → page 1

[49] S. Newell, R. W. Sanson-Fisher, A. Girgis, and A. Bonaventura. How well do Medical oncologists perceptions' reflect their patients' reported physical and psychosocial problems? 83(8):1640–1651. → pages 2, 46

[50] J.-J. Nunez, T. T. Nguyen, Y. Zhou, B. Cao, R. T. Ng, J. Chen, B. N. Frey, R. Milev, D. J. Müller, S. Rotzinger, C. N. Soares, R. Uher, S. H. Kennedy, and R. W. Lam. Replication of machine learning methods to predict treatment outcome with antidepressant medications in patients with major depressive disorder from STAR*D and CAN-BIND-1. 16(6):e0253023. ISSN 1932-6203. doi:10.1371/journal.pone.0253023. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0253023. → page 45

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang,

Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html. → page 14

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. 12:2825–2830. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v12/pedregosa11a.html. → page 24

[53] B. Pillay, S. J. Lee, L. Katona, S. Burney, and S. Avery. Psychosocial factors predicting survival after allogeneic stem cell transplant. 22(9):2547–2555. ISSN 0941-4355, 1433-7339. doi:10.1007/s00520-014-2239-7. URL http://link.springer.com/10.1007/s00520-014-2239-7. → page 1

[54] M. Pinquart and P. R. Duberstein. Depression and cancer mortality: A meta-analysis. 40(11):1797–1810. ISSN 0033-2917. doi:10.1017/S0033291709992285. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935927/. → page 1

[55] J. D. Posada, A. J. Barda, L. Shi, D. Xue, V. Ruiz, P.-H. Kuan, N. D. Ryan, and F. R. Tsui. Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records. 75: S94–S104. ISSN 15320464. doi:10.1016/j.jbi.2017.05.019. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046417301181. → page 8

[56] J. Pyo, M. Ock, M. Lee, J. Kim, J. Cheon, J. Cho, J. H. Kwon, H. Kim, H.-S. Im, Y. J. Min, and S.-J. Koh. Unmet needs related to the quality of life of advanced cancer patients in Korea: A qualitative study. 20(1):58. ISSN 1472-684X. doi:10.1186/s12904-021-00749-8. URL https://doi.org/10.1186/s12904-021-00749-8. → page 1

[57] K. Rajput, G. Chetty, and R. Davey. Performance Analysis of Deep Neural Models for Automatic Identification of Disease Status. In *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, pages 136–141. doi:10.1109/iCMLDE.2018.00033. → page 9

[58] A. Rios and R. Kavuluru. Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles. In *Proceedings*

of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, pages 258–267. ACM, . ISBN 978-1-4503-3853-0. doi:10.1145/2808719.2808746. URL https://dl.acm.org/doi/10.1145/2808719.2808746. → pages 8, 20

[59] A. Rios and R. Kavuluru. Ordinal Convolutional Neural Networks for Predicting RDoC Positive Valence Psychiatric Symptom Severity Scores. 75 Suppl:S85–S93, . ISSN 1532-0464. doi:10.1016/j.jbi.2017.05.008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5682241/. → pages 8, 20

[60] C. I. Ripamonti, D. Santini, E. Maranzano, M. Berti, F. Roila, and on behalf of the ESMO Guidelines Working Group. Management of cancer pain: ESMO Clinical Practice Guidelines. 23:vii139–vii154. ISSN 0923-7534, 1569-8041. doi:10.1093/annonc/mds233. URL https://academic.oup.com/annonc/article-lookup/doi/10.1093/annonc/mds233. → page 2

[61] R. Sanson-Fisher, A. Girgis, A. Boyes, B. Bonevski, L. Burton, P. Cook, and S. C. R. Group. The unmet supportive care needs of patients with cancer. 88 (1):226–237. → page 1

[62] J. Savard, H. Ivers, and M.-H. Savard. Capacity of the Edmonton Symptom Assessment System and the Canadian Problem Checklist to screen clinical insomnia in cancer patients. 24(10):4339–4344. ISSN 1433-7339. doi:10.1007/s00520-016-3273-4. URL https://doi.org/10.1007/s00520-016-3273-4. → page 7

