Subgrouping Ulcerative Colitis Patients using Administrative Claims Data

by

Heather Berlin

B.S., Washington University in St. Louis (2016)
S.M., Washington University in St. Louis (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 14, 2021

Certified by .....
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Subgrouping Ulcerative Colitis Patients using Administrative Claims Data

by

Heather Berlin

Submitted to the Department of Electrical Engineering and Computer Science
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Approximately 3 million patients in the US have been diagnosed with Ulcerative Colitis, a chronic inflammatory disease affecting the colon. Uncovering patient subgroups could improve treatment guidelines and help physicians choose an appropriate treatment plan for a patient. Here, we outline a Python implementation to generate a cohort from a dataset in the OMOP Common Data Model (CDM), propose a patient timeline visualization tool, create and analyze a cohort of Ulcerative Colitis patients using a claims dataset. We extract patient features and use dimensionality reduction techniques along with clustering to identify patient subgroups. We observe four patient subgroups consisting of distinct patient characteristics, most prominently age, insurance type, sex, and type of initial conventional therapy prescription.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Peter Szolovits. Pete's continuous encouragement, guidance, and insight has been invaluable. Without Pete, this thesis would not have been possible.

I would like thank our Takeda collaborators for their clinical guidance and for granting access to the Optum dataset for this thesis. In particular, I would like to thank Pravin Kamble for his help defining the cohort inclusion criteria. This research was supported by Millennium Pharmaceuticals, Inc. (a subsidiary of Takeda Pharmaceuticals).

Finally, I would like to thank my friends and family for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The prevalence and incidence of Ulcerative Colitis are both rising worldwide. A variety of medications may be used in treatment plans for patients with Ulcerative Colitis, including 5-aminosalicylic drugs, corticosteroids, immunomodulators, and biologics [26, 39]. With the number of drugs available to patients increasing, understanding potential subgroupings of patients can improve guidelines in disease management as well as strengthen the community's understanding of the disease. In this thesis, we outline a cohort generation tool to process data in the OMOP CDM format and visualize patients on an individual-level as well as on a cohort-level. We analyze a cohort of newly diagnosed Ulcerative Colitis patients, and demonstrate the potential to use clustering methods on claims data to uncover distinct patient subgroups.

Chapter 2 provides background on longitudinal claims data and the standardization framework of the OMOP Common Data Model (CDM), Ulcerative Colitis (UC), and machine learning methods used for clustering and discovering patient subtypes. Chapter 3 discusses the cohort generation process used to create patient cohorts from data transformed to the OMOP CDM. The cohort of newly diagnosed Ulcerative Colitis (UC) patients, patient outcomes of interest, and the methods used are described in Chapter 4. Chapter 5 discusses the experimental results and examines the patient subgroups uncovered by using clustering on features from the patient cohort. A discussion of our results, study limitations, and future work is included in Chapter 6.

Chapter 2

Background and Related Work

2.1 Standardization of Longitudinal Health Data

Health data can come from multiple sources, including sources like surveys, administrative and medical records, vital records, disease registries, peer-reviewed literature [31]. Medical records are collected directly from health care providers and include information about events and transactions between patients and healthcare providers. Medical records include information on services including diagnoses, procedures, and lab tests. Electronic health records (EHR) were introduced in the 1960's but gained widespread adoption in 2009.

Longitudinal health records contain patient health information generated by one or more encounters in a clinical care setting. Large-scale electronic healthcare databases provide opportunity to generate large-scale insights about patients and healthcare settings.

2.1.1 Overview of Standardization Methods

While electronic medical records are widely used and the databases contain vast amounts of patient information and health-related events, the databases are often difficult to leverage by researchers. Electronic patient record systems were not designed for researchers and are organized to fit the local structure of the institution of

deployment. For example, hospitals, clinics, health insurance companies, and pharmacies have different codes and structures used, all of which contribute to differing electronic patient records. Electronic Health Record (EHR) data supports clinical practice at the point of care, whereas insurance claims data is built for the insurance reimbursement process.

Methods to handle the variety, velocity, and volume of medical and health information systems have been actively researched. Solutions include the HL7 clinical document architecture [5], the OpenEHR platform [11], and the OMOP Common Data Model [32].

OpenEHR is a two-level modeling approach to extract data from various medical databases by separating operations of medical experts and software engineers OpenEHR is designed for exporting and reusing data from distributed EHR systems, but a major limitation of OpenEHR is that it's often incompatible with non-EHR data. Additionally, there is a steep learning curve and a lack of documentation [18].

