The Use of Natural Language Processing and Machine Learning for Early Diagnosis of Lung and Ovarian Cancer

Grace Turner

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2022

Reading Committee:
Meliha Yetisgen, Chair
Matthew Thompson

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education

University of Washington

Abstract

The Use of Natural Language Processing and Machine Learning for Early Diagnosis of Lung and Ovarian Cancer

Grace Turner

Chair of the Supervisory Committee:
Meliha Yetisgen
Biomedical Informatics and Medical Education

Cancer is a serious diagnosis and diagnostic delay is correlated with reductions in survival rates following treatment. For many cancers, providers can only rely on symptoms and signs to diagnose patients. These details are recorded primarily free text clinical notes. Natural language processing (NLP) can be used to extract symptoms/signs from these notes for population level diagnosis screening. This creates opportunity for machine learning to alert providers earlier in the diagnostic process using existing, but easily overlooked information.

Thus, the focus of this thesis was to determine opportunities for reducing diagnostic delay in ovarian and lung cancer. A symptom extraction model trained on a primarily COVID-19 population was adapted to lung and ovarian cancer populations. The model then extracted symptoms/signs from a retrospective case-control study (ovarian) developed as part of this work as a well a leveraged study (lung). Symptom frequencies for ovarian cancer were then explored across different routes to diagnosis. Finally, this thesis developed experiments using machine learning models to predict lung and ovarian cancer prior to diagnosis. This work showed early prediction using symptoms was only possible on the lung cohort. Nevertheless, both cohorts had significantly higher "next step" recommendations in cases as compared to controls, even 6 months prior to diagnosis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to the University of Washington and advisors Meliha Yetisgen and Matthew Thompson. Many thanks to all the unsung heroes who kept the author sane doing a Master's through the COVID lockdown. Finally, none of this work would have been possible without the author's funding sources, the Gordon and Betty Moore Foundation and the Cancer Research UK CanTest Initiative.

DEDICATION

This work is dedicated to those diagnosed with ovarian cancer, at the University of Washington Medicine and elsewhere.

Chapter 1

INTRODUCTION

Cancer caused the deaths of 609,640 people in the United States in 2018, and is one of the top causes of death in the United States[22]. That being said, the importance of early diagnosis cannot be overstated. There are marked differences in survival rates depending on the stage of cancer at diagnosis[22]. Later stage cancers are generally larger, and many have metastasized, or spread to surrounding organ systems. Patients diagnosed with later stage cancers have a worsened prognosis compared to earlier stage cancers[22].

The slow, often hidden, progression of many cancers may result in delays in diagnosis. The longer the delay before diagnosis, the longer the delay before treatment and the longer the cancer has to grow. For example, with ovarian cancer, early stage 5 year survival rates are 80-90%, while later stage survival rates are only 20-30%[24]. However, in ovarian cancer the proportion of patients with a later stage diagnosis (stage 3-4) is close to 70%[24]. Similarly, with lung cancer, the 5 year survival rate is 40-60% and 10-20% respectively[18]. As many as 55% of these patients are diagnosed at a later stage[28]. This disconnect is what drives much of the research into diagnostic delay, with the hopes of improving patient outcomes by increasing the number of early stage diagnoses.

There are two main sources of diagnostic delay. The first source of delay is the time between the patient entering the disease state and initial suspicion by a healthcare provider. For cancers without a screening test, this is the delay between when the patient first experiences a symptom and when a provider first suspects cancer. In many cases, the first indication that a provider is suspicious of cancer is when they order or recommend the appropriate "next step" on the diagnostic pathway. In the case of ovarian cancer, for example, this is an ultrasound (US), CT, or gynecology referral[19]. In the case of lung cancer, this typically

is a Chest CT, Chest X-Ray, or Chest MRI[2]. The second segment of delay is between the first indication of suspicion by the provider and the actual diagnosis. The former aspect of delay invites automated diagnostic prediction to augment provider intuition.

And so, the primary goal of this work was to ascertain the potential for symptoms/signs to address the former aspect of delay through machine learning. This was done with two cancer types: lung and ovarian cancer. Unlike breast or cervical cancer, where a screening test can start the diagnostic process prior to any symptomatic expression, cancers like ovarian or lung do not have a recommended general population screening test[23]. This means that, for now, symptoms/signs are the earliest possible signal to begin the diagnostic process for these two cancers.

A common way to uncover patterns in symptomatic expression prior to cancer diagnosis is to use electronic health record (EHR) data. Data stored in the EHR typically follows two forms: discrete, or coded information, and notes, or textual information[20]. Discrete information, such as diagnosis codes in the form of ICD-9 or ICD-10 codes, combine with textual to form the entire picture of a patient as perceived by the health system. Despite ICD-9 and ICD-10 including a variety of codes for what typically are known as symptoms/signs, e.g. "abdominal pain" (R10.9, 789.00), many symptoms/signs tend to be severely undercoded. In previous work, the author's lab found that symptoms/signs extracted from the note were far more prevalent than their coded equivalents on the same patient cohort. This pattern was found in several other studies, including Chan et. al.[4]. Thus, the note is generally a richer source of symptoms/signs, but practical considerations limit their use in secondary purposes due to the need for manual chart review. However, advances in natural language processing (NLP) allow the extraction of symptoms/signs automatically from the note. This enables the use of this rich data source in a cost-effective, scalable manner.

In order to accurately predict both ovarian and lung cancer it is important to understand the diseases and their progression. In 2018, ovarian cancer was diagnosed in approximately 22,240 patients with 14,070 deaths in the United States[26]. Despite only 2.5% of women diagnosed with cancer experiencing ovarian cancer, it accounted for 5% of the deaths[26]. In

short, it is a relatively rare disease, but often deadly to the patients it afflicts.

Symptoms/signs of ovarian cancer that have had significant odds ratios in prior work include fatigue, abdominal pain, pelvic pain, post-menopausal bleeding, distention, ascites, bloating, loss of appetite, early satiety, weight loss, urinary urgency, incontinence, frequency, other issues with bowel habits, and nausea [7][1][25]. A discovery of an abdominal or pelvic mass during exam might also indicate cancer[8][7]. Ovarian cancer symptoms/signs are easily confused with other, far more common indications, as well as simple menopause and aging. As an example, menopause can contribute to a loss of appetite, while a UTI or simple weakening of the bladder muscles can cause many of the symptoms/signs related to urinary urgency[24]. Nevertheless, from the literature many patients experience these symptoms/signs 6 or more months prior to diagnosis[8]. In an analysis by Goff et. al., high frequencies (more than 12 times a month) of pelvic, abdominal pain and bloating all had positive odds ratios greater than 6 months prior to diagnosis[8].

Lung cancer is a disease that was diagnosed in 234,030 people in the USA in 2018, with 121,680 deaths attributed to lung cancer[22]. While it is common in certain sub-populations like smokers, it is not exclusive to those patients and the symptoms/signs experienced are varied and complex[22]. The symptoms/signs that are most consistently mentioned in the literature include cough, wheezing, dyspnea or shortness of breath (SOB), fever, clubbing, and weight loss[3]. Other symptoms/signs include various kinds of bone and back pain, fatigue, and more[3]. To note is that many of these symptoms/signs could be confused with gastroesophageal reflux disease (coughing), influenza (coughing, fever, fatigue, shortness of breath), and other more common disease states.

The complexity and confusing nature of the symptomatic expression of both diseases indicates the potential use case for a machine learning classification task, so long as the symptoms can be extracted appropriately from the note. In this work, a NLP symptom extraction model was adapted to these two cancer contexts. An existing case-control dataset for lung cancer was leveraged and used as a template for a similar retrospective case-control dataset observing the year prior to diagnosis for patients diagnosed with ovarian cancer. Then

the symptom extraction model extracted symptoms/signs for both the lung and the ovarian cancer cohort. Then uni-variate analyses on the ovarian cancer cohort were performed. Another part of the team developed the lung case-control cohort and performed the uni-variate analysis of lung cancer, and so that work was not included in this thesis. Finally, the extracted symptoms/signs were used to develop predictive machine learning models in experiments on both the ovarian and lung cancer cohorts.

The rest of this thesis is organized in the following manner. First, Chapter 2 (Related Work) reviews different types of predictive tools and their effectiveness for cancer prediction. It also reviews the different potential mechanisms for symptom information extraction. Then, Chapter 3 (Methods) discusses the datasets used in this work and the details of the experiments conducted in this work. Chapter 4 (Results) discusses the direct results from the experiments. Finally, Chapter 5 (Discussion) explores the potential future work and proposes some potential rationale for the discoveries unearthed in the results section.

Chapter 2

RELATED WORK

This chapter reviews both the current body of work related to symptom extraction as well as work related to early prediction of cancer. Both bodies of knowledge have experienced several modern advances in the past few years, which in part enables the experimentation discussed in further chapters.

2.0.1 Symptom Extraction Using NLP

NLP is a body of research that focuses on enabling computers to process human language via an automated, and thus faster approach. A subfield within NLP is information extraction. The goal of information extraction is to extract concepts ("information") consistently and efficiently through an automated process.

Much of this work focuses on extracting information from textual documents. This typically follows a multi-step process that starts with creating an annotation schema for the concepts of interest. An annotation schema is a formal model of the information at hand. This is similar to ontology work that endeavors to create a consistent frame of reference for all parties[5]. Using this schema, it is relatively straightforward to construct a gold standard corpora for a specific extraction task. With this gold standard in hand, designers can test model architectures of varying complexity and determine their efficacy in a standard and objective process.

There have been many attempts to extract symptoms/signs from medical text. A systematic review found several methods currently in use[10]. The simplest method is to use a "rule-based" approach, such as accomplished by Iqbal et. al. and Greenwald et. al.[10]. These works benefit from simplicity and efficiency, but they lose much of the nuance needed

in order to accurately interpret the information. For example, while a rule based system is sufficient to extract the phrase "pain", further work is needed to determine if pain is negated, if it is associated with the abdomen or the pelvis, or if it is burning or throbbing in nature. Needless to say, in order to create an accurate diagnosis, these details are necessary and so it is insufficient to apply a simpler system.

Another standard approach is to filter the output of a generalized concept extractor, such as MetaMaps, C-Takes, MedLEE or others, for a certain set of concepts related to the symptomatic task at hand. Gundlapalli et. al. and Wang X et. al. are all works that have taken this approach[10]. While this approach takes advantage of prior work and likely requires less up front cost, it does require a clinical expert to determine the concepts to analyze. This can be appropriate, but it also can be inappropriate depending on the downstream task. If the task involves constructing disease definitions, than it can be difficult to determine which concepts to remove and which to keep.

Finally, there have been many recent work endeavoring to extract more nuanced phenomena using sophisticated machine learning methods. Lee et. al., Luan et. al., and Wadden et. al., are all papers that helped developed the framework for generic event extraction with related attributes [13][15][27]. This work was adapted by Kevin Lybarger et. al.[16] to fit a symptomatic context. In Lybarger et. al., a model was built to extract symptoms/signs from medical text, using a large dataset of COVID-19 related patient notes[16]. Crucially, annotated symptoms were not limited to known COVID-19 symptoms, but included all possible symptoms in the notes. These symptoms/signs were extracted as events, with a trigger phrase such as "pain", and related attributes such as "no" (role of assertion) or "abdominal" (role of anatomy). Extracted attributes included the symptom trigger, the assertion, anatomy, characteristics, duration, severity, frequency and change, although the model performance varied depending on the role.

The model created by Lybarger et. al. is span-based; it works end-to-end with a multi-layer extraction model that predicts all event roles in a joint fashion, including the event type, spans, argument types, role linkages, and argument labels if any. It functions by first

encoding sentences using Bio+Clinical BERT, with a bi-LSTM layer to reduce computational cost. A single classifier predicts non-label arguments, while distinct classifiers predict span labels for arguments that include labels. Role based scoring is also performed with separate classifiers for arguments with labels and one boolean classifier for span-only arguments.

The model performance was 0.83 F1 (0.86 F1 annotator agreement) for the symptom trigger, 0.79 F1 (0.83 F1 annotator agreement) for the assertion, and 0.61 F1 (0.81 F1 for annotator agreement) for anatomy. Other features performed less accurately due to the dearth and inconsistency of training data, and so were not prioritized in this work or downstream functions. This model outperformed MetaMapLite++, with a trigger F1 of 0.54 and assertion of 0.44.

Regardless of how symptoms are extracted, the goal is to then use them in a downstream task such as diagnosis prediction. The fundamental goal of automated diagnosis prediction is to reduce the time to diagnosis and prevent diagnostic error. This problem is a well known phenomena in the medical literature, and is especially concerning for cancer. As outlined by Lyratzopoulos et. al., the specificity of cancer symptoms/signs tend to be quite low, and so many patients appear in primary care over multiple consultations before they are eventually diagnosed with cancer[17]. This is because this relatively rare, but serious set of conditions have symptoms/signs that are easily confused with common, less serious conditions.

One paradigm of automated diagnostic prediction focuses on creating a simple index that can be easily used in a clinical setting. Answering a set of questions generates a score for that patient, which might deem the patient high or low risk. This can be a useful method in low resource settings, as it can typically be accomplished with only pen and paper. A study developed by Goff et. al. provides an example of this style of risk scoring[8]. In this work, a questionnaire was developed with scoring logic to create a ovarian cancer risk score. Using questions like “How long did this symptom persist?” and “How many days per month did you experience this symptom?”, providers can ascertain information that may not be uncovered naturally in the patient visit.

An alternative paradigm of diagnostic prediction is using more standard machine learning

approaches to create a classification algorithm. Ayer T. et. al, and Listgarden J. et. al. used artificial neural networks (ANN) and support vector machines (SVM) respectively to predict susceptibility of breast cancer[11]. Waddell M. et. al. performed a similar SVM based experiment on multiple myeloma[11]. SVM, along with Random Forest and Linear Regression are all examples of standard classification algorithms that can be easily applied to clinical data. The input feature set can be driven by lab values, demographic information, genetic markers, or even less quantitative information like symptoms/signs so long as it is properly normalized into a matrix.