[63] L. R. Schover. The impact of breast cancer on sexuality, body image, and intimate relationships. 41(2):112–120. ISSN 1542-4863. doi:10.3322/canjclin.41.2.112. URL https://onlinelibrary.wiley.com/doi/abs/10.3322/canjclin.41.2.112. → page 1

[64] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. ISBN 978-1-107-29801-9. doi:10.1017/CBO9781107298019. URL http://ebooks.cambridge.org/ref/id/CBO9781107298019. → page 23

[65] S. Singer. Psychosocial Impact of Cancer. In U. Goerling and A. Mehnert, editors, *Psycho-Oncology*, Recent Results in Cancer Research, pages 1–11. Springer International Publishing. ISBN 978-3-319-64310-6. doi:10.1007/978-3-319-64310-6_1. URL https://doi.org/10.1007/978-3-319-64310-6_1. → page 1

[66] A. Smrke, B. Leung, A. Srikanthan, M. McDonald, A. Bates, and C. Ho. Distinct Features of Psychosocial Distress of Adolescents and Young Adults with Cancer Compared to Adults at Diagnosis: Patient-Reported Domains of Concern. 9(4):540–545. ISSN 2156-5333. doi:10.1089/jayao.2019.0157. URL https://www.liebertpub.com/doi/full/10.1089/jayao.2019.0157. → page 7

[67] W. K. W. So, C.-L. Wong, K.-C. Choi, C. W. H. Chan, J. C. Y. Chan, B. M. H. Law, R. W. M. Wan, S. S. S. Mak, W.-M. Ling, W.-T. Ng, and B. W. L. Yu. A Mixed-Methods Study of Unmet Supportive Care Needs Among Head and Neck Cancer Survivors. 42(1):67–78, January/February 2019. ISSN 0162-220X. doi:10.1097/NCC.0000000000000542. URL https://journals.lww.com/cancernursingonline/Abstract/2019/01000/A_Mixed_Methods_Study_of_Unmet_Supportive_Care.9.aspx. → page 1

[68] R. Steele and M. I. Fitch. Why patients with lung cancer do not want help with some needs. 16(3):251–259. ISSN 0941-4355. doi:10.1007/s00520-007-0301-4. → page 2

[69] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. URL http://arxiv.org/abs/1703.01365. → page 27

[70] W. Söllner, A. DeVries, E. Steixner, P. Lukas, G. Sprinzl, G. Rumpold, and S. Maislinger. How successful are oncologists in identifying patient distress, perceived social support, and need for psychosocial counselling? 84(2): 179–185. ISSN 0007-0920. doi:10.1054/bjoc.2000.1545. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2363697/. → pages 2, 46

[71] J. S. Temel, J. A. Greer, A. Muzikansky, E. R. Gallagher, S. Admane, V. A. Jackson, C. M. Dahlin, C. D. Blinderman, J. Jacobsen, W. F. Pirl, J. A. Billings, and T. J. Lynch. Early Palliative Care for Patients with Metastatic Non–Small-Cell Lung Cancer. 363(8):733–742. ISSN 0028-4793. doi:10.1056/NEJMoa1000678. URL https://doi.org/10.1056/NEJMoa1000678. → page 53

[72] L. P. M. Thomas, E. A. Meier, and S. A. Irwin. Meaning-Centered Psychotherapy: A Form of Psychotherapy for Patients With Cancer. 16(10): 488. ISSN 1535-1645. doi:10.1007/s11920-014-0488-2. URL https://doi.org/10.1007/s11920-014-0488-2. → page 53

[73] T. Tran and R. Kavuluru. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. 75:

S138–S148. ISSN 1532-0464. doi:10.1016/j.jbi.2017.06.010. URL https://www.sciencedirect.com/science/article/pii/S1532046417301338. → page 8

[74] H. Wang, Y. Li, S. A. Khan, and Y. Luo. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. 110:101977. ISSN 0933-3657. doi:10.1016/j.artmed.2020.101977. URL https://www.sciencedirect.com/science/article/pii/S0933365720312422. → page 9