The OMOP Common Data Model (CDM) was developed to be an international standard for observational research and represents healthcare data from multiple sources in a standardized way. The CDM has demonstrated utility for multiple databases spanning a variety of data types. Since the OMOP CDM is open-sourced, documentation and tools are readily available and frequently updated by the community. Because not all source data can be converted into a standardized vocabulary, there will be loss of information when transforming a dataset into the OMOP CDM.

2.2 The OMOP Common Data Model (CDM)

The Observational Health Data Sciences and Informatics (OHDSI, pronounced "Odyssey") initiative spun out of the Observational Medical Outcomes Partnership (OMOP), which aimed to build a ready-to-use database in a standard common data model and accelerate clinical research leveraging electronic health records [9]. The OMOP Common Data Model (CDM) harmonizes coding systems to a standard vocabulary allowing information from a variety of clinical databases to be combined and accessed

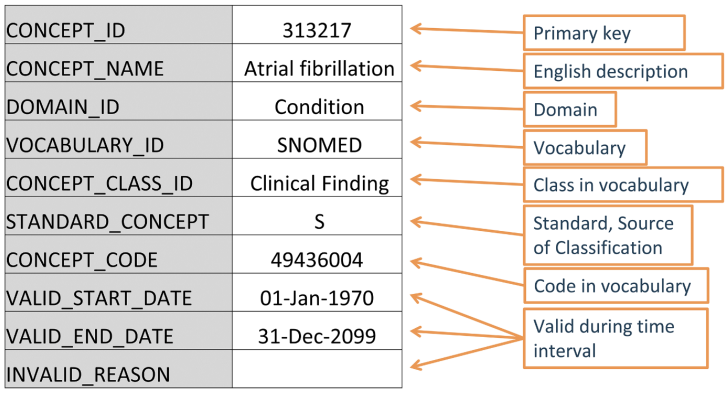

Figure 2-1: Overview of the OMOP CDM table structure. Standard representation of vocabulary concepts in the OMOP CDM. Example is a record from the CONCEPT table for the SNOMED code for atrial fibrillation sourced from [29]

in a standardized manner [30]. The OMOP-CDM was first developed in 2008 as a public-private partnership between the FDA, multiple pharmaceutical companies, and healthcare providers for safety surveillance studies. After successfully completing the project over the course of five years, OMOP transitioned to expand focus to incorporate other clinical domains [30]. OMOP harmonizes coding systems to a standard vocabulary. Once a database is converted to the OMOP CDM, tools that work on one OMOP dataset will work on any other dataset in the CDM format.

2.2.1 CDM Tables

Standard CDM tables include: Conditions, Drugs, Procedures, Measurements, Observations, and Visits. The CDM is a "person-centric" model, linking all event tables (e.g., `visit_occurrence`, `condition_occurrence`, `drug_exposure`, `measurement`) to the PERSON table. A longitudinal patient-level view can be generated by linking a patient's event records along with the start and end dates of each event.

To standardize the content of diverse healthcare records, the CDM employs Standardized Vocabularies, which are derived from public and proprietary terminologies and contain appropriate standard health concepts. Standardized Vocabularies are utilized to assign concepts across a variety of domains: Condition (SNOMED, ICD-10), Procedure (SNOMED, CPT4, HCPCS, ICD10PCS, ICD9Proc, OPCS4), Mea-

surement (SNOMED, LOINC), Drug (RxNorm, RxNorm Extension, CVX), Device (SNOMED), Observation (SNOMED), and Visit (CMS Place of Service, ABMT, NUCC) [30]. While all codes are mapped to the Standardized Vocabularies, the original source codes are maintained in the dataset to ensure no information is lost.

2.2.2 Hierarchical Relationships

The `CONCEPT_ANCESTOR` table is generated from the `CONCEPT_RELATIONSHIP` table, traversing all possible concepts connected through hierarchical relationships. Any two concepts that have a defined relationship are stored in the `CONCEPT_RELATIONSHIP` table to map the type of relationship. For example, a standard SNOMED concept id for hypertension has a "maps to" relationship with the non-standard ICD10 concept id for hypertension.

The `CONCEPT_ANCESTOR` table is automatically generated from the `CONCEPT_RELATIONSHIP` table and contains "Is a" - "Subsumes" pairs, and other relationships connecting hierarchies across vocabularies. Currently, there are comprehensive concept hierarchies available for drug and condition concepts [30]. For example, the SNOMED concept id for "Atrial fibrillation" is related to the SNOMED concept id for "Atrial arrhythmia" through a "Is a" relationship. Both "Atrial fibrillation" and "Atrial arrhythmia" have the same attributes (shown as ancestors and descendants in Figure 2-2), except for the type of arrhythmia, which is fibrillation in one and atrial arrhythmia in the other.