An example of this style of approach is Levitsky et. al.[14]. In this work, researchers wished to predict at the time of patients being referred to a lung cancer diagnosis and treatment center whether a patient suspected of lung cancer has lung cancer. To determine this, they used a variety of features including demographic information, symptomatic information, and laboratory values. The gathered this information in a similar fashion as Goff et. al., through a questionnaire. This questionnaire information served as a feature set that was fed into a multivariate regression model. They experimented with different sets of features, achieving sensitivity in the 76.1-84.8 range at the time the patient was referred to the diagnosis treatment center and specificity in the 51.9-66.7 range.

Chapter 3

METHODS

This chapter describes the corpora and datasets that were developed. It also reviews the relevant details of leveraged corpora. Finally, it discusses the methods proper, namely first the annotation adaptation experiments, second the case-control study development and uni-variate analyses, and third the predictive model experimentation.

3.1 Research Datasets

The following sections cover the corpora used to adapt the symptom extraction model developed in prior work, the case-control dataset developed for ovarian cancer, and the leveraged case-control dataset for lung cancer.

3.1.1 NLP Corpora for Symptom Extraction

Over the course of this work, one clinical annotated corpora was leveraged and two corpora were developed. Each corpora describes a separate patient population.

COVID-19 Annotated Text (CACT) Corpus: The CACT was built from clinical notes of 230,000 patients who received treatment at the University of Washington Medical Center (UWM) between May-June 2020. The CACT corpus includes telephone encounter notes, progress notes of all kinds including outpatient and emergency department notes.

Lung Cancer Annotated Text (LACT) Corpus: LACT was built using notes written on lung cancer patients in the 24 months prior to diagnosis. LACT utilized notes from an existing dataset of 4,673 lung cancer patients who were diagnosed with lung cancer at UWM and Seattle Cancer Care Alliance (SCCA) between 2012-2020. LACT includes the following note types: outpatient progress notes, admission notes, emergency and discharge

notes.

Ovarian Cancer Annotated Text (OACT) Corpus: The Ovarian corpus was built using notes written on ovarian cancer patients in the 12 months prior to diagnosis. OACT utilized notes from an existing dataset of 173 ovarian cancer patients who were diagnosed at UWM and SCCA between 2012-2021. OACT included outpatient progress notes, admission, emergency, discharge, and gynecology notes.

Annotations

In prior work, a symptom annotation schema was created[16]. It was designed to generalize across domains and serve as a global representation of symptoms/signs and their nuanced event attributes. This annotation schema was designed in an "event" structure, where one role, the trigger, is the nexus of the symptom, while other roles serve as nuance providing detail.

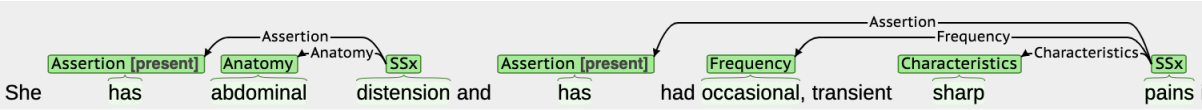

Figure 3.1: Annotation Schema Example

In Table 3.1, there are two required roles, assertion and trigger. The trigger is what makes the symptom unique, and can be either a single word such as "pain" or an atomic phrase such as "short of breath". Other, optional, roles are less common and so less represented in the corpus. Figure 3.1 shows how symptoms/signs can have related anatomy ("abdominal"), characteristics ("sharp"), and frequency ("occasional"). Some of these roles, like assertion, have values associated, while other roles, like anatomy, only have spans associated with the role.

Table 3.1: Symptom Annotation Schema

Role	Labels	Examples
Trigger (required)	None	“pain”, “cough”
Assertion (required)	present, absent, hypothetical	“reports”, “no”, “prn”
	not patient, possible, conditional	
Anatomy	None	“abdominal”, “chest”
Characteristics	None	“watery”, “pink”
Severity	mild, moderate, severe	“mild”, “severe”
Change	improving, no change, worsened, resolved	“worsening”, “increasing”
Duration	None	“for 3 days”
Frequency	None	“every day”

Table 3.2: Corpora Comparison

Characteristics	CACT	LACT	OACT
Dataset Type	Baseline	New	New
Training Size (# notes)	1028	170	110
Test Size (# notes)	444	100	100
Annotator Agreement (Trigger F1)	0.85	0.83	0.82
Annotator Agreement (Assertion F1)	0.83	0.79	0.80
Annotator Agreement (Anatomy F1)	0.8	0.79	0.78
Annotator Agreement (Change F1)	0.51	0.62	0.37
Annotator Agreement (Severity F1)	0.41	0.47	0.38
Annotator Agreement (Frequency F1)	0.65	0.38	0.21
Annotator Agreement (Duration F1)	0.56	0.30	0.55
Annotator Agreement (Characteristics F1)	0.53	0.45	0.41

To determine annotator agreement and model performance, three different approaches were used, depending on the type of the role. For the trigger, a true positive was defined as an exact index match between trigger in the gold and test corpus. For a value role, such as assertion or change, an exact match on the trigger as well as a value match on the role (present = present, for example) were both required to count as a true positive. Finally, for span-only roles such as anatomy, a partial overlap on the anatomy span as well as an exact match on the trigger was sufficient for a true positive.

Table 3.2 compares the new corpora with the existing corpus for key differences. The relatively small sizes of training data in LACT and OACT as compared to CACT, 170 and 110 compared to 1028 respectively, are due to the smaller dataset requirements needed to do transfer learning. The annotator agreements were similar across corpora, between 0.82 and 0.85 trigger level F-1. This agreement serves as a ceiling for model performance, as trigger accuracy is essential for any other role to be correct.

3.1.2 Case-Control Datasets for Ovarian Cancer and Lung Cancer

As part of this work, one case-control dataset was developed, for ovarian cancer, and one case-control dataset was leveraged, for lung cancer.

Ovarian Cancer Case-Control Dataset (Ovarian Cancer Cohort): The Ovarian Cancer Cohort was created using data from University of Washington Medicine (UWM), which is an academic health science center comprised of ambulatory care clinics, urgent care facilities, emergency departments, and hospitals primarily serving western Washington, United States of America (USA). Electronic medical record systems were in use in both the inpatient and outpatient settings for the duration of the study; data is stored centrally in an integrated data repository. The UWM hospitals and associated laboratories all participate in the Cancer Surveillance System/Surveillance, Epidemiology and End Results (CSS/SEER) tumor registry, which includes data on all the newly-diagnosed cancers (except non-melanoma skin cancers) in 13 western Washington counties. The Institutional Review Board (IRB) of the University of Washington approved this research study.

The Ovarian Cancer Cohort was developed to describe the pre-diagnosis clinical presentation of patients with ovarian cancer in ambulatory settings compared to control patients without a diagnosis of ovarian cancer that were seen at a similar ambulatory care location. Cases were patients aged 18 years or greater at date of diagnosis, with a first, primary ovarian cancer diagnosed between Jan 1, 2012 and Dec 31, 2020, with an established relationship with UWM ambulatory care prior to the diagnosis of cancer. The logistics of such a case selection are described below. For the purposes of this study, ovarian cancer patients were initially defined as patients with at least one of the relevant ICD-10/ICD-9 codes seen in Table 3.3.

Table 3.3: Ovarian Case Inclusion Diagnoses/Control Exclusion Diagnoses

ICD-9	ICD-10
183.0 Malignant neoplasm of ovary	C56 Malignant neoplasm of ovary C56.1 Malignant neoplasm of right ovary C56.2 Malignant neoplasm of left ovary C56.9 Malignant neoplasm of unspecified ovary
183.2 Malignant neoplasm of the fallopian tube	C57.0 Malignant neoplasm of the fallopian tube C57.00 Malignant neoplasm of unspecified fallopian tube C57.01 Malignant neoplasm of right fallopian tube C57.02 Malignant neoplasm of left fallopian tube
158.8 Malignant neoplasm of specified parts of peritoneum	C48.1 Malignant neoplasm of specified parts of the peritoneum
158.9 Malignant neoplasm of peritoneum, unspecified	C48.2 Malignant neoplasm of the peritoneum, unspecified
236.2 Neoplasm of uncertain or unknown behaviour of the ovary	D39.1 Neoplasm of uncertain or unknown behaviour of the ovary

This preliminary cohort was shared securely with CSS/SEER through identifiable information on identified cases, such as name, date of birth, Medical Record Number (MRN), and an ID generated for this study. CSS/SEER matched cancer registry records with primary ovarian cancer tumors to this cohort. Requested data fields were extracted and de-identified by CSS/SEER. CSS/SEER de-identified data linked only by the UWM study ID was returned

to the UW research team. Date of diagnosis of ovarian cancer was defined as the date of diagnosis listed in CSS/SEER (in most cases, the date of histologic/cytologic confirmation). The patient cohort was then filtered by the following criteria.

First, patients without notes in the year prior to the CSS/SEER diagnosis date were removed. Many patients are referred into the UW medical system upon diagnosis of cancer. Since only patients already active within the UW medical system are useful for this study, this filter was used to remove many of these "referral" patients.

Second, patients without a documented recommendation showing evidence of suspected ovarian cancer (a CT, an ultrasound (US), or a referral to Gynecology) were removed. Documented recommendations were extracted from discrete order data or using NLP. Similar to symptoms/signs, test or referral recommendations can be far more present in the note compared to discretely recorded orders. The model used was developed by Wilson et. al. to classify sentences within radiology notes as recommendations. The recommendations extractor is a binary classifier implemented with Hierarchical Attention Network[12]. The model learns to identify a recommendation based on the aggregated attention weights of word contexts within the sentences. The model was trained on a multi-modal radiology corpus and achieved 0.92 F1[12].

The clinic or provider type suggesting the initial recommendation was mapped to assign the patient to a specific diagnosis route. This was accomplished through meta data about the provider specialty and the location where the notes were written (encounter facility, department specialty). The provider specialty and department were mapped to one of four categories: 1) Primary Care, which includes Family Medicine, Internal Medicine, International Medicine, and Urgent Care, 2) Women's Health, which includes Women's Health and Obstetrics and Gynecology, 3) Emergency, which includes Emergency Medicine, and 4) Other, which includes all other encounter or facility types. The earliest mappable recommendation in the year prior to diagnosis was used to determine the patient's route.

The result of this filtering led to a cohort of size N=136. Controls were randomly selected from an eligible pool using a 10:1 matched sample. Women were eligible for selection as a

control if they had a birthday within 3 years of the case patient, had at least one visit to the same type of route as the patient, had no history of ovarian, fallopian, or peritoneal cancer (using the ICD codes noted in Table 3.3), and no history of bilateral salpingo-oophorectomy Z90.722 (ICD-10) and 656 (ICD-9). The visit to the same route as the patient served as the lookback date for the data pull, similar to the diagnosis date for the patient. No other exclusions were used. Figure 3.2 explores the patients removed at each step.

For both control and case patients, the symptom extractor extracted all notes in the year prior to the relevant lookback date, as well as any orders, labs, radiology reports, and patient weight. The notes were processed by a symptom extraction model, created in other work and tuned to the ovarian cancer context, as well as a recommendation extraction model, used as is.

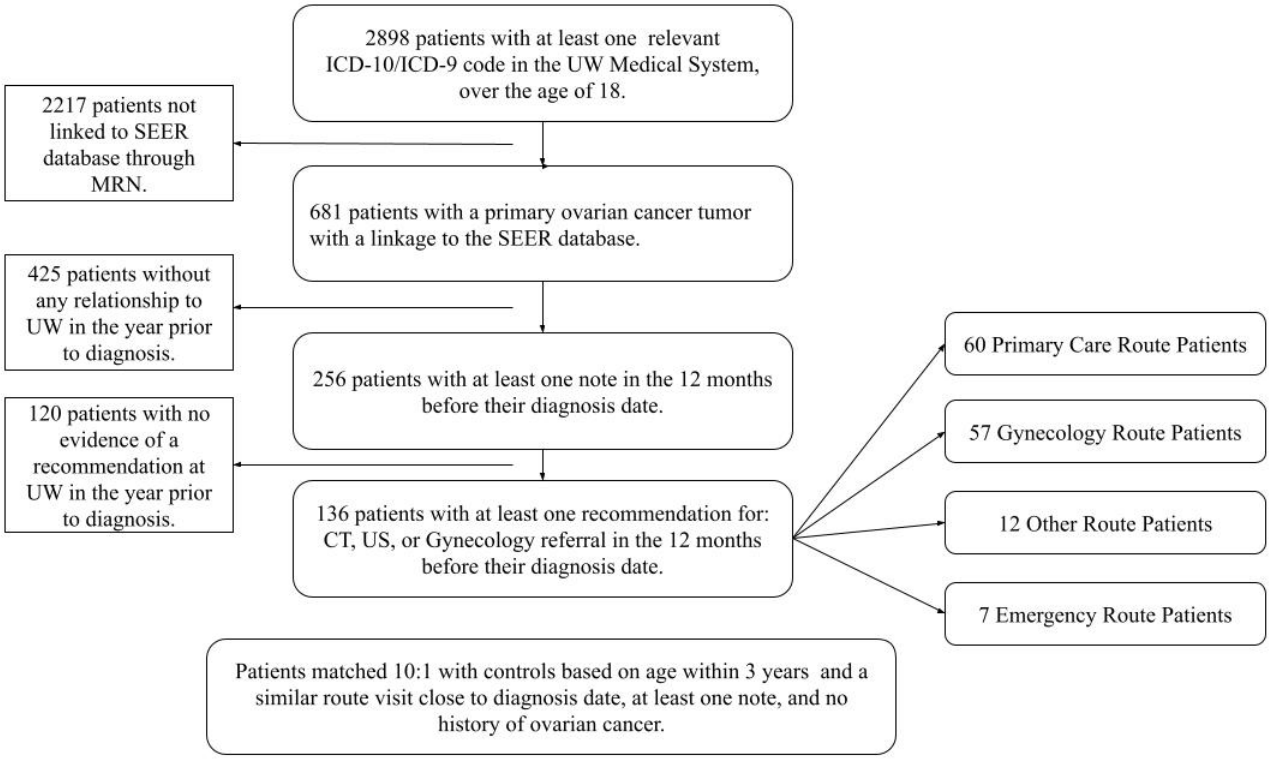

Figure 3.2: Case-Control Definition for the Ovarian Cancer Cohort

Lung Cancer Case-Control Dataset (Lung Cancer Cohort): The Lung Cancer Cohort was developed in prior, as of yet unpublished work by Zigman Suchsland et. al. in the author's lab. As part of that unpublished work, uni-variate analyses were conducted and a case-control cohort was developed. Similar to the Ovarian Cancer Cohort, this was created with patients attending and diagnosed through UWM, linked to CSS/SEER records, and then mapped to control patients. The case cohort was filtered differently, but the intention was identical, with the goal to remove patients who were referred to UWM for lung cancer treatment. In this cohort, there were 711 case patients and 6841 control patients, or approximately 10:1 matched sample. Thus, there are several key differences between the lung and ovarian cancer cohorts. The first is that lung cancer patients are far more numerous and the resulting dataset size is approximately 5 times larger than the Ovarian Cancer Cohort. The second is that the patients are not further subdivided into different routes based on referral patterns. Both of these facts have potential downstream impacts when performing the same experiments in the two different contexts.