[75] L. Watson, S. Qi, A. DeIure, C. Link, L. Chmielewski, A. Hildebrand, K. Rawson, and D. Ruether. Using Autoregressive Integrated Moving Average (ARIMA) Modelling to Forecast Symptom Complexity in an Ambulatory Oncology Clinic: Harnessing Predictive Analytics and Patient-Reported Outcomes. 18(16):8365. doi:10.3390/ijerph18168365. URL https://www.mdpi.com/1660-4601/18/16/8365. → page 6

[76] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6. → pages 14, 25

[77] C.-S. Wu, C.-J. Kuo, C.-H. Su, S. Wang, and H.-J. Dai. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. 260:617–623, . ISSN 01650327. doi:10.1016/j.jad.2019.09.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0165032719306172. → page 8

[78] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu. Deep learning in clinical natural language processing: A methodical review. 27(3):457–470, . doi:10.1093/jamia/ocz200. URL https://academic.oup.com/jamia/article/27/3/457/5651084. → page 9

[79] Q. Yuan, T. Cai, C. Hong, M. Du, B. E. Johnson, M. Lanuti, T. Cai, and D. C. Christiani. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal

Cohort of Patients With Lung Cancer. 4(7):e2114723–e2114723. ISSN 2574-3805. doi:10.1001/jamanetworkopen.2021.14723. URL https://doi.org/10.1001/jamanetworkopen.2021.14723. → page 9

[80] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big Bird: Transformers for Longer Sequences. URL http://arxiv.org/abs/2007.14062. → page 54

[81] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T.-S. Chung. A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model. 2022:e7893775. ISSN 1687-5265. doi:10.1155/2022/7893775. URL https://www.hindawi.com/journals/cin/2022/7893775/. → page 8

[82] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. Dive into deep learning. → pages 20, 21

Appendix A

Supporting Materials

A.1 Specific Details for Data Preparation and Processing

A.1.1 Patient Data and Selection

We extracted patient data from the provided `mt2a_cancer_details.csv`, and PSSCAN-R data from `m2c_PSSCAN_details.csv`. To exclude patients with more than one cancer treated at BC Cancer, we removed patients with duplicate entries in either of these files.

A.1.2 Document Preparation

We selected documents used in this work based on the data provided in `q18-0193_document_details.csv`.

To select only consultation documents, we filtered documents by `document_type`, requiring that it be `ONC` or `H&P`, which corresponds to consultations.

To only capture documents produced by treating physicians at BC Cancer, we required their `fasr_discipline` be one of the following, corresponding to the following medical or surgical disciplines:

1. `MED ONCOLOGY` (Medical Oncology)
2. `RAD ONCOLOGY` (Radiation Oncology)

3. HEMATOLOGY (Hematology)

4. DERMATOLOGY (Dermatology)

5. GASTROENTER (Gastroenterology and Gastrointestinal Surgery)

6. MUSCULOSKELE (Orthopedic Surgery)

7. NEUROLOGY (Neurology and Neurosurgery)

8. SURGERY (Cancer and General Surgery)

9. GYNECOLOGY (Gynecology)

A.1.3 Target Preprocessing

Seeing a Discipline

We judged a patient as having seen a discipline if they have a document where fasr_discipline was equal to either PSYCHIATRY, for seeing psychiatry, and SOCIAL WORK, for seeing counselling. Documents generated by counsellors are marked in this fashion, regardless of whether the clinicians were trained as social workers or counselling psychologists.

A.1.4 Staging and Metastatic Status

We determined stage by using a patient's stage summary *COL_AJCC_Stage_Sum*. However, some patients are missing this data. Future work could recover some of this missing staging data by looking at the clinical and pathological staging data, and calculating the summary stage. This is not a trivial task, as it would require using the desired AJCC staging handbook rules. As this was not integral to this thesis project, we have deferred this for now.

We determined metastatic status simply by using *COL_AJCC_Stage_Sum*, assigning that a patient has metastatic disease if they are Stage IV, but are non-metastatic if Stage I-III.