2.2.3 Adoption of the OMOP CDM

The OMOP CDM is gaining adoption on an international level. The Heart Institute (InCor) of São Paulo, Brazil, which contains longitudinal health data for more than 1.3 million patients spanning two decades, was integrated to the OMOP CDM; the integration demonstrated that the new database, standardized to the international CDM, was consistent with the data in the original database [19].

Longitudinal Korean nationwide health insurance data was successfully transformed to the OMOP-CDM [45]. The source data is documented with Korean national

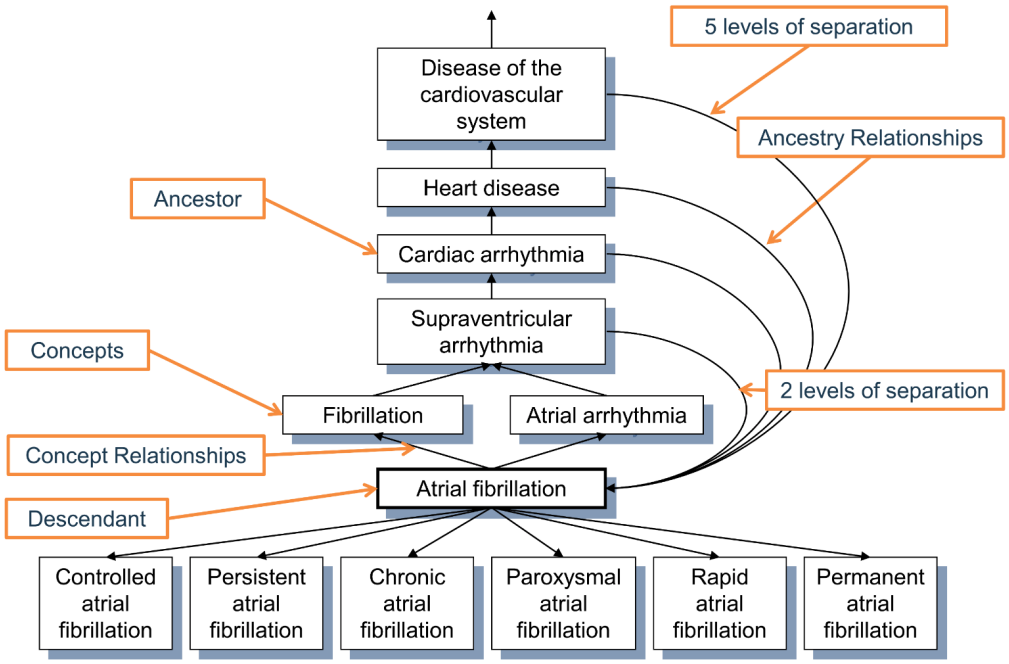

Figure 2-2: Example of a relationship hierarchy in OMOP CDM concepts. Sourced from [30].

medical code system, but by converting the data into the OMOP-CDM, the data is available in an internationally standardized OMOP-CDM format, allowing more clinical investigators to generate evidence that is applicable to Asian populations. The conversion of Korean clinical documents from the HL7 clinical document architecture to the OMOP-CDM has also been explored, which would facilitate the use of health information exchange information in longitudinal clinical studies [10]. Data from the Clinical Practice Research Datalink (CPRD), a UK EHR database, was effectively converted to the OMOP CDM format with acceptable information loss [24].

As more researchers work with data collected for different purposes, by different organizations that use different terms to describe the same clinical concept, transforming these data into the OMOP CDM for downstream use cases is increasingly attractive. By migrating more databases to the OMOP CDM, clinical researchers will have more standardized data readily accessible to aid in large-scale collaborative analyses and support research initiatives to further the community's understanding of clinical conditions.

2.3 Ulcerative Colitis

Ulcerative colitis (UC) is a chronic and recurrent inflammatory disease affecting the colon. Inflammation in the rectum spreads from the distal to proximal colon in UC patients. UC most commonly affects adults aged 30-40 years old [39].

2.3.1 Symptoms

Common symptoms of new UC or recurrent flare-ups frequently include abdominal pain, bloody diarrhea, and/or mucous diarrhea [38] [26]. The UC patient's trajectory is characterized by periods of flare ups and remission, which can occur in response to treatment changes and/or sporadically without known cause.