3.2 Experimental Design

The experiments are divided into three main sections. The sections are as follows: the adaptation of the symptom extraction model, the uni-variate case-control analysis of ovarian cancer, and finally the prediction task.

Adaptation of a Symptom Extraction Model to Two Separate Corpora: The original model created by Lybarger et. al. was developed solely on the CACT corpus. As part of this work, a series of experiments was run. First, the model was trained from scratch solely on the CACT training set, and then tested it on the CACT, LACT, and OACT test set. These experiments served as a baseline of performance without domain adaptation. Subsequently, in-domain training data was included with the out-domain training data, and then the performance on the respective in-domain test set was considered. The following experiment combined CACT and LACT training data and tested it on the Lung test set. The final experiment combined CACT and OACT training data and tested the new model on

the OACT test set. In all experiments the model was retrained from scratch, but otherwise no other optimizations of the model's hyperparameters beyond the parameters derived from Lybarger et. al. were performed. The goal of this experimentation was to determine the limitations of the dataset itself, and so the experimentation was designed to avoid tuning beyond the inclusion of in-domain or out-domain data.

For the domain adaptation experimentation, only trigger level performance was considered. This is due to the nature of the annotation schema. Without the trigger being accurate, no other role can be accurate, and so the priority is solely trigger performance in this experimentation. Finally, as part of the error analysis the impacts of span level trigger performance changes were explored.

Uni-variate Analysis of Symptoms and Signs in Ovarian Cancer: This part of the analysis first explored the differences in symptomatic expression between mapped ICD-10 codes and normalized note-extracted symptoms/signs found in the Ovarian Cancer Cohort. First, the model adapted using the OACT corpus extracted un-normalized symptomatic events. Then, the unique trigger and anatomy spans were normalized to one of the core symptoms/signs or relevant anatomy that were known to be affected by ovarian cancer. Core symptoms/signs were symptoms that appeared in one or more studies on ovarian cancer. If the span was not known to be related, it was mapped as Unknown. For complex symptoms/signs such as "abdominal pain", a normalized symptom was found if it had both a trigger mapped to the normalized trigger "Pain" ("pain", "painful", "hurts"), and a linked anatomy role that mapped to the normalized anatomy region "Abdominal Region" ("abd", "abdomen", "abdominal"). In very rare cases, where the trigger was one word and implied both symptom and anatomy, it was mapped directly (e.g. "spotting", which uniformly refers to bleeding from the vaginal region).

This set of core symptoms/signs can be seen in Table 3.4. Uni-variate analyses were performed for both symptomatic and demographic differences between cases and controls as well as between early and late stage cancer. Finally, uni-variate symptomatic analyses were performed across time for both the overall Ovarian Cancer Cohort and the two larger route

sub-cohorts, Primary Care and Gynecology.

Table 3.4: Ovarian Cancer Core Symptoms and Signs

Symptom (References)	ICD-9, ICD-10 Codes	Textual Examples
Pain-Abdominal Region[7][1][25]	789.0, 789.01, 789.03, 789.04, 789.09, R10.9, R10.1, R10.10, R10.3, R10.30, R10.8, R10.84, R10.9	"abdominal pain", "pain in the abdomen"
Pain-Pelvic Region[7]	789.5, 789.51, 789.59, R18, R18.8, R18.0	"pelvic pain", "pain in the pelvic area"
Ascites[7]	789.5, 789.51, 789.59, R18, R18.8, R18.0	"ascites"
Bleeding-Vaginal Area[1][9][25]	N95.0, N93.9, N93.8, N93, N92.4, N92.3, 627.0, 627, 627.1, 626.7, 626.5	"vaginal bleeding", "spotting"
Nausea	787.02, 787.01, 787.04, R11.0, R11, R11.2, R11.14	"nausea"
Fatigue[7]	780.79, R53, R53.0, R53.1, R53.8, R53.81, R53.82,R53.83	"tired", "fatigue"
Distention[7][1][9][25]	787.3, R14.0, R14.1, R14.2, R14.3	"distended", "distention"
Bloating[7][1][9][25]	787.3, R14.0, R14.1, R14.2, R14.3	"bloated"
Weight Loss[7]	783.2, R63.4, 783.21	"weight loss"
Change in Bowel Habits[7][25][9]	787.99, 564.5, 564.00, 564.01, 564.02, R19.4, 787.91, K59.1, R19.7, 564.0, K59.09, K59, K59.04, K59.01, K59.02, K59.0	"loose bms", "constipation"
Loss of Appetite[7][1]	783.0, R63.0	"lack of appetite"
Early Satiety[7][1][25]	780.94, R68.81	"feeling full"
Urinary Frequency[7][9][25]	788.41, R35, R35.1, R35.0, R35.8	"urinary frequency"
Incontinence-Urinary[7]	R32, 788.30, 788.31, 788.33, 788.38, 625.6, 788.39, R39.81, 788.91	"leaking"
Urinary Urgency[7][25]	788.63, R39.15	"urinary urgency"
Dysuria[7]	R30.0, 788.1	"painful urination"
Masses-Abdominal Region[7][9]	R19.01, R19.02, R19.03, R19.04, R19.00, R19.09, R19.05, 789.31, 789.32, 789.33, 789.34, 789.30, 789.35, 789.37, 789.39, R19.00, R19.07, R19.09	"abdominal mass"
Masses-Pelvic Region[7][9]	R19.00, R19.09, R19.07, 789.3, 789.30, 789.39	"pelvic mass"

3.2.1 Predictive Model Experimentation

Three distinct experiments of disease classification using both the Ovarian Cancer Cohort and Lung Cancer Cohort were performed. The primary goal was to determine the potential for early cancer prediction prior to diagnosis, but also to explore the effect that different data types and models might have on the predictive task.

Each patient's year prior to diagnosis was represented as an input row in the training and test sets. The output value was a binary indication of whether the patient belonged to the case or control cohort. Thus, the classification task was to determine which cohort the patient belonged given the input features. The input features were modified across the experiments, but they were all variations of the core symptoms/signs for each cancer generated from a literature review. Fundamentally, the classification task was to differentiate between cancer and matched controls at different points in time, most commonly at the time of diagnosis.

In the first set of experiments, performance comparisons were performed with random forest, SVM, and linear regression models on the same datasets with the goal of using the model type with the highest sensitivity in the following, more crucial, experiments. In the experimentation, each patient was represented by the set of core, normalized symptoms/signs for ovarian cancer. These symptoms/signs were driven by prior literature on symptoms/signs potentially correlated with either cancer. Given the year prior to diagnosis, a patient would have a 1 if the present symptom/sign appeared at least once in the year prior to diagnosis. There would be a 0 for that feature if the present symptom/sign did not appear. A true positive of the positive class, cases, would indicate that the model predicted the positive class using the feature set for that patient. NLP-only, ICD-10 only, and features that included both NLP and ICD-10 expression were considered as sub-categories. The features were the normalized, present symptoms/signs of ovarian cancer. The effect of the inclusion of NLP was analyzed through these deviations.

The second series of experiments considered the difference in model performance with the addition of absent symptomatic information that can be observed using the NLP extracted

symptoms/signs. Changes in performance were evaluated with the inclusion of explicitly absent symptoms/signs. Each symptom/sign was represented by two features: a feature for explicit presence, and a feature for explicit absence. A patient would receive a one in the present feature if a provider explicitly indicated presence at least once in the time period, and a one in the absent feature if a provider explicitly indicated absence at least once in the time period. Thus, a patient could both have a positive indication in explicitly absent and explicitly present features for the same symptom/sign. Only symptoms/signs extracted using NLP were included, as discrete data for the Lung Cancer Cohort was not available for these experiments. This was done in order to appropriately compare across the two cancers.

The third series of experiments considered how model performance changed over different time series filters, given the normalized symptoms/signs. 5 alternative filters were applied to both the Lung and Ovarian CAncer Cohorts. Input data was developed given the full year, only the data from 30-365 days (1 Month) before diagnosis, 60-365 days (2 Months) before diagnosis, 90-365 days (3 Months) before diagnosis, 180-365 days (6 Months) before diagnosis, and 270-365 (9 Months) before diagnosis. Thus, the classification task changed slightly with each dataset, namely predicting cancer at different time points instead of at the date of diagnosis. Fundamentally, the goal was to determine the potential for prediction as an early warning system for providers during the initial diagnostic process.

Training and test splits were performed by randomly selecting 20% of the respective input set and setting it aside as the test dataset. Models were first trained using 10-fold cross validation (CV). The best underlying designs found in 10-fold CV were used in the following experiments on the test dataset.

Chapter 4

RESULTS

This chapter is divided into three major sections, coinciding with the three major experimental themes of this work. First, the results of the adaptation of the symptom extraction model are discussed. Second, the results of the ovarian cancer uni-variate analyses are reviewed. Third, the results of the classification task predicting cancer diagnoses at different time points are provided.

4.1 Adaptation of a Symptom Extraction Model to Two Separate Corpora

This set of experiments focuses on the adaptation of a symptom extraction model to two out-domain contexts. The results of the different experiments can be seen below in Table 4.1. When changing to an out-domain, the model trained only on CACT achieves a precision change of -0.01-0.06, with a recall drop of 0.13-0.14. With the addition of in-domain training data, recall improves from the prior nadir by 0.11-0.12. This recall drop seems to drive the overall F1 drop prior to the inclusion of in-domain training data.

Table 4.1: Domain Adaptation Performance for Symptom Trigger Extraction.

Training	Test	P	R	F1
CACT	CACT	0.76	0.76	0.76
CACT	LACT	0.75	0.62	0.68
CACT	OACT	0.82	0.63	0.71
CACT, LACT	LACT	0.71	0.73	0.72
CACT, OACT	OACT	0.82	0.75	0.79

Table 4.2: Top 15 Symptoms/Signs with Largest Recall Delta Impact. TP represents true positives. Key symptoms/signs for lung cancer and ovarian cancer are in bold.

Most Improved Triggers in LACT		Most Improved Triggers in OACT	
Trigger Span	Δ TP	Trigger Span	Δ TP
pain	18 (261 → 279)	**pain**	30 (285 → 315)
constipation	9 (14 → 23)	**nausea**	27 (103 → 130)
gallops	8 (1 → 9)	**vomiting**	23 (73 → 96)
lesions	7 (2 → 9)	masses	19 (4 → 23)
masses	6 (0 → 6)	neuropathy	19 (3 → 22)
problems	6 (1 → 7)	**ascites**	14 (1 → 15)
cyanosis	6 (5 → 11)	murmur	11(0 → 11)
edema	5 (40 → 45)	**bleeding**	9 (22 → 31)
rubs	5 (0 → 5)	rash	8 (11 → 19)
coughing	5 (3 → 8)	incontinence	7 (7 → 14)
weight loss	5 (22 → 27)	murmurs	7 (0 → 7)
adenopathy	5 (0 → 5)	lesions	7(14 → 21)
murmurs	5 (0 → 5)	**distended**	6 (10 → 16)
suicidal ideation	4 (1 → 5)	**tenderness**	6 (24 → 30)
rash	4 (37 → 41)	ulcers	6 (11 → 17)

There are two main patterns when observing the most changed triggers between the model trained solely on out-domain data and the model trained on both in- and out-domain data. First, triggers tend to be common symptoms/signs such as pain that appear in new, uncommon contexts, such as cancer pain management. Second, most improved triggers tend to be uncommon symptoms/signs that appear more commonly in the new domain, such as bleeding or adenopathy. In Table 4.2, symptom triggers that have been found in the literature to be key symptoms/signs for their relevant domains are in bold. To note is how many of these symptoms/signs are represented in the top 15 most changed symptoms/signs, and thus

would be under-represented without in-domain training data. Not all of these triggers are necessarily symptoms associated with the out-domain, but the fact that many of them are indicates the differences between in- and out-domain data.

Table 4.3: Final Adapted Model Performance on Symptomatic Roles

Roles	LACT F1	OACT F1
Trigger	0.72	0.78
Assertion	0.65	0.73
Anatomy	0.61	0.62
Change	0.06	0.16
Severity	0.36	0.45
Frequency	0.59	0.52
Duration	0.27	0.45
Characteristics	0.26	0.27

The performance on all roles for the adapted models are recorded in Table 4.3. These were the models used in the following uni-variate analysis and early diagnostic prediction.