Table A.1: Choosing Bag-of-Word's number of words and C regularization hyperparameter. We based this on the performance when predicting which patients will see a psychiatrist within five of the generation of the initial oncologist cancer document used by the models. AUC: Receiver-operator-curve area-under-curve.

Model	Words	C	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BoW	1000	0.1	0.70	0.71	0.71	0.08	0.72	0.70
BoW	1000	1	0.78	0.70	0.70	0.09	0.61	0.78
BoW	1000	2	0.80	0.68	0.68	0.09	0.56	0.80
BoW	5000	0.1	0.77	0.72	0.72	0.10	**0.67**	0.77
BoW	5000	0.2	0.80	**0.73**	**0.73**	0.11	0.66	0.81
BoW	5000	0.3	0.82	0.71	0.71	0.11	0.60	0.82
BoW	5000	0.5	0.85	0.68	0.68	0.11	0.51	0.85
BoW	5000	1	**0.8**7	0.67	0.67	0.12	0.46	**0.88**
BoW	10000	0.01	0.61	0.62	0.62	0.06	0.62	0.61
BoW	10000	0.10	0.78	**0.73**	**0.73**	0.10	**0.67**	0.78
BoW	10000	0.5	0.87	0.69	0.69	**0.13**	0.51	**0.88**

A.2 Specific Details for Model Training and Tuning

A.2.1 Bag-of-Words Hyperparameters

For our BOW, we used a common L2-regularized logistic regression classifier. We set the number of words and regularization strength C hyperparameters empirically, as shown in Table A.1. We choose 5000 words as further increases let to incrementally less improvement, at the possible risk of overfitting and computational costs, and used a regularization constant C of 0.2.

A.2.2 Comparing Techniques for Dealing with Class Imbalance

We compared two techniques to deal with class imbalance, undersampling, and weighing our cross-entropy loss. To compare performance, we compared our models when predicting whether subjects will see a psychiatrist within the five years following when the document used for training was generated, at the start of their cancer. We show these results in Table A.2.

Table A.2: Comparison two different methods for dealing with class imbalance using Bag-of-Word (BoW), Convolutional Neural Network (CNN), Long short term memory (LSTM) and bidirectional encoder representations from transformers (BERT) models. Here we show the results when using the models to predict whether a patient will see a psychiatrist within five years. AUC: Receiver-operator-curve area-under-curve.

Model	Fix	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BoW	Undersampling	0.80	**0.73**	0.73	0.11	0.06	0.66
BoW	Weighted Loss	0.80	**0.73**	0.73	0.11	0.06	0.66
CNN	Undersampling	0.87	0.69	0.78	**0.13**	**0.07**	0.50
CNN	Weighted Loss	0.83	**0.73**	**0.80**	0.12	**0.07**	0.63
LSTM	Undersampling	**0.98**	0.50	0.43	0.0	0.0	0.0
LSTM	Weighted Loss	0.58	0.63	0.71	0.06	0.03	**0.70**
BERT	Undersampling	0.79	0.67	0.70	0.09	0.05	0.54
BERT	Weighted Loss	0.73	0.68	0.73	0.08	0.04	0.63

A.3 Additional Result Tables

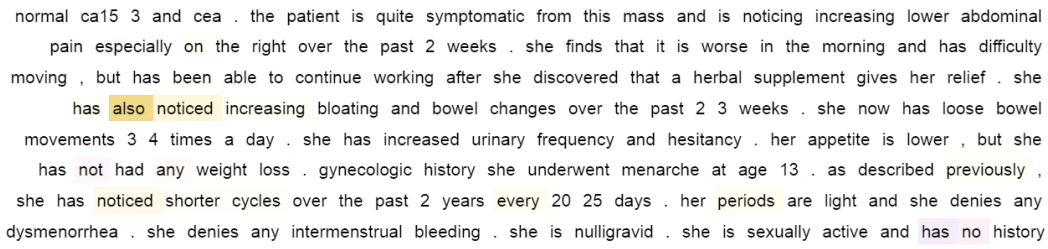
This section contains additional tables reporting our results.

Table A.3: Comparison of model performance when whether a subject reports a minimum threshold of emotional needs, using Bag-of-Words (BoW) and Convolutional Neural Networks (CNN). AUC: Receiver-operator-curve area-under-curve.