2.3.2 Treatment

The aim of therapy for UC patients is to induce and maintain remission. The treatment strategy may be guided by the clinician's assessment of disease aggressiveness. For mild to moderate UC, aminosalicylates are the most common treatment choice; a first-line of therapy for adult patients with mild to moderate UC, the American College of Gastroenterology recommends oral therapy with 5-aminosalicylic acids (5-ASAs) like mesalamine, balsalizine, sufasalazine, and olsalazine [15].

To treat UC flares, topical and systemic steroids can be used. In moderate to severe cases of UC, immunosuppressants and biologics may be used [26]. Among adult population UC longitudinal cohort studies, salycilates were the most consumed medication, followed by systemic steroids and immunosuppressants [22].

Combination therapy has been shown to be more effective in UC maintenance than isolated oral therapy or isolated topical therapy, specifically in patients with left-sided colitis and pancolitis [7]. Recent results demonstrate that 5-ASAs are useful for inducing UC remission as well as preventing relapse without safety risks from long-term use [2].

Ulcerative colitis is a recurrent condition which is only fully removed by means of colectomy; a colectomy is needed in up to 15% of UC patients [22], even if receiving

treatment. UC patients may also need to undergo procedures to treat complications of dysplasia [38].

2.3.3 Open Questions

There are many open questions about patients with Ulcerative Colitis. Among IBD patients, much less is known about Ulcerative colitis in terms of factors that contribute to worsening disease than is known about Crohn's disease. In Crohn's disease patients, which may be misdiagnosed as UC patients and vice versa due to overlapping symptoms, a study found that factors such as young age of diagnosis, history of smoking, and early steroid use were predictors of disabling disease outcomes [17]. The optimal medication, dose, and route of administration for UC patients are not entirely understood, so this work aims to investigate UC patient subtypes and examine differences between patients based on the type of conventional treatment initiated.

2.4 Unsupervised Clustering on Healthcare Data

While supervised learning involves learning a function mapping from a set of input variables to a known target variable, unsupervised learning involves a set of input variables without corresponding labels. Unsupervised clustering can be useful for revealing hidden structures summarizing and aggregating complex high-dimensional datasets. The goal of clustering is to group similar data points together in a cluster distinct from dissimilar data points.There are many types of clustering techniques; each technique generally falls in one of five categories: centroid-based methods, connectivity-based methods, density based methods, low dimensional embeddings, and probabilistic clustering methods [16]

There are many clustering methods and a variety of quality metrics used to evaluate the clusters produced [16]. We will not go into detail on each of the various clustering techniques and quality metrics but for more information, the methods are documented in detail in the literature and [16] provides a detailed overview. For our purposes, we will use k-Means, a centroid-based method which requires the number

of clusters and distance metric to be specified.

Unsupervised clustering has been performed on longitudinal healthcare data for a variety of applications, some of which include identifying distinct immunological patient profiles that influenced the evolution of COVID-19 in patients admitted to the hospital with COVID-19 [20], identifying phenotypic subgroups in a cohort of heart failure patients [35], and identifying phenotypic subgroups of patients with Amyotrophic lateral sclerosis (ALS) [37]

Unsupervised learning has demonstrated utility in identifying patterns and relationships from EHR data without requiring human-specified labels. On claims and EHR data specifically, some of the use-cases for unsupervised clustering include discovering latent disease clusters and patient subgroupings [41] and identifying implausible observations [6]. When applying an unsupervised clustering algorithm to EHR data on laboratory tests to identify implausible observations, clustering resulted in fewer false positives than conventional anomaly detection approaches [6].

2.4.1 Identifying Patient Subgroups

Unsupervised clustering has been used to identify patient subgroups for a variety of diseases, including Alzheimer's Disease [1], breast cancer [8], diabetes [23] heart failure [35], and COVID-19 [20].

Claims data has been used to characterize patient trajectories and outcomes [12]. Recently, administrative claims data was used to identify outcomes across health systems in patients who underwent bariatric surgery [21].