4.2 Uni-variate Analysis of Symptoms and Signs in Ovarian Cancer

In this section the results from the uni-variate analyses of ovarian cancer symptoms/signs in the year prior to diagnosis are reviewed. Comparisons include cases and controls, early and late stage cancer, and changes caused by the inclusion of different symptomatic data types.

Changes in Patient Coverage Due To The Inclusion of NLP Extracted Core Symptoms and Signs for Ovarian Cancer

In Table 4.4, the combination of both ICD and NLP leads to a higher frequency than either alone, although for many of the core symptoms/signs NLP significantly outperforms ICD in recall, with little net benefit from the inclusion of ICD. This lack of benefit is true for

symptoms/signs such as Distention, Loss of Appetite, and Early Satiety. Only two symptoms/signs, Masses-Abdominal Region and Masses-Pelvic Region, achieves a higher incidence with ICD than with NLP alone in the cases of the Ovarian Cancer Cohort.

	Case			Control		
Symptom	ICD, NLP	ICD	NLP	ICD, NLP	ICD	NLP
Ascites	23.5% (32)	14.0% (19)	16.9% (23)	1.4% (19)	1.2% (17)	0.6% (8)
Bleeding-Vaginal Area	41.2% (56)	15.4% (21)	38.2% (52)	15.4% (209)	7.1% (97)	11.5% (156)
Bloating	52.2% (71)	10.3% (14)	50.7% (69)	9.6% (131)	3.4% (46)	7.2% (98)
Change in Bowel Habits	73.5% (100)	20.6% (28)	71.3% (97)	37.8% (514)	16.2% (221)	30.7% (417)
Distention	55.1% (75)	10.3% (14)	52.2% (71)	7.4% (101)	3.4% (46)	4.7% (64)
Dysuria	12.5% (17)	8.8% (12)	6.6% (9)	14.2% (193)	11.1% (151)	5.6% (76)
Early Satiety	27.2% (37)	2.2% (3)	26.5% (36)	2.4% (33)	0.5% (7)	2.0% (27)
Fatigue	69.9% (95)	23.5% (32)	66.2% (90)	43.2% (588)	19.3% (263)	34.7% (472)
Incontinence-Urinary	16.2% (22)	5.1% (7)	12.5% (17)	12.3% (167)	7.4% (100)	7.4% (101)
Loss of Appetite	46.3% (63)	0.7% (1)	46.3% (63)	11.8% (161)	1.4% (19)	11.4% (155)
Masses-Abdominal Region	62.5% (85)	58.8% (80)	27.9% (38)	4.8% (65)	2.6% (36)	2.5% (34)
Masses-Pelvic Region	60.3% (82)	58.1% (79)	35.3% (48)	2.9% (39)	2.5% (34)	0.8% (11)
Nausea	67.6% (92)	17.6% (24)	61.8% (84)	34.1% (464)	9.7% (132)	30.0% (408)
Pain-Abdominal Region	76.5% (104)	30.9% (42)	74.3% (101)	26.8% (364)	13.1% (178)	18.7% (254)
Pain-Pelvic Region	41.9% (57)	29.4% (40)	27.2% (37)	21.3% (290)	17.3% (235)	7.4% (100)
Urinary Frequency	23.5% (32)	14.7% (20)	16.2% (22)	11.0% (149)	8.8% (120)	3.9% (53)
Urinary Urgency	5.9% (8)	2.2% (3)	5.1% (7)	5.9% (80)	3.7% (50)	2.6% (35)
Weight Loss	31.6% (43)	4.4% (6)	30.9% (42)	14.8% (201)	4.4% (60)	11.8% (161)

Table 4.4: Frequency of Different Core Symptoms and Signs, Ovarian, ICD vs. NLP

In Table 4.5, some symptoms/signs only achieve significant odds ratios with the inclusion of NLP, namely Change in Bowel Habits, Early Satiety, Fatigue, Incontinence-Urinary, Loss of Appetite, and Weight Loss. Even in cases where a symptom is significant without NLP, the inclusion typically increases the odds ratios. In some cases, it increases it to the point that the confidence intervals no longer overlap, such as Distention and Pain-Abdominal Region. Only in two cases do the odds ratios drop slightly, namely Masses-Abdominal Region and

Masses-Pelvic Region, although the confidence intervals continue to overlap.

Table 4.5: Odds Ratios for Cases vs. Controls, Comparison of Different Data Sources

	ICD, NLP		ICD		NLP	
Symptom	Odds Ratio	P-Value	Odds Ratio	P-Value	Odds Ratio	P-Value
Ascites	21.7 (11.9-39.6)	**0.001**	12.8 (6.5-25.4)	**0.001**	34.4 (15.0-78.7)	< **0.001**
Bleeding-Vaginal Area	3.9 (2.7-5.6)	< **0.001**	2.4 (1.4-4.0)	**0.002**	4.8 (3.3-7.0)	< **0.001**
Bloating	10.2 (7.0-15.0)	< **0.001**	3.3 (1.8-6.1)	**0.001**	13.3 (8.9-19.7)	< **0.001**
Change in Bowel Habits	4.6 (3.1-6.8)	< **0.001**	1.3 (0.9-2.1)	0.226	5.6 (3.8-8.3)	< **0.001**
Distention	15.3 (10.3-22.7)	< **0.001**	3.3 (1.8-6.1)	**0.001**	22.1 (14.5-33.7)	< **0.001**
Dysuria	0.9 (0.5-1.5)	0.698	0.8 (0.4-1.4)	0.473	1.2 (0.6-2.4)	0.562
Early Satiety	15.0 (9.0-25.1)	< **0.001**	4.4 (1.1-17.1)	0.055	17.8 (10.4-30.5)	< **0.001**
Fatigue	3.0 (2.1-4.5)	< **0.001**	1.3 (0.8-2.0)	0.258	3.7 (2.5-5.3)	< **0.001**
Incontinence-Urinary	1.4 (0.8-2.2)	0.222	0.7 (0.3-1.5)	0.483	1.8 (1.0-3.1)	**0.044**
Loss of Appetite	6.4 (4.4-9.4)	< **0.001**	0.5 (0.1-3.9)	1.0	6.7 (4.6-9.8)	< **0.001**
Masses-Abdominal Region	33.2 (21.7-50.9)	< **0.001**	52.5 (32.7-84.5)	< **0.001**	15.1 (9.1-25.1)	< **0.001**
Masses-Pelvic Region	51.4 (32.2-82.2)	< **0.001**	54.1 (33.4-87.5)	< **0.001**	66.9 (33.6-133.3)	< **0.001**
Nausea	4.0 (2.8-5.9)	< **0.001**	2.0 (1.2-3.2)	**0.007**	3.8 (2.6-5.4)	< **0.001**
Pain-Abdominal Region	8.9 (5.9-13.5)	< **0.001**	3.0 (2.0-4.4)	< **0.001**	12.6 (8.4-18.9)	< **0.001**
Pain-Pelvic Region	2.7 (1.8-3.8)	< **0.001**	2.0 (1.3-3.0)	**0.001**	4.7 (3.1-7.2)	< **0.001**
Urinary Frequency	2.5 (1.6-3.8)	< **0.001**	1.8 (1.1-3.0)	**0.03**	4.8 (2.8-8.1)	< **0.001**
Urinary Urgency	1.0 (0.5-2.1)	1.0	0.6 (0.2-1.9)	0.623	2.1 (0.9-4.7)	0.097
Weight Loss	2.7 (1.8-3.9)	< **0.001**	1.0 (0.4-2.4)	1.0	3.3 (2.2-5.0)	< **0.001**

4.2.1 Differences in Demographics Between Cases and Controls and Early and Late Stage Ovarian Cancer

The Ovarian Cancer Cohort consists of 136 cases and 1360 matched controls. Of the 136 cases, there were 2 patients with Stage 0, 30 patients with Stage 1, 11 patients with Stage 2, 70 patients with Stage 3, and 16 patients with Stage 4. There were 7 patients with no staging information available.

Table 4.6: Demographic Analysis of Cases vs. Controls, Ovarian Cancer

Characteristic	Cases	Controls
Age	60.7	60.3
Race-Ethnicity: African American	8 (5.9%)	95 (7.0%)
Race-Ethnicity: American Indian or Alaska Native	2 (1.5%)	12 (0.9%)
Race-Ethnicity: Asian	13 (9.6%)	108 (7.9%)
Race-Ethnicity: Native Hawaiian or Other Pacific Islander	1 (0.7%)	10 (0.7%)
Race-Ethnicity: Other	1 (0.7%)	8 (0.6%)
Race-Ethnicity: Unknown	3 (2.2%)	158 (11.6%)
Race-Ethnicity: White	104 (76.5%)	927 (68.2%)
Race-Ethnicity: White (Hispanic)	4 (2.9%)	42 (3.1%)
Smoking Status: Former or Current Smoker	37 (27.2%)	336 (24.7%)
Smoking Status: Never Smoker	58 (42.6%)	642 (47.2%)
Smoking Status: Unknown	41 (30.1%)	382 (28.1%)
Mean Number of Consultation Days Per Patient: (-10, 366) Days before Diagnosis	18.83	13.62
Mean Number of Consultation Days Per Patient: (-10, 89) Days before Diagnosis	11.17	6.03
Mean Number of Consultation Days Per Patient: (-10, 29) Days before Diagnosis	8.63	3.41
Mean Number of Consultation Days Per Patient: (30, 59) Days before Diagnosis	1.46	1.44
Mean Number of Consultation Days Per Patient: (60, 89) Days before Diagnosis	1.08	1.18
Mean Number of Consultation Days Per Patient: (90, 179) Days before Diagnosis	2.38	2.81
Mean Number of Consultation Days Per Patient: (180, 269) Days before Diagnosis	2.94	2.42
Mean Number of Consultation Days Per Patient: (270, 366) Days before Diagnosis	2.35	2.36
Elixhauser Comorbidity Mean	3.36	1.61

In Table 4.6, case patients seem to have a higher Elixhauser comorbidity mean than the control patients. There is also a sharper increase in the number of consultation days in the last 30 days prior to diagnosis compared to controls, while between 180-365 days there are far fewer differences between cases and controls. This sparcity of data past 30 days is explored in a later section.

Table 4.7: Demographic Analysis of Early vs. Late Stage Ovarian Cancer

Characteristic	Stage 1-2	Stage 3-4
Age	61.7	60.1
Race-Ethnicity: African American	0 (0.0%)	8 (9.3%)
Race-Ethnicity: American Indian or Alaska Native	1 (2.3%)	1 (1.2%)
Race-Ethnicity: Asian	6 (14.0%)	7 (8.1%)
Race-Ethnicity: Native Hawaiian or Other Pacific Islander	0 (0.0%)	1 (1.2%)
Race-Ethnicity: Other	0 (0.0%)	1 (1.2%)
Race-Ethnicity: Unknown	0 (0.0%)	2 (2.3%)
Race-Ethnicity: White	36 (83.7%)	62 (72.1%)
Race-Ethnicity: White (Hispanic)	0 (0.0%)	4 (4.7%)
Smoking Status: Former or Current Smoker	11 (25.6%)	24 (27.9%)
Smoking Status: Never Smoker	16 (37.2%)	38 (44.2%)
Smoking Status: Unknown	16 (37.2%)	24 (27.9%)
Mean Number of Consultation Days Per Patient: (-10, 366) Days before Diagnosis	19.65	18.45
Mean Number of Consultation Days Per Patient: (-10, 89) Days before Diagnosis	11.77	11.08
Mean Number of Consultation Days Per Patient: (-10, 29) Days before Diagnosis	8.65	8.78
Mean Number of Consultation Days Per Patient: (30, 59) Days before Diagnosis	1.72	1.3
Mean Number of Consultation Days Per Patient: (60, 89) Days before Diagnosis	1.4	1.0
Mean Number of Consultation Days Per Patient: (90, 179) Days before Diagnosis	2.02	2.48
Mean Number of Consultation Days Per Patient: (180, 269) Days before Diagnosis	3.79	2.48
Mean Number of Consultation Days Per Patient: (270, 366) Days before Diagnosis	2.07	2.42
Elixhauser Comorbidity Mean	2.74	3.62

Table 4.7 provides similar demographic patterns as Table 4.6, with two exceptions. First, patients diagnosed with later stage cancer seem to generally have a higher Elixhauser comorbodity mean. Second, although the size of the cohort is small, it is important to note that all 12 Hispanic and African American patients are diagnosed with later stage cancer.

4.2.2 Odds Ratios Comparing Cases and Controls and Early and Late Stage Ovarian Cancer, Entire Year Prior to Diagnosis

Table 4.8 shows significant differences (in bold) between cases and controls when observing patients with at least one symptom in ICD or NLP in the year prior to diagnosis. This is true across all symptoms/signs, with the exception of Dysuria, Urinary Urgency and Incontinence-Urinary.