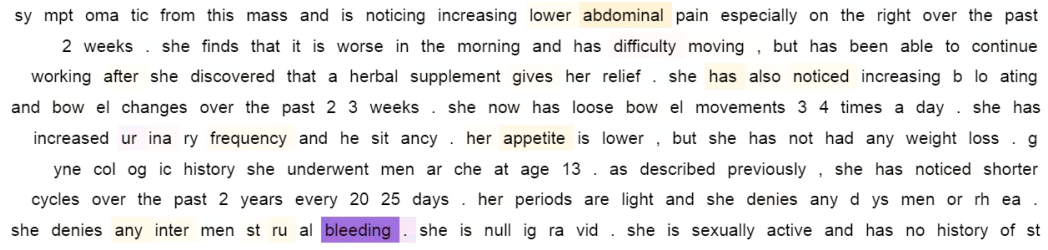
Model	Needs	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	1	0.63	0.63	0.63	0.67	0.69	0.65
BOW	2	0.64	0.65	0.65	0.51	0.41	0.67
BOW	3	0.65	0.65	0.65	0.34	0.23	0.65
BOW	4	0.69	0.66	0.66	0.18	0.10	0.63
BOW	5	0.78	0.62	0.62	0.07	0.04	0.46
CNN	1	0.65	0.65	0.70	**0.68**	**0.70**	0.66
CNN	2	0.67	**0.67**	**0.73**	0.54	0.44	0.69
CNN	3	0.67	0.66	0.72	0.36	0.25	0.65
CNN	4	0.42	0.58	0.62	0.12	0.07	**0.77**
CNN	5	**0.95**	0.55	0.64	0.09	0.07	0.12

Table A.4: Comparison of model performance when predicting whether a subject has a minimum number of informational needs, using Bag-of-Words (BOW) and Convolutional Neural Networks (CNN). AUC: Receiver-operator-curve area-under-curve.

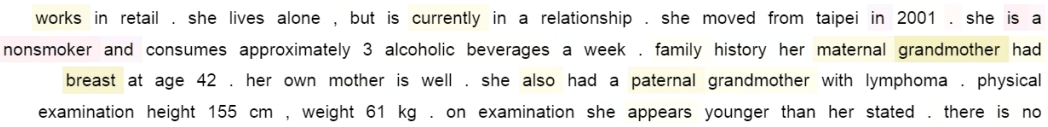
Model	Needs	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	1	0.61	0.60	0.60	0.64	**0.64**	0.64
BOW	2	0.59	0.59	0.59	0.47	0.38	0.61
BOW	3	0.59	0.59	0.59	0.35	0.25	0.58
BOW	4	0.61	0.60	0.60	0.27	0.18	0.59
CNN	1	0.60	0.60	0.63	**0.65**	0.63	**0.69**
CNN	2	0.58	**0.61**	0.64	0.49	0.38	0.67
CNN	3	**0.67**	0.61	**0.65**	0.36	0.28	0.51
CNN	4	0.62	0.60	0.64	0.28	0.18	0.58



(a) Excerpt of a clinical document describing symptoms with CNN word importance



(b) Excerpt of a clinical document describing symptoms with BERT word importance



(c) Excerpt of a clinical document describing family and social history with CNN word importance

Figure A.1: Red-green colour-blindness accessible version of Figure 4.1. Interpretation of our convolutional neural network (CNN) and bidirectional encoder representations from transformers (BERT) models, showing the importance of words when predicting whether a patient will see a psychiatrist, according to interpretation with integrated gradients. We show excerpts from a clinical document that has been anonymized for presentation here, with any potentially identifying words, names, dates or numbers changed. We show only some relevant excerpts, showing a portion describing symptoms in (a) and (b). We also show a segment describing social and family history in (c) which is only used by the CNN model due to BERT's token limits. The darker the orange highlighting, the more predictive of seeing a psychiatrist, while blue similarly corresponds to not seeing a psychiatrist.

Table A.5: Model performance when using Bag-of-Word (BOW) and Convolutional Neural Network (CNN) models to predict surviving the specified number of months after the document used by the model was generated. AUC: Receiver-operator-curve area-under-curve.