2.4.2 Dimensionality Reduction

Unsupervised feature learning has been used to represent patients by a set of features inferred automatically from large-scale EHR databases [27]. Feature learning algorithms such as principal component analysis (PCA), k-means clustering (K-Means), and Gaussian mixture models (GMMs) have demonstrated utility in reducing dimensionality to obtain feature embeddings [27]. Principal Component Analysis (PCA) is a

linear feature extraction technique that can be used to extract a low dimensional set of features from a high dimensional data set while retaining as much information as possible. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique commonly used to visualize high-dimensional data in lower dimensions [42, 40]. The t-SNE method preserves local structure by converting Euclidean distances between data points into conditional probabilities that represent similarity measures and minimizing the sum of the difference in conditional probabilities, converging these probability distributions of neighborhoods around points to a lower-dimensional mapping. t-SNE has been used to visualize high-dimensional patient data in healthcare domains such as medical imaging data, sequencing data, and longitudinal electronic health records [4, 46, 3].

Chapter 3

Cohort Creation Pipeline

Our cohort-creation pipeline is an alternative to the pipeline developed by ATLAS, which also works with OMOP-formatted data [9]. To create a cohort, we generated and leveraged cohort-specific concept sets, defining groups of entries that belonged to a particular event of interest (e.g., drugs that are immunomodulating medications).

3.1 Defining Cohort Inclusion Criteria

We defined the patient cohort using a rule-based approach. In our case, the patient cohort included Ulcerative colitis (UC) patients. The inclusion criteria were translated into a PostgreSQL query.

3.2 Building Cohort-Specific Concept Sets

To aid in generating the patient cohort, we generated concept sets, which represented groups of concepts that can be reused for future analyses. We created a table to store our concept sets, `concept_sets`. It contained all of the concept records from the standard OMOP `CONCEPT` table which belong in each of our concept sets, along with a `concept_set_name` field labeling the concept set an entry belongs to.

To make sure a concept set includes all OMOP-equivalent concepts, we included all direct concept mappings in a concept set (e.g., `UC_Condition` included ICD9,

ICD10, and SNOMED codes for Left sided ulcerative colitis), as shown in figure 3-4.

When building a concept set, we did not require patient records to be present for each concept included; this gave us the ability to reuse concept sets on different datasets in the future. Since a dataset may contain patient records of a specific format (e.g., SNOMED codes for conditions), and our concept sets included multiple sources (ICD9, ICD10, SNOMED), we did not expect patient records to be available for all concepts.

To ensure the concept sets can be referenced and reproduced on different datasets, we saved a .sql file containing the PostgreSQL query used to build the concept sets.

3.2.1 Defining a Concept Set

We defined the contents of a concept set with one input file which specified the codes and corresponding code vocabulary of the concepts to include. The input file could describe concepts in any vocabulary the user preferred (ICD10, ICD9, SNOMED, etc.), since all equivalent concepts would later be extracted automatically to ensure all equivalent concepts across vocabularies were included. The structure and an example input file defining a concept set, `<concept_set>.csv`, is shown in figure 3-1, which contains example contents for an `autoimmune_disease` concept set.

3.2.2 Extracting Direct Concept Mappings

Using `<concept_set>.csv`, we extracted all concepts that directly correspond to each `(vocabulary_id, concept_code)` pair. Next, we extracted all of the concepts that directly mapped to or from any of the concepts that appeared in the first set of mapping results. Figure 3-2 displays the generated PostgreSQL query and query results after extracting direct mappings from the example `<concept_set>.csv` in figure 3-1.

autoimmune_disease_condition.csv

	vocabulary_id	concept_code
0	ICD9	714.0
1	ICD9	714.1
2	ICD9	714.2
3	ICD9	714.30
4	ICD9	714.31
..	...	...
83	ICD10	H30.12
84	ICD10	H30.13
85	ICD10	H30.9
86	ICD10	H30.8
87	ICD10	H20.82

[88 rows x 2 columns]

Figure 3-1: Example contents of an input file used to specify a concept set with source codes.

3.2.3 Storing Concept Sets

Once a concept set was created, each `concept_set` entry was stored in the `concept_sets` table along with the name of the concept set. This framework was developed to allow a concept set to be easily modified, added, or removed from the `concept_sets` table for future use. Additionally, having the concept sets readily available encouraged reproducible results and helped reduce querying time when building the cohort or otherwise querying records satisfying multiple conditions involving multiple concept sets.

Looking at the last entry of figure 3-5, which is a `UC_Procedure` concept, we see that `2109048` is the unique OMOP `concept_id` for "Colectomy, total abdominal, without proctectomy; with ileostomy or ileoproctostomy", a procedure specified by the procedure CPT4 code=44150.