Table 4.8: Odds Ratios for Cases vs. Controls (merged NLP and ICD Codes)

Symptom	Case	Control	Odds Ratio	P-Value
Masses-Pelvic Region	60.3% (82)	2.9% (39)	51.4 (32.2 - 82.2)	< **0.001**
Masses-Abdominal Region	62.5% (85)	4.8% (65)	33.2 (21.7 - 50.9)	< **0.001**
Ascites	23.5% (32)	1.4% (19)	21.7 (11.9 - 39.6)	< **0.001**
Distention	55.1% (75)	7.4% (101)	15.3 (10.3 - 22.7)	< **0.001**
Early Satiety	27.2% (37)	2.4% (33)	15.0 (9.0 - 25.1)	< **0.001**
Bloating	52.2% (71)	9.6% (131)	10.2 (7.0 - 15.0)	< **0.001**
Pain-Abdominal Region	76.5% (104)	26.8% (364)	8.9 (5.9 - 13.5)	< **0.001**
Loss of Appetite	46.3% (63)	11.8% (161)	6.4 (4.4 - 9.4)	< **0.001**
Change in Bowel Habits	73.5% (100)	37.8% (514)	4.6 (3.1 - 6.8)	< **0.001**
Nausea	67.6% (92)	34.1% (464)	4.0 (2.8 - 5.9)	< **0.001**
Bleeding-Vaginal Area	41.2% (56)	15.4% (209)	3.9 (2.7 - 5.6)	< **0.001**
Fatigue	69.9% (95)	43.2% (588)	3.0 (2.1 - 4.5)	< **0.001**
Pain-Pelvic Region	41.9% (57)	21.3% (290)	2.7 (1.8 - 3.8)	< **0.001**
Weight Loss	31.6% (43)	14.8% (201)	2.7 (1.8 - 3.9)	< **0.001**
Urinary Frequency	23.5% (32)	11.0% (149)	2.5 (1.6 - 3.8)	< **0.001**
Incontinence-Urinary	16.2% (22)	12.3% (167)	1.4 (0.8 - 2.2)	0.222
Urinary Urgency	5.9% (8)	5.9% (80)	1.0 (0.5 - 2.1)	1.0
Dysuria	12.5% (17)	14.2% (193)	0.9 (0.5 - 1.5)	0.698

Table 4.9: Odds Ratios in Stage 0-2 vs. Stage 3-4 (merged NLP and ICD Codes)

Symptom	Stage 0-2	Stage 3-4	Odds Ratio	P-Value
Dysuria	18.6% (8)	8.1% (7)	2.6 (0.9 - 7.7)	0.09
Bleeding-Vaginal Area	58.1% (25)	36.0% (31)	2.5 (1.2 - 5.2)	**0.023**
Distention	60.5% (26)	53.5% (46)	1.3 (0.6 - 2.8)	0.573
Masses-Pelvic Region	65.1% (28)	59.3% (51)	1.3 (0.6 - 2.7)	0.569
Pain-Pelvic Region	46.5% (20)	40.7% (35)	1.3 (0.6 - 2.6)	0.574
Bloating	55.8% (24)	52.3% (45)	1.2 (0.6 - 2.4)	0.852
Masses-Abdominal Region	65.1% (28)	62.8% (54)	1.1 (0.5 - 2.4)	0.848
Fatigue	69.8% (30)	69.8% (60)	1.0 (0.5 - 2.2)	1.0
Incontinence-Urinary	16.3% (7)	17.4% (15)	0.9 (0.3 - 2.5)	1.0
Urinary Urgency	4.7% (2)	7.0% (6)	0.7 (0.1 - 3.4)	0.718
Change in Bowel Habits	69.8% (30)	76.7% (66)	0.7 (0.3 - 1.6)	0.4
Early Satiety	23.3% (10)	31.4% (27)	0.7 (0.3 - 1.5)	0.411
Nausea	62.8% (27)	72.1% (62)	0.7 (0.3 - 1.4)	0.316
Urinary Frequency	16.3% (7)	25.6% (22)	0.6 (0.2 - 1.5)	0.27
Pain-Abdominal Region	67.4% (29)	81.4% (70)	0.5 (0.2 - 1.1)	0.12
Loss of Appetite	30.2% (13)	53.5% (46)	0.4 (0.2 - 0.8)	**0.015**
Weight Loss	20.9% (9)	38.4% (33)	0.4 (0.2 - 1.0)	**0.049**
Ascites	11.6% (5)	30.2% (26)	0.3 (0.1 - 0.9)	**0.028**

In Table 4.9 Stage 0-2 and Stage 3-4 in the cohort are compared. To note is that only four symptoms/signs achieve significant odds ratios distinguishing earlier and later stages: Weight Loss, Ascites, Bleeding-Vaginal Area and Loss of Appetite. Out of these, only Bleeding-Vaginal Area is more strongly associated with earlier stage than later stage ovarian cancer. Given the demographics of the patient population, with an average age of 61.7, instances of Bleeding-Vaginal Area most commonly refer to post-menopausal bleeding.

4.2.3 Odds Ratios Comparing Cases and Controls and Early and Late Stage Ovarian Cancer Over Different, Overlapping Time Intervals

In Figure 4.1, odds ratios are significant across almost all symptoms/signs when considering the entire year, they lose significance after 30 days, with the exception of Pain-Abdominal Region, Masses-Abdominal Region, and Masses-Pelvic Region. This somewhat contradicts prior research into the length and severity of symptoms/signs prior to diagnosis. It would be expected that the majority of symptoms show no significant odds ratios past, 6 months, as that follows findings provided by Goff et. al.[8]. However, the fact that significant odds ratios are not present at 30 days does contradict prior research.

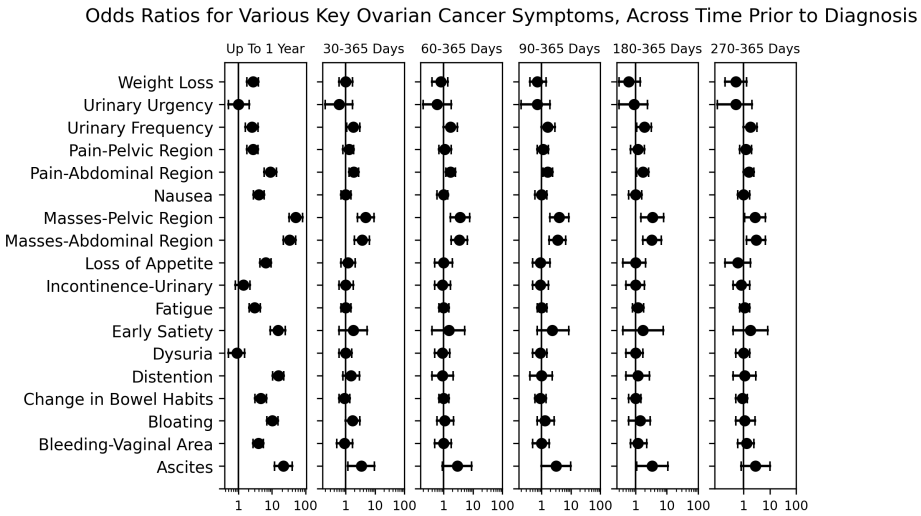

Figure 4.1: Odds Ratios for Ovarian Cancer, Entire Cohort, Across the Year

When the Ovarian Cancer Cohort is further subdivided into the two main routes, Primary Care (N=60) and Gynecology (N=57), a two definitely different patterns emerge from the odds ratios across different time intervals.

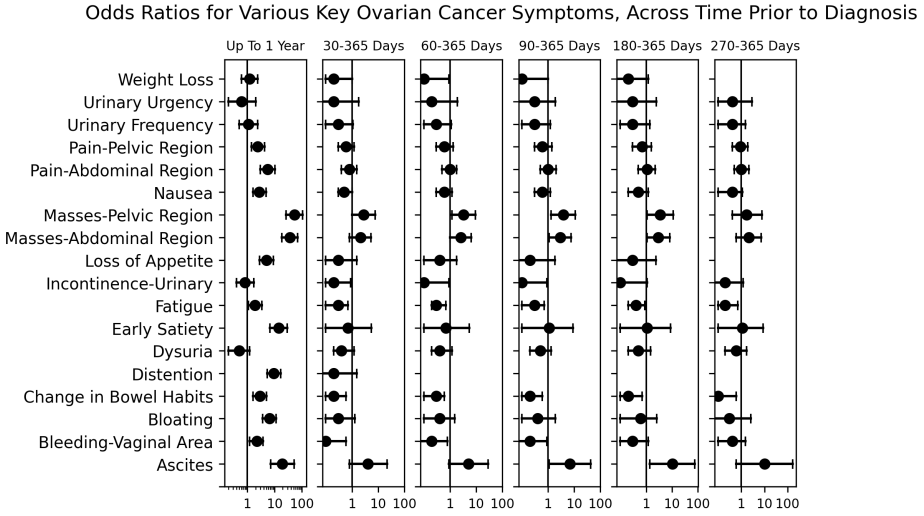

Figure 4.2: Odds Ratios for Ovarian Cancer, Gynecology Route, Across the Year

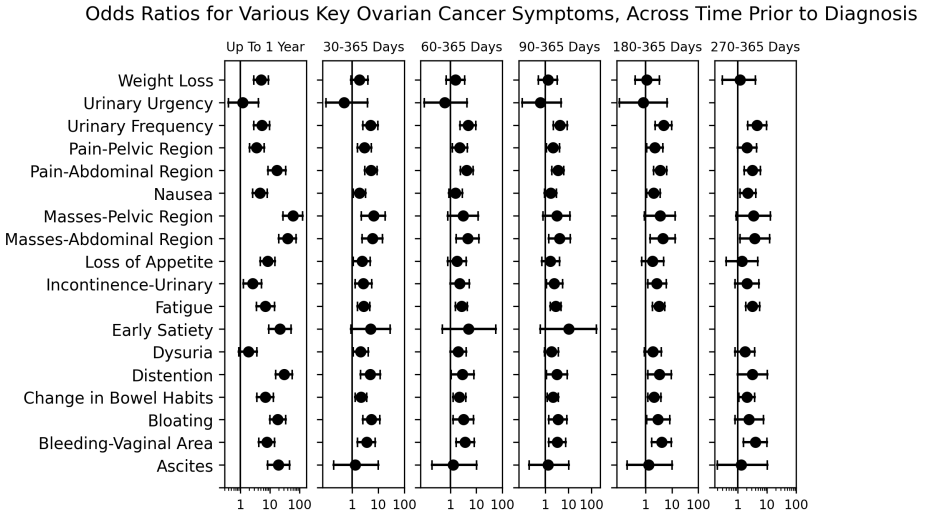

Figure 4.3: Odds Ratios for Ovarian Cancer, Primary Care Route, Across the Year

As can be seen in the figures above, the two largest route sub-cohorts provide significantly different patterns in sign/symptoms over time. In Figure 4.2, covering Gynecology, almost all symptoms/signs are not significant past 30 days. Patients on the Gynecology route, for

whatever reason, do not have significant odds ratios compared to their matched controls in earlier months. However, in 4.3, many symptoms/signs are significant even 270 days prior to diagnosis. Primary Care patients have a symptomatic expression that coincides with the literature, but because both cohorts are about equal in size, the significance is lost in the undivided cohort. The other two smaller cohorts, Emergency and Other, are too small to gain significant insights from their performance.

In the 270-365 Day time window, Bleeding-Vaginal Area, Bloating, Change in Bowel Habits, Distention, Fatigue, Nausea, Pain-Abdominal Region, and Urinary Frequency all have significant odds ratios. In the 180-365 Day Bleeding-Vaginal Area, Bloating, Change in Bowel Habits, Distention, Fatigue, Nausea, Pain-Abdominal Region, Pain-Pelvic Region, and Urinary Frequency all have significant odds ratios. All told, there are 9 symptoms/signs that have significant odds ratios 6 months prior to diagnosis in the primary care context, and 8 symptoms/signs with significant odds ratios 9 months prior to diagnosis in the primary care context.

4.3 Predictive Model Experimentation

This section reviews the 3 experiments undertaken around diagnostic prediction. First, a preliminary set of experiments were conducted to determine the appropriate model framework to use, as well as observe differences in how NLP and ICD can be useful for symptomatic prediction. Second, a set of experiments were performed around the inclusion of explicit absence in models predicting both lung and ovarian cancer. Finally, time series experiments were performed to determine how early prior to diagnosis a given model can predict cancer.

4.3.1 NLP vs. ICD for Explanatory Power

Across all model types there is better performance with the inclusion of both ICD and NLP information, as can be seen in Table 4.10. Understanding the causes of this hinges on the fact that while the majority of symptoms/signs, NLP achieves a higher incidence than ICD alone, in almost all symptoms/signs the combination of both NLP and ICD has a higher

incidence than either separate. This can be observed in Table 4.4. The specificity range (recall of the negative class) is quite high regardless of the experiment, between 0.97-1.0 for all experimental combinations, but this may be due to the relatively high ratio of positive to negative samples compared to prior work.

A Random Forest model achieves the best sensitivity, at 0.68. For the following experiments, only Random Forest models were used due to its higher sensitivity in this experimentation.

Model Type	Data Types	Sensitivity (Training)	Specificity (Training)	Precision (Training)
Random Forest	ICD, NLP	**0.68 (0.449)**	0.99 (0.977)	0.85 (0.682)
Random Forest	ICD	0.47 (0.55)	0.99 (0.973)	0.79 (0.687)
Random Forest	NLP	0.39 (0.454)	0.98 (0.98)	0.65 (0.75)
SVM	ICD, NLP	0.56 (0.428)	**1.0 (0.985)**	**0.93 (0.744)**
SVM	ICD	0.25 (0.396)	0.99 (0.981)	0.73 (0.696)
SVM	NLP	0.39 (0.382)	0.99 (0.989)	0.79 (0.806)
Linear Regression	ICD, NLP	0.64 (0.481)	0.99 (0.981)	0.89 (0.724)
Linear Regression	ICD	0.47 (0.572)	0.97 (0.978)	0.65 (0.74)
Linear Regression	NLP	0.39 (0.436)	0.99 (0.989)	0.79 (0.823)

Table 4.10: 1-Hot Core Present symptoms/signs for Ovarian, Entire Year, Different Data Sources

4.3.2 Presence and Absence in Lung and Ovarian

In this next set of experiments, 1-Hot feature sets were constructed for both present and absent symptoms/signs in both Lung and Ovarian cohorts. To ensure direct comparisons across cancers, these experiments were performed using only NLP extracted symptoms/signs. The inclusion of absent core symptoms/signs improved sensitivity from 0.39 to 0.6 and 0.37 to 0.54 for Ovarian and Lung, respectively. This can be observed in Table 4.11. Specificity remained approximately the same regardless of the experimentation, oscillating between 0.96 and 0.99.