Model	Months Survived	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	6	0.81	0.82	0.82	0.88	**0.97**	0.80
BOW	36	0.82	0.82	0.82	0.83	0.86	0.80
BOW	60	**0.84**	0.83	0.83	0.77	0.72	0.82
CNN	6	0.83	0.81	0.90	**0.90**	0.96	0.84
CNN	36	0.83	0.83	0.91	0.85	0.86	0.83
CNN	60	**0.84**	**0.85**	**0.93**	0.78	0.71	**0.88**

Table A.6: Model performance when predicting whether a patient will see psychiatry in the five years after the document used by the models was generated. Here we compare the results of using Bag-of-Words (BOW) and Convolutional Neural Networks (CNN) models when training with randomly selected smaller numbers of patients (n). AUC: Receiver-operator-curve area-under-curve.

Model	n	Accuracy	Balanced Accuracy	AUC	F1	Precision	Recall
BOW	3000	0.84	0.65	0.65	0.09	0.05	0.45
BOW	4000	0.88	0.63	0.63	0.10	0.06	0.38
BOW	6000	0.87	0.66	0.66	**0.11**	0.06	0.45
BOW	9000	0.83	0.69	0.69	**0.11**	0.06	0.55
BOW	12000	0.82	0.70	0.70	0.10	0.06	0.57
BOW	15000	0.83	0.68	0.68	0.10	0.06	0.54
BOW	20000	0.82	0.71	0.71	**0.11**	0.06	0.59
BOW	30953	0.80	**0.73**	0.73	**0.11**	0.06	**0.66**
CNN	3000	0.79	0.58	0.58	0.06	0.03	0.35
CNN	4000	**0.98**	0.52	0.66	0.07	**0.14**	0.05
CNN	6000	0.85	0.62	0.65	0.08	0.05	0.38
CNN	9000	0.88	0.65	0.68	**0.11**	0.06	0.40
CNN	12000	0.71	0.63	0.67	0.07	0.04	0.55
CNN	15000	0.81	0.64	0.67	0.08	0.05	0.48
CNN	20000	0.79	0.66	0.68	0.09	0.05	0.52
CNN	30953	0.83	0.71	**0.75**	**0.11**	0.06	0.59

pelvic ultrasound performed on march 18 , 2014 , revealing complex left adnexal lesion with a cystic component measuring 8 . 3 cm and a solid irregular marrow component measuring 3 . 1 cm . the uterus and right ovary appeared normal . she went on to see dr . muir in late april and an endometrial biopsy was negative for malignancy . she had repeat ultrasound on april 29 , 2014 , which revealed that the left adnexal lesion was enlarging , now measuring 10 . 2 x 10 . 1 x 9 . 6 cm with multiple peripheral solid nodules with the largest

(a) Excerpt of a clinical document describing a patient's cancer with CNN word importance

canada , she had a repeat pe l vic ultrasound performed on april 14 , 2012 , revealing complex right ad ne xa l les ion with a cy stic component measuring 8 . 5 cm and a solid irregular marrow component measuring 3 . 3 cm . the ut erus and left o vary appeared normal . she went on to see dr . cohen in late april and an end ome tri al bio psy was negative for mali gnan cy . she had repeat ultrasound on may 22 , 2013 , which revealed that the right ad ne xa l les ion was en lar ging , now measuring 10 . 1 x 10 . 0 x 9 . 3 cm with multiple peripheral

(b) Excerpt of a clinical document describing a patient's cancer with BERT word importance

Figure A.2: Red-green colour-blindness accessible version of Figure 4.2.Interpretation of our convolutional neural network (CNN) and bidirectional encoder representations from transformers (BERT) models, showing the importance of words when predicting whether a patient will survive for five years, according to interpretation with integrated gradients. We show excerpts from a clinical document that has been anonymized for presentation here, with any potentially identifying words, names, dates or numbers changed. We show an excerpt describing a patient's cancer. The darker the orange highlighting, the more predictive of surviving five years, while blue similarly corresponds to not surviving.