3.3 Defining Patient Outcomes

Once the patient cohort was defined, we labeled patient outcomes. Patient outcomes could be defined with the same rule-based approach used to define a patient cohort, and concept sets were similarly leveraged.

```
filename = 'autoimmune_disease_condition.csv'
concept_set_name = filename.split('.')[0] # remove .csv ending

source_dict, df = load_source_concept_codes(filename, bucket)

concept_set_query = get_concept_id_query(source_dict, concept_set_name)
results = pd.read_sql(concept_set_query, conn)

print(concept_set_name)
print("------------")
print(concept_set_query)
print("------------")
results
```

```
autoimmune_disease_condition
------------

      SELECT *
      FROM cdm_data_cuts_optum.concept
      WHERE vocabulary_id ilike '%ICD10%'
      AND concept_code IN ('M069', 'M0500', 'M0530', 'M0560', 'M061', 'M0800', 'M083', 'M0840', 'M0840', 'M1200', 'M0510', 'M064', 'M064', 'M459', 'M4600', 'M4980', 'M4680', 'L732', 'L4054', 'L4059', 'L400', 'L401', 'L402', 'L403', 'L404', 'L408', 'L410', 'L411', 'L418', 'L42', 'L440', 'L305', 'L448', 'H30.2', 'H44.11', 'H35.02', 'H35.06', 'H30.0', 'H30.10', 'H30.11', 'H30.12', 'H30.13', 'H30.9', 'H30.8', 'H20.82')
      UNION

      SELECT *
      FROM cdm_data_cuts_optum.concept
      WHERE vocabulary_id ilike '%ICD9%'
      AND concept_code IN ('714.0', '714.1', '714.2', '714.30', '714.31', '714.32', '714.33', '714.4', '714.81', '714.89', '714.9', '720.0', '720.1', '720.81', '720.89', '705.83', '696.0', '696.1', '696.2', '696.3', '696.4', '696.5', '696.8', '363.21', '360.12', '362.12', '362.18', '363.00', '363.01', '363.03', '363.04', '363.05', '363.06', '363.07', '363.08', '363.10', '363.11', '363.12', '363.13', '363.14', '363.20', '363.22', '364.24')

ORDER BY vocabulary_id, concept_code
------------
```

	concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code	valid_start_date	valid_end_date	invalid_reason
0	45567021	Focal chorioretinal inflammation	Condition	ICD10	ICD10 code	None	H30.0	1990-05-01	2099-12-31	None
1	45557388	Posterior cyclitis	Condition	ICD10	ICD10 code	None	H30.2	1990-05-01	2099-12-31	None
2	45596059	Other chorioretinal inflammations	Condition	ICD10	ICD10 code	None	H30.8	1990-05-01	2099-12-31	None
3	45581638	Chorioretinal inflammation, unspecified	Condition	ICD10	ICD10 code	None	H30.9	1990-05-01	2099-12-31	None
4	45538573	Pityriasis rosea	Condition	ICD10	ICD10 Hierarchy	None	L42	1990-05-01	2099-12-31	None
...	...	...	...	...	...	...	...	...	...	...
56	44832629	Unspecified inflammatory polyarthropathy	Condition	ICD9CM	4-dig billing code	None	714.9	1970-01-01	2099-12-31	None
57	44835004	Ankylosing spondylitis	Condition	ICD9CM	4-dig billing code	None	720.0	1970-01-01	2099-12-31	None
58	44837355	Spinal enthesopathy	Condition	ICD9CM	4-dig billing code	None	720.1	1970-01-01	2099-12-31	None
59	44832655	Inflammatory spondylopathies in diseases classified elsewhere	Condition	ICD9CM	5-dig billing code	None	720.81	1970-01-01	2099-12-31	None
60	44825704	Other inflammatory spondylopathies	Condition	ICD9CM	5-dig billing code	None	720.89	1970-01-01	2099-12-31	None

61 rows × 10 columns

Figure 3-2: Querying concepts that directly map to the input concepts defining the autoimmune disease concept set

	concept_set_name	concept_count
0	Autoimmune_Disease_Condition	103
1	CD_Biologic_Drug	1451
2	CD_Biosimilar_Drug	187
3	CD_Condition	59
4	CT_Aminosalicylate_Drug	4745
5	CT_Corticosteroid_Drug	47529
6	CT_Immunomodulator_Drug	20543
7	Emergency_Visit	8
8	Injectable_Corticosteroid_Drug	519
9	Inpatient_Visit	108
10	Oral_Corticosteroid_Drug	1459
11	Rectal_Corticosteroid_Drug	79
12	Topical_Corticosteroid_Drug	1288
13	UC_Biologic_Drug	1129
14	UC_Biosimilar_Drug	187
15	UC_Condition	95
16	UC_Procedure	10