Model Type	Data Types	Sensitivity (Training)	Specificity (Training)	Precision (Training)
Ovarian				
Random Forest	Present	0.39 (0.454)	**0.98 (0.98)**	0.65 (0.75)
Random Forest	Absent, Present	**0.6 (0.366)**	0.97 (0.985)	**0.72 (0.728)**
Lung				
Random Forest	Present	0.37 (0.365)	0.96 (0.971)	0.54 (0.611)
Random Forest	Absent, Present	**0.54 (0.593)**	**0.99 (0.988)**	**0.87 (0.865)**

Table 4.11: 1-Hot Core Present Symptoms/Signs vs. Core All Assertion symptoms/signs for Both Lung and Ovarian, NLP Only, Entire Year

The following shap plots, Figures 4.4 and 4.5, explore the features that the Absent,Present models determined to be the most relevant in both cancer contexts. In shap plots, the color indicates the value of the feature while the x-axis indicates how strong the impact, whether positive or negative, on the probability that the patient belongs to the positive class.

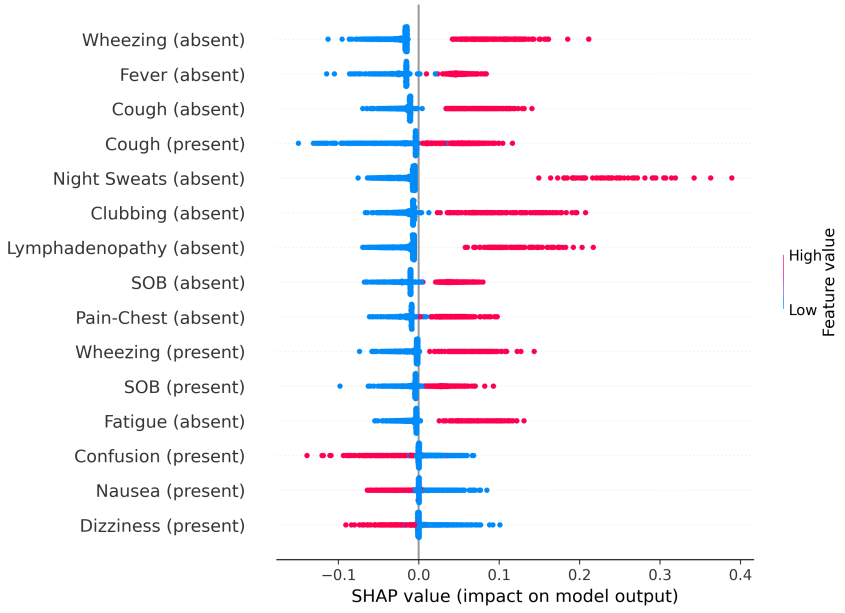

Figure 4.4: Key Features for Lung Cancer, Absent and Present Model

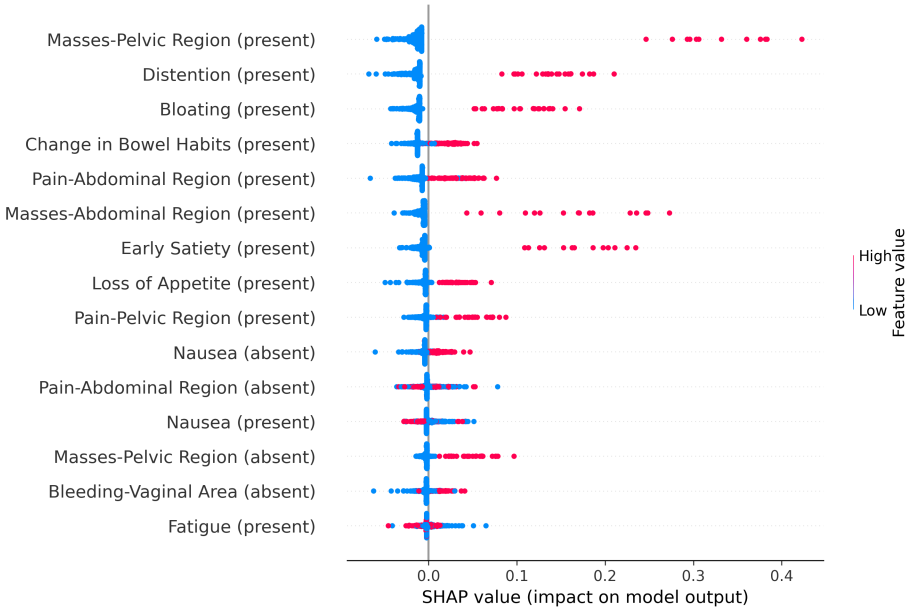

Figure 4.5: Key Features for Ovarian Cancer, Absent and Present Model

In Figure 4.4, both explicitly absent and present wheezing, explicitly absent and present cough, and several other explicitly absent symptoms/signs like night sweats, clubbing, and lymphadenopathy have a strong positive impact on the model when present, and a mild negative impact when absent. Similarly, in Figure 4.5, explicitly absent Nausea, Masses-Pelvic Region, and Bleeding-Vaginal Area have a positive indication when positive, while other symptoms/signs like explicitly absent Pain-Abdominal Region are more mixed.

Table 4.12: Feature Level Performance for Some Ovarian Absent Indicators

Symptom	Correct	Incorrect	A Template
Bleeding-Vaginal Area (Absent)	49	1	0
Nausea (Absent)	45	2	3

In Table 4.12, the precision performance for the specific symptom-assertion combination for two randomly sampled sets of 50 indications. In both Bleeding-Vaginal Area and Nausea,

there is minimal error, with at least 0.9 in precision. It seems unlikely that model inaccuracy is driving the pattern where the presence of explicit absence has a positive impact on the likelihood a patient is a case patient. Another possible explanation is that the models are intuiting provider diagnostic logic. For example, anecdotally "night sweats" is not a routine symptom in a typical patient exam. The fact that the provider is bothering to record the absence of an uncommon symptom could imply concern about an uncommon illness, such as cancer.

4.3.3 Overtime Prediction for Lung and Ovarian

The final experiment undertook comparisons of a random forest model, trained on both present and absent core symptoms/signs, and performance over time.

Table 4.13: 1-Hot Core symptoms/signs, Present and Absent, With Different Duration Filters

Model Type	Data Types	Sensitivity (Training)	Specificity (Training)	Precision (Training)
Random Forest	Full Year, Ovarian	**0.6 (0.366)**	0.97 (0.985)	**0.72 (0.728)**
Random Forest	Full Year, Lung	**0.54 (0.593)**	0.99 (0.988)	**0.87 (0.865)**
Random Forest	Day 30-365, Ovarian	0.0 (0.0)	1.0 (0.994)	0 (0.0)
Random Forest	Day 30-365, Lung	0.32 (0.332)	0.97 (0.977)	0.62 (0.64)
Random Forest	Day 60-365, Ovarian	0.0 (0.0)	0.99 (0.989)	0.0 (0.0)
Random Forest	Day 60-365, Lung	0.24 (0.337)	0.97 (0.976)	0.49 (0.648)
Random Forest	Day 90-365, Ovarian	0.0 (0.0)	1.0 (0.991)	0 (0.0)
Random Forest	Day 90-365, Lung	0.3 (0.281)	0.97 (0.976)	0.56 (0.616)
Random Forest	Day 180-365, Ovarian	0.09 (0.0)	1.0 (0.995)	0.67 (0.0)
Random Forest	Day 180-365, Lung	0.2 (0.254)	0.99 (0.983)	0.69 (0.66)
Random Forest	Day 270-365, Ovarian	0.0 (0.0)	0.99 (0.992)	0.0 (0.0)
Random Forest	Day 270-365, Lung	0.25 (0.214)	0.98 (0.983)	0.58 (0.634)

To note in Table 4.13 is that while lung cancer achieves diminishing returns for predictability past 30 days prior to diagnosis, sensitivity and precision drops to essentially 0

for predicting the positive class of the Ovarian Cancer Cohort. This indicates that with the Ovarian Cancer Cohort, the model assigns all instances to the negative class using these features prior to 30 days. This is likely due to variety of compounding factors, but the strongest hypothesis is explored in the subsequent section.

As for lung cancer, the important features change over time across the different time windows, with the sensitivity remaining more or less stable up until 6 months prior to diagnosis.

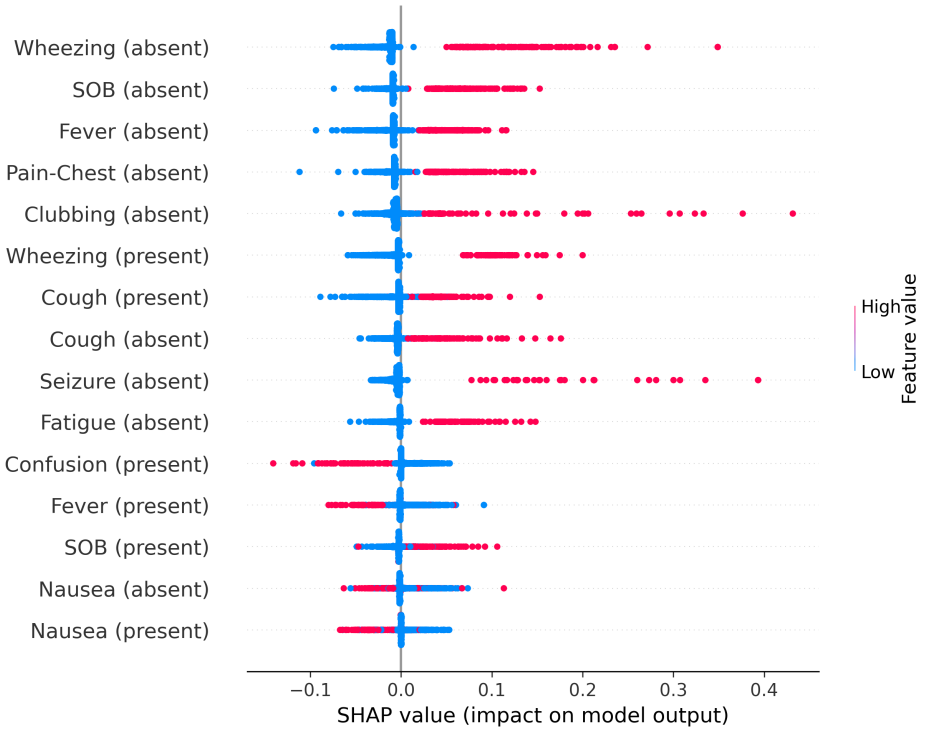

Figure 4.6: Key Features for Lung Cancer, 90-365 Days

As seen in Figure 4.6, by 90 days, absent night sweats is no longer a key symptom, despite its strong impact in the prior experimentation. On the other hand, absent lymphadenopathy and clubbing remain key features even until 180 days (6 months) prior to diagnosis, seen in 4.7. If these explicit absences are indeed an indication that the doctor is concerned about case

patients more than controls, that could imply that for a strong minority of patients, doctors are concerned even 6 months prior to diagnosis. This topic of concern and recommendation is explored in the Discussion chapter.

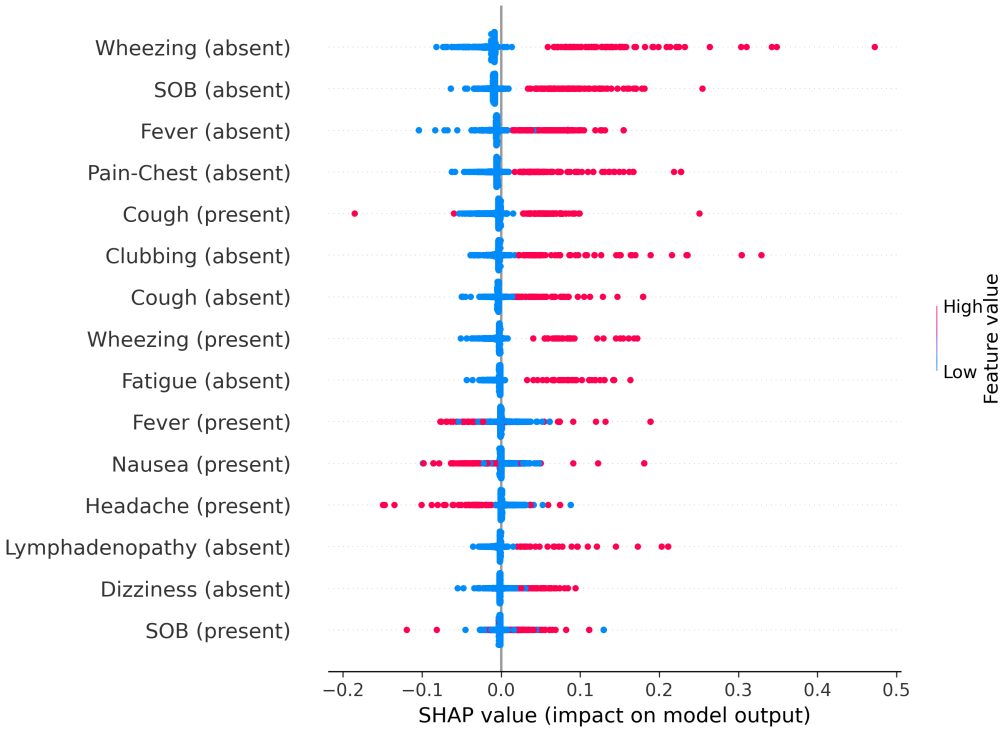

Figure 4.7: Key Features for Lung Cancer, 180-365 Days

Chapter 5

DISCUSSION

Both lung and ovarian cancer have complex symptomatic profiles which lend themselves to delays in diagnosis.This work has delved deeply into the symptomatic profile of ovarian cancer across the year prior to diagnosis. Experimentation was performed in order to explore how symptoms can be useful for prediction of both lung and ovarian cancer. Several findings remain difficult to explain. First, what is causing the lack of predictability of ovarian cancer past 30 days from diagnosis? Second, if the usefulness of explicit absent symptoms/signs indeed implies that the model is inferring cancer from provider suspicion, are there other forms of delay more readily reduced? If a provider is indeed already concerned, than it becomes harder to justify a predictive model alerting the provider to such a concern.