Figure 3-3: Unique concept counts in the claims dataset for each of our 17 IBD-related concept sets.

	concept_set_name	concept_id	concept_name	vocabulary_id	condition_occurrence_record_count
0	CD_Condition	201606	Crohn's disease	SNOMED	2293785
1	UC_Condition	81893	Ulcerative colitis	SNOMED	1707419
2	CD_Condition	46269889	Complication due to Crohn's disease	SNOMED	697061
3	CD_Condition	195585	Crohn's disease of small intestine	SNOMED	442839
4	CD_Condition	194684	Crohn's disease of large bowel	SNOMED	434719

Figure 3-4: Condition occurrence counts for the most prevalent conditions included in our IBD-related concept sets (UC and CD) across the entire Optum dataset.

	concept_set_name	concept_id	concept_name	domain_id	vocabulary_id	standard_concept	concept_code
0	Autoimmune_Disease_Condition	72705	Pauciarticular juvenile rheumatoid arthritis	Condition	SNOMED	S	74391003
1	CD_Biologic_Drug	28496	Infliximab	Drug	MeSH	None	D000069285
2	CD_Biosimilar_Drug	783887	infliximab Injectable Solution [Zessly]	Drug	RxNorm Extension	S	OMOP4831716
3	CD_Condition	194684	Crohn's disease of large bowel	Condition	SNOMED	S	7620006
4	CT_Aminosalicylate_Drug	505783	Mesalazine 300mg/5ml oral suspension 1 ml	Drug	SNOMED	None	32863611000001100
5	CT_Corticosteroid_Drug	505455	Prednisolone 1mg gastro-resistant tablets (A A H Pharmaceuticals Ltd)	Drug	SNOMED	None	33425811000001105
6	CT_Immunomodulator_Drug	505410	Mycophenolate mofetil 500mg powder for concentrate for solution for infusion vials (Accord Healthcare Ltd)	Drug	SNOMED	None	33341411000001104
7	Emergency_Visit	8870	Emergency Room - Hospital	Visit	CMS Place of Service	S	23
8	Injectable_Corticosteroid_Drug	1506430	Methylprednisolone 40 MG Injection	Drug	RxNorm	S	311659
9	Inpatient_Visit	8676	Nursing Facility	Visit	CMS Place of Service	S	32
10	Oral_Corticosteroid_Drug	783225	Budesonide 3 MG Delayed Release Oral Capsule [Entocort] Box of 100	Drug	RxNorm Extension	S	OMOP4831056
11	Rectal_Corticosteroid_Drug	928002	Pramoxine hydrochloride 10 MG/ML Rectal Foam [Proctofoam]	Drug	RxNorm	S	828367
12	Topical_Corticosteroid_Drug	975171	Hydrocortisone 25 MG/ML Topical Lotion	Drug	RxNorm	S	197785
13	UC_Biologic_Drug	28537	Adalimumab	Drug	MeSH	None	D000068879
14	UC_Biosimilar_Drug	783887	infliximab Injectable Solution [Zessly]	Drug	RxNorm Extension	S	OMOP4831716
15	UC_Condition	75580	Chronic ulcerative proctitis	Condition	SNOMED	S	52231000
16	UC_Procedure	2109048	Colectomy, total, abdominal, without proctectomy; with ileostomy or ileoproctostomy	Procedure	CPT4	S	44150

Figure 3-5: A sampled concept from each of our IBD-related concept sets.

Chapter 4

Methodology

4.1 Data Source

We used United States health insurance claims data from the Optum database which was transformed to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). While the Optum Research Database contains claims data for millions of commercial and Medicare Advantage patients, a subset of this database was used for this analysis. The Optum Claims dataset we used for our analyses was a data cut of patients with Inflammatory Bowel Disease (IBD), which encompasses both Ulcerative colitis and Crohn's Disease diagnoses. The Optum claims data cut included anonymized patient data for 623,063 patients spanning six sources: insurance membership, medical claims, pharmacy claims, lab tests, inpatient visits, and provider data. The insurance billing codes data included demographics, office visits, hospitalizations, conditions, procedures, drugs, and in some cases lab test results. Since this work focused on Ulcerative colitis patients, patients with Crohn's Disease conditions were omitted from our analysis. The resulting dataset used included $64.40\%(n = 401,226)$ distinct patients with a UC condition.