5.1 Explanatory Factors for Lack of Predictability for Ovarian Cohort Past 30 Days

A main finding in the Results is the lack of predictive capability more than 30 days prior to diagnosis with the Ovarian Cancer Cohort. This is strikingly different from the Lung Cancer Cohort. The following differences in route types for present symptoms/signs in the note may indicate an explanation.

As one can see in Figure 5.1, there is very little noticeable difference between cases and controls when considering all patients as a whole past 30 days. When further subdividing across different routes in Figures 5.2 and 5.3, Primary Care case patients have a distinctly different symptomatic curve than their relevant controls as compared to Gynecology. At 6 months, Primary Care patients have 46.6% of their cohort with at least one symptom vs. 23% for controls; while Gynecology patients have 15.8% percent of their cohort with at least

one symptom vs. 20.7% for controls. There is a different diagnosis pattern for these two sub-cohorts, which cumulatively provide the pattern seen in the overall cohort.

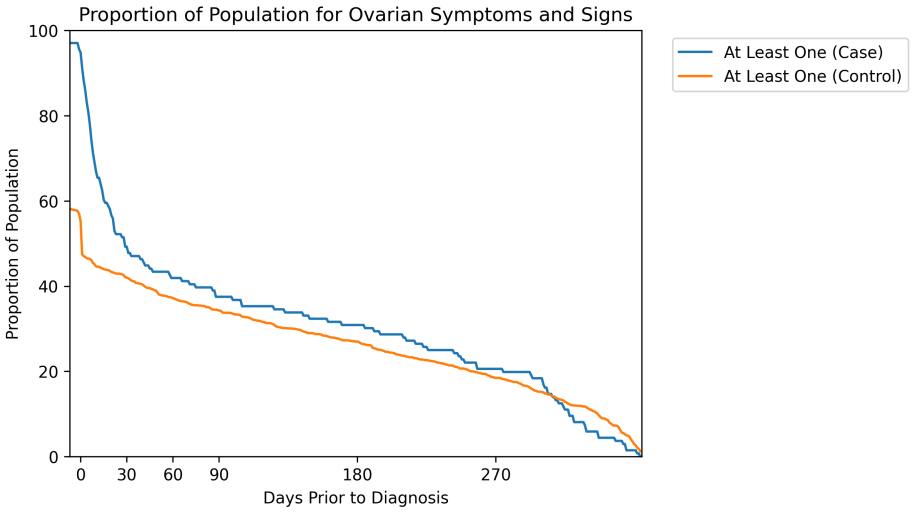

Figure 5.1: Frequency of At Least One Symptom/Sign in Ovarian Cancer. In Notes, Cases and Controls, Year Prior to Diagnosis

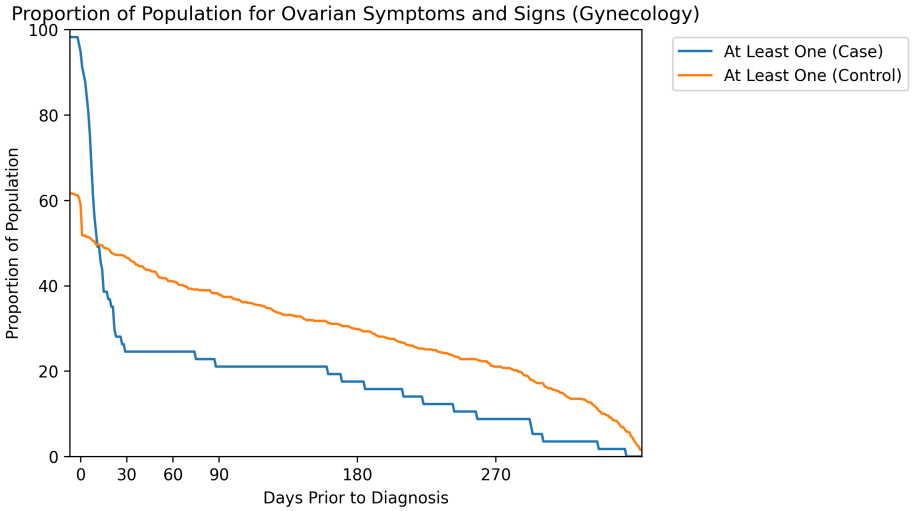

Figure 5.2: Frequency of At Least One Symptom/Sign in Ovarian Cancer (Gynecology). In Notes, Cases and Controls, Year Prior to Diagnosis

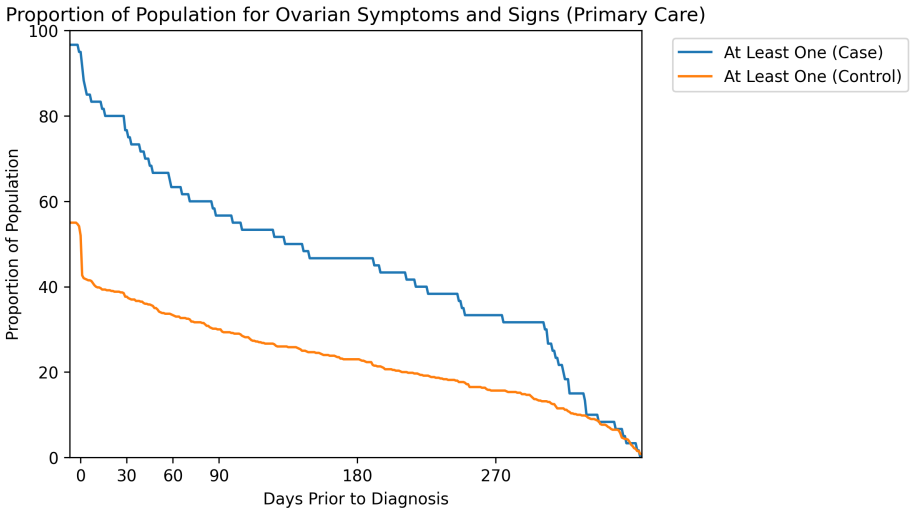

Figure 5.3: Frequency of At Least One Symptom/Sign in Ovarian Cancer (Primary Care). Notes, Cases and Controls, Year Prior to Diagnosis

One potential explanatory factor is patient visit rates. Note frequencies are very different in the Overall, Primary Care, and Gynecology cohorts, as seen in the Figures 5.4, 5.6, and 5.5.

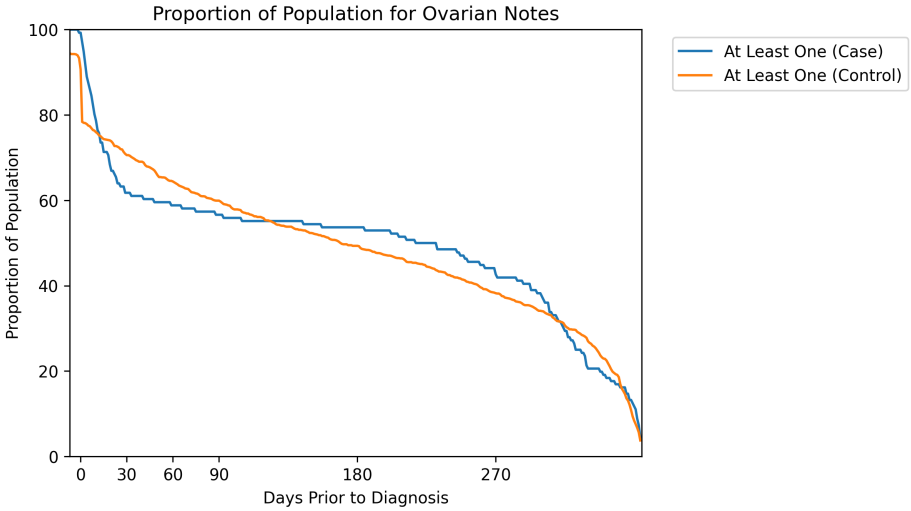

Figure 5.4: Frequency of At Least One Note in Ovarian Cancer. In Notes, Cases and Controls, Year Prior to Diagnosis

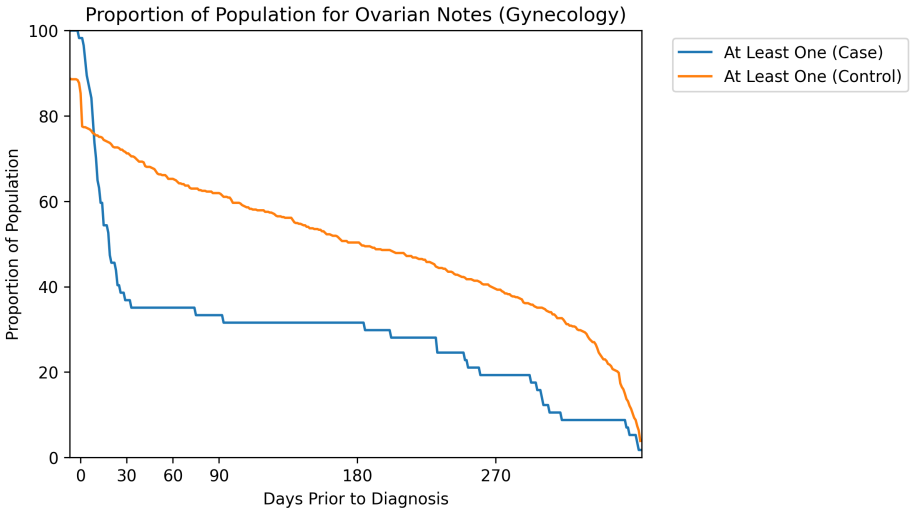

Figure 5.5: Frequency of At Least One Note in Ovarian Cancer (Gynecology). In Notes, Cases and Controls, Year Prior to Diagnosis

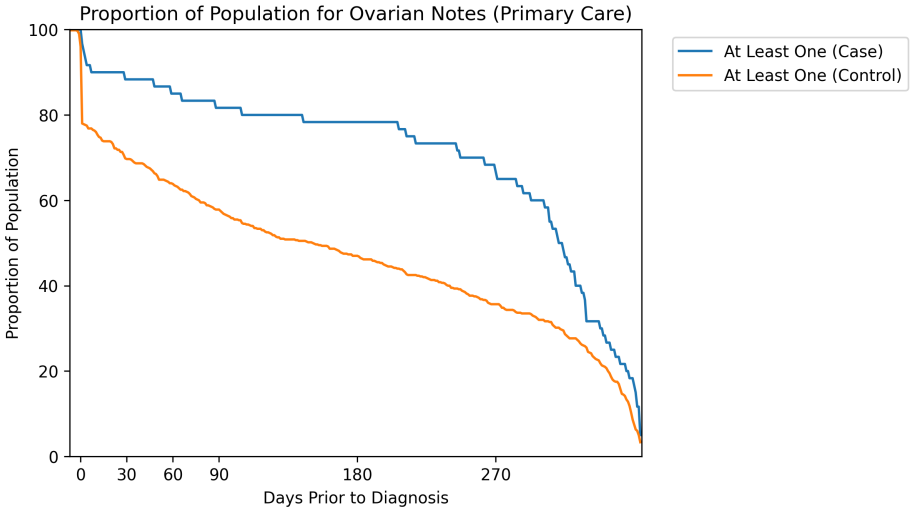

Figure 5.6: Frequency of At Least One Note in Ovarian Cancer (Primary Care). Notes, Cases and Controls, Year Prior to Diagnosis

By 30 days, only 36.8% of Gynecology case patients have a clinical note, while 88.3% of Primary Care patients have one. Further research is needed to determine why Gynecology

patients visits differ from Primary Care patients. Some potential explanations may be patients with an external referral, Gynecology route patients receiving primary care outside the UW system, or other reasons. This variation in routes, combined with the small sample size, conspires to prevent the model from correctly predicting Ovarian cancer given the symptom feature set.

5.2 Other Causes of Delay in Lung and Ovarian Cancer Diagnosis

In the following section provider recommendations for the "next step" on the patient's diagnosis pathway were analyzed to determine if there were any significant differences between cases and controls. As part of this work, the recommendation extraction model extracted recommendation sentences from the notes belonging to both cases and controls in both Cohorts. These sentences were normalized for references to "Chest CT", "Chest MRI", and "Chest X-Ray" for the Lung Cancer Cohort, and "CT", "Ultrasound" and "Gynecology" for the Ovarian Cancer Cohort. Then, odds ratio and frequency analyses were performed when considering patients with at least one of the relevant recommendations.

The proportion of patients with at least one recommendation in the Ovarian and Lung Cancer Cohort is consistently higher in cases than the respective controls, despite the difficulty discerning symptoms/signs past 30 days. The proportion in the Ovarian cases is 20.6% at 6 months prior to diagnosis compared to 11.1% at 6 months for controls. Similarly, the proportion is 38.2% at 6 months for Lung cases compared to 26.7% for Lung controls. In Table 5.1 the odds ratios for Ovarian (Primary Care) and Ovarian overall continue to be significant well into 9 months, similar to Lung. However, patients in the Ovarian (Gynecology) route do not have significant odds ratios even at 3 months, implying a dramatically different route pattern.

When the Ovarian cohort is divided further, Primary Care route patients experience a far higher incidence of recommendations than their Gynecology counterparts. At 6 months, Primary Care case patients have 78.3% with at least one recommendation, while Gynecology route patients have 14%. At 9 months, Primary Care case patients have 66.6% with at least

one recommendation, while Gynecology route patients have 8.7%. Thus, the majority of the difference in the overall Ovarian Cancer Cohort likely driven by the Primary Care route.

Further research is needed to determine if this overall pattern of delay holds true across other cancers and disease states in the UW context, or if it is only for the Lung and Ovarian cohorts. There also needs to be research determining the proportion of potentially spurious recommendations, as in both Cohorts controls receive non-zero relevant "next step" recommendation levels, which could imply conservative recommendations on the part of UWM physicians. That being said, if this pattern holds true across several disease states, than operational stakeholders need to be engaged in order to prevent this potentially life altering delay in care.

Patient Cohort	Full Year OR	3 Month OR	6 Month OR	9 Month OR
Lung	18.7 (15.5 - 22.4)	6.1 (4.8 - 7.8)	5.2 (3.8 - 7.1)	4.1 (2.7 - 6.3)
Ovarian	35.8 (19.1 - 67.1)	2.0 (1.3 - 3.2)	2.3 (1.4 - 3.8)	3.2 (1.8 - 5.6)
Ovarian (Gynecology)	65.8 (15.9 - 272.9)	1.1 (0.5 - 2.3)	1.4 (0.6 - 3.2)	1.6 (0.6 - 4.2)
Ovarian (Primary Care)	29.0 (13.4 - 62.7)	3.8 (2.1 - 7.2)	4.8 (2.3 - 9.6)	5.4 (2.3 - 12.5)

Table 5.1: Odds Ratios for At Least One Recommendation Across Year Prior to Diagnosis

5.3 Limitations

There are several limitations to this work. A primary limitation is in the adaptation of the recommendation extraction model. This model, while it has the necessary precision, has an unknown recall in the context of non-radiology notes. Since the use of this model was upstream at the point of creating the case routes, a lower recall could have downstream impacts on the results. Similarly, much of the results are driven by the output of adapted symptom extraction models, which, while achieving similar performance as human annotators, are not perfect by any means. For example, although other roles were extracted, such as duration, they were not used in the final analysis because of both the poor annotator agreement and

subsequently poor performance of the extraction model on such roles. This means that much of the nuance around duration, frequency, and other roles present in the free text of the note was ignored during normalization.

Another limitation is the ultimate size of the Ovarian cohort and the nature of the UWM dataset. Ovarian cancer is relatively rare, and unfortunately as a major referral center, many ovarian cancer patients diagnosed at the University of Washington experienced the majority of their care prior to diagnosis elsewhere. The use of recommendations is intended to eliminate some of these direct referral patients, but it is possible that patients in the Gynecology route experienced lower odds ratios not because they experience fewer symptoms/signs relative to their matched controls, but because some of them may have been referred into the UWM system midway through their diagnostic process. Thus, similarly to the size of the dataset, the selection of cases and controls likely contributed to variations in the results. Reasons for exclusion or inclusion of patients were well intentioned but could easily have introduced bias into the dataset. As especially seen through analysis of the Gynecology route, the absence of evidence is not evidence of absence when it comes to data in the EHR. This makes it difficult to draw conclusions about the generalizable nature of our study population. Indeed, many of the findings related to recommendation and route patterns may simply be an artifact of the UWM system.

A final limitation is in the underlying bias within the notes themselves. This limitation is two fold. First, the implication that the extraction model uses explicit absence of symptoms/signs as a marker for cancer. If it is true that this implies that the model is picking up on provider suspicion, than potentially any model using extracted symptoms/signs from the note to do early prediction is fraught, because it would be difficult to unwind a true indication of a disease or provider suspicion of a disease. If the model is simply predicting based on provider suspicion, the model would be limited in usefulness. Second, as seen in prior work by Goff, for ovarian cancer the duration and frequency of the symptoms/signs matter. However, both duration and frequency are not always explicitly recorded in the note. While some information is better than no information at all, such inconsistency makes

building a model difficult. Finally, absence of evidence is not evidence of absence. Many patients receive care outside of any given system, and so gaps in care are normal and even the default. These gaps make it difficult to determine if the patient is asymptomatic, but instead is receiving care elsewhere and is symptomatic.

5.4 Conclusions and Future Work

This work explores the importance of information extraction in work that relies on symptomatic information for analyses. Many symptoms that are known to be correlated with ovarian cancer do not have significant odds ratios when observing only the ICD-10 codes, such as weight loss. This work underscores the finding that analyses of coded symptoms are incomplete without the inclusion of symptoms present within the note. Any work that solely relies on coded information for symptomatic analyses would likely be an incomplete profile of the patient population, to the detriment of the work.

To that effect, this work also quantifies the difficulty of transfering a symptom extraction model to a new domain. Adapting the model to a new domain does require new data from the out-domain in order to avoid a drop in recall. This adapation process is important, because much of the change in recall is driven by symptoms associated with the out-domain. Using a model without adapting it to the new domain could potentially mean missing symptoms crucial to that new domain that are relatively rare in other domains.

This work also undertook the creation of a case-control dataset of the year prior to diagnosis for ovarian cancer. It endeavored to avoid patients that are "referral" patients through the use of recommendations for the "next step" in the diagnosis pathway. It endeavored to match patients appropriately to their controls through the use of the route – the type of clinic or provider that had the initial suspicion of the diagnosis in the year prior. This clustering of patients in different routes uncovered significant differences in symptoms/signs, recommendation, and note expression between routes. These differences were masked when observing the cohort as a whole. The differences uncovered indicate the need for caution when observing patterns in patient population data. If these variations are so large, and

so easily hidden, it begs the question what other variations might be hidden, caused by the relative proportions of different clinical referral patterns. When observing solely the Primary Care route, several symptoms/signs are significantly higher up to 270 days prior to diagnosis. In the Gynecology route, such symptoms/signs are not significant. However, this may be due to a dearth in data in these Gynecology patients, and not a dearth of symptoms/signs within those patients, because the note proportions indicate that a higher proportion of the Gynecology case patients receive care outside the UWM system.

Finally, the Ovarian Cancer Cohort and the Lung Cancer Cohort were leveraged to perform predictive experimentation. In both Lung and Ovarian, the inclusion of explicit absence for core symptoms has a positive impact on positive class recall (sensitivity). One hypothesis is that the models are correlating relevant provider suspicion and diagnostic thinking with case patients. Thus, it may be that the models are not simply predicting cancer, but predicting suspicion of cancer. Ultimately, in Lung the prediction of patients was possible, although with deteriorating quality, several months prior to diagnosis. In Ovarian, however, prediction was not possible even 30 days prior to diagnosis. It seems very likely that the lack of predictability in Ovarian is due to differences in symptom expression across routes.

The future work can be divided into a handful of sub-categories. First, it would be extremely useful to attempt to improve symptom model extraction performance on certain key roles, such as duration. The use of symptom duration could be crucial in reducing the impact that lack of data has on observing symptoms over time. So long as providers were explicit about the duration of a certain symptom, it would not matter that prior visits occurred outside the system. The relevant data could still be processed appropriately. If such duration information could be extracted reliably, a re-evaluation of the uni-variate and prediction experiments on ovarian cancer, specifically the overtime concerns, may be advised.

Second, given that it is unclear when such roles could be reliably extracted, and whether they appear in the notes enough to provide value, other avenues should be considered. It does appear that provider recommendations for the “next step” in cancer diagnoses appear

at far higher rates in cases than controls, even 6 or 9 months prior to diagnosis. Of course, further research is needed to determine whether this correlation is spurious or rooted in fact. If it is rooted in fact, deliberation must be made on what policy changes might be useful, if any, to reduce this potential source of delay. One consideration is that too much ordering can be equally problematic as ordering too little. While recommendations are higher in case patients, they are still present in control patients, potentially indicating a conservative recommendation pattern by providers.

All cancers are very serious diagnoses, and time is of the essence in preventing poor outcomes. Thus, it behooves everyone to remove as much diagnostic delay as possible from our healthcare system. One potential route is to improve our capabilities of prediction as a clinical decision support tool for overworked providers. This might reduce the time between a patient first reports a symptom and when a provider indicates suspicion of cancer. Another potential route is by reducing the time between when a provider first indicates suspicion and when the patients is finally diagnosed. Both intervention points have potential for improvement, and further work is needed to determine when and how the system could intervene. However, this work does seem to indicate that there are several points of concern that make it difficult to address the former. The most effective change point might be the latter, depending on the root causes behind the delay between initial suspicion and ultimate diagnosis.

BIBLIOGRAPHY

[1] Clare Bankhead, C. Collins, H. Stokes-Lampard, P. Rose, S. Wilson, A. Clements, D. Mant, S. Kehoe, and J. Austoker. Identifying symptoms of ovarian cancer: A qualitative and quantitative study. *BJOG : an international journal of obstetrics and gynaecology*, 115:1008–14, 07 2008.

[2] Jacqueline Barrett and William Hamilton. Pathways to the diagnosis of lung cancer in the uk: a cohort study. *BMC Family Practice*, 9, 2008.

[3] Michael A. Beckles, Stephen G. Spiro, Gene L. Colice, and Robin M. Rudd. Initial evaluation of the patient with lung cancer*: Symptoms, signs, laboratory tests, and paraneoplastic syndromes. *Chest*, 123(1, Supplement):97S–104S, 2003.

[4] Lili Chan, Kelly Beers, Amy A. Yau, Kinsuk Chauhan, Áine Duffy, Kumardeep Chaudhary, Neha Debnath, Aparna Saha, Pattharawin Pattharanitima, Judy Cho, Peter Kotanko, Alex Federman, Steven G. Coca, Tielman Van Vleck, and Girish N. Nadkarni. Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients. *Kidney International*, 97(2):383–392, 2020.

[5] Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. Building gold standard corpora for medical natural language processing tasks. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:144–153, 11 2012.

[6] Sayon Dutta, William J. Long, David F.M. Brown, and Andrew T. Reisner. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. *Annals of Emergency Medicine*, 62(2):162–169, 2013.

[7] Garth Funston, Helena O'Flynn, Neil A J Ryan, Willie Hamilton, and Emma J Crosbie. Recognizing gynecological cancer in primary care: Risk factors, red flags, and referrals. *Advances in therapy*, pages 577–589, 2018.

[8] Barbara A. Goff, Lynn S. Mandel, Charles W. Drescher, Nicole Urban, Shirley Gough, Kristi M. Schurman, Joshua Patras, Barry S. Mahony, and M. Robyn Andersen. Development of an ovarian cancer symptom index. *Cancer*, 109(2):221–227, 2007.

[9] William Hamilton, Tim Peters, Clare Bankhead, and Deborah Sharp. Risk of ovarian cancer in women with symptoms in primary care: population based case-control study. *BMJ*, 339, 2009.

[10] Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364–379, 02 2019.

[11] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[12] Wilson Lau, Thomas H. Payne, Ozlem Uzuner, and Meliha Yetisgen. Extraction and analysis of clinically important follow-up recommendations in a large radiology dataset. *AMIA Joint Summits on Translational Science proceedings*, page 335–344, 2020.

[13] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[14] Adrian Levitsky, Maria Pernemalm, Britt-Marie Bernhardson, Jenny Forshed, Karl Kölbeck, Maria Olin, Roger Henriksson, Janne Lehtiö, and Carol Tishelman. Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model. *Nature Research Scientific Reports*, 9, 2019.

[15] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[16] Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. Extracting covid-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *Journal of Biomedical Informatics*, (144-153), 2021.

[17] Georgios Lyratzopoulos, Jane Wardle, and Greg Rubin. Rethinking diagnostic delay in cancer: how difficult is the diagnosis? *BMJ*, 349, 2014.

[18] Tsuguo Naruke, Ryosuke Tsuchiya, Haruhiko Kondo, Hisao Asamura, and Haruhiko Nakayama. Implications of staging in lung cancer. *Chest*, 112(4, Supplement):242S–248S, 1997.

[19] Carolyn Rooth. Ovarian cancer: risk factors, treatment and management. *British Journal of Nursing*, 22(Sup17):S23–S30, 2013. PMID: 24067270.

[20] Shahid Munir Shah and Rizwan Ahmed Khan. Secondary use of electronic health record: Opportunities and challenges. *IEEE Access*, 8:136947–136965, 2020.

[21] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5—29, 2015.

[22] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(1):7–30, 2018.

[23] Robert A. Smith, Vilma Cokkinides, Durado Brooks, Debbie Saslow, and Otis W. Brawley. Cancer screening in the united states, 2010: A review of current american cancer society guidelines and issues in cancer screening. *CA: A Cancer Journal for Clinicians*, 60(2):99–119, 2010.

[24] Christine Stewart, Christine Ralyea, and Suzy Lockwood. Ovarian cancer: An integrated review. *Seminars in Oncology Nursing*, 35(2):151–156, 2019. Gynecology Oncology.

[25] Elizabeth Suh-Burgmann and Mubarika Alavi. Detection of early stage ovarian cancer in a large community cohort. *Cancer Medicine*, 8, 09 2019.

[26] Lindsey A. Torre, Britton Trabert, Carol E. DeSantis, Kimberly D. Miller, Goli Samimi, Carolyn D. Runowicz, Mia M. Gaudet, Ahmedin Jemal, and Rebecca L. Siegel. Ovarian cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(4):284–296, 2018.

[27] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics.

[28] Sarah Walters, Camille Maringe, Michel P Coleman, Michael D Peake, John Butler, Nicholas Young, Stefan Bergström, Louise Hanna, Erik Jakobsen, Karl Kölbeck, Stein Sundstrøm, Gerda Engholm, Anna Gavin, Marianne L Gjerstorff, Juanita Hatcher,

Tom Børge Johannesen, Karen M Linklater, Colleen E McGahan, John Steward, Elizabeth Tracey, Donna Turner, Michael A Richards, Bernard Rachet, and the ICBP Module 1 Working Group. Lung cancer survival and stage at diagnosis in australia, canada, denmark, norway, sweden and the uk: a population-based study, 2004–2007. *Thorax*, 68(6):551–564, 2013.

VITA

Grace Turner works as a software developer. She likes to solve problems in healthcare using common sense, NLP, and perhaps far too many graphs.