4.2 Study Population

4.2.1 Newly Diagnosed Ulcerative Colitis Patients

Our cohort was defined using rule-based inclusion criteria and consisted of newly diagnosed Ulcerative Colitis (UC) patients who initiated conventional therapy between 2011 and 2018. The initial prescription fill for treatment was defined as the patient's index date. The drug concepts which we classified as a drug used for Conventional Therapy (CT) included: aminosalicylates, corticosteroids, and immunomodulators. All patients in the cohort were naive to therapies conventionally used to treat UC, including aminosalicylates, corticosteroids, or immunomodulators, before the index date. For each patient, we gathered data from one year before and one year after the index date; continuous healthcare coverage was required throughout this time period to ensure we were not missing patient data due to lack of insurance coverage. Cohort inclusion criteria and patient attrition during the cohort creation process is detailed in table 4.1.

Inclusion Criteria	Description	Patient Count
1	Initial prescription fill for Conventional Therapy (CT) on or after Jan 1 2011	186,053
2	Age 18 or older on initial CT prescription fill date (index date)	179,960
3	Continuous health benefits for 365 days on or before index date (baseline period)	60,390
4	Continuous health benefits for 365 days after index date (follow-up period)	43,443
5	One UC diagnosis on or before index date	13,986
6	No autoimmune disease diagnoses on or before index date	13,002
7	No CD diagnoses on or before index date	11,507
8	No CD diagnoses after index date	10,717
9	No biologic drug exposure on or before index date	10,619
10	Age and Gender available	10,599
	Total Patients	**10,599**

Table 4.1: Cohort attrition table for UC patients initiating CT.

IBD Concept Sets

We identified UC patients using ICD-9-CM, ICD-10, SNOMED codes in an OMOP CDM formatted claims database. These diagnosis codes have been used to identify UC patients in other studies leveraging claims data [28]. As described in Chapter 3, we defined 17 IBD-related concept sets covering the conditions, treatments, procedures,

and visits that we used to a) define the cohort of Ulcerative Colitis (UC) patients and b) label outcomes of interest.

4.3 Visualizing Patient Timelines

To visually examine individual patient timelines, we developed a patient timeline visualization method that takes as input concept sets of interest (e.g., aminosalicylate drugs and corticosteroid drugs), a `person_id` specifying the person to plot, and start/end dates for the timeline. This visualization tool was useful to better understand trends in the data and also proved useful in discussions around cohort inclusion/exclusion criteria. An example of patient timeline is shown in figure 4-1.

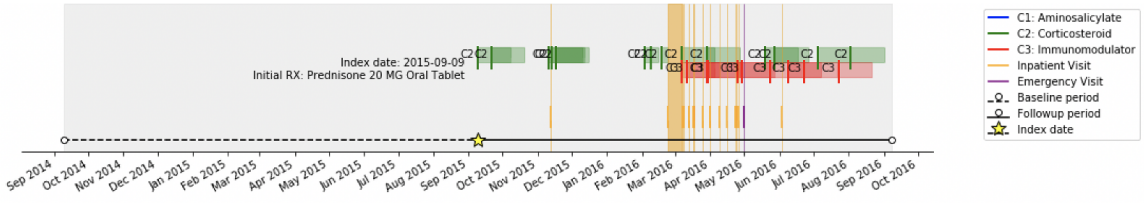

Figure 4-1: Example of a patient timeline for a patient prescribed corticosteroids as initial treatment.

4.4 Defining Adverse Events of Interest

We investigated adverse event outcomes for patients in the cohort using previously defined adverse events of interest among IBD patients receiving conventional treatment [43, 25]. An adverse event was defined as a non-outpatient occurrence within 6 months after a patient initiates treatment of events including the following: acute hepatic failure, Jaundice, Cholestasis, Diabetes mellitus, Congestive heart failure, Abscess of liver, cirrhosis. We exclude non-inpatient events to help ensure that the adverse event labels correspond to moderate to severe patient instances. Additionally, to better identify adverse events due to the initial conventional therapy, an adverse event was only counted if a patient had no history of adverse event occurrences.

	Concept Name	Concept ID
1	Abscess of liver	201901
2	Acute hepatic failure	4026032
3	Aspergillosis	434281
4	Autoimmune chronic active hepatitis	4026125
5	Cholestasis	4143915
6	Congestive heart failure	319835
7	Cryptococcosis	440035
8	Diabetes mellitus	201820
9	Disorder of bone	75909
10	Hepatic infarction	194417
11	Hepatic vein thrombosis	4301208
12	Hepatomegaly	4167902
13	Jaundice	137977
14	Macronodular cirrhosis	4184779
15	Operation on liver	4171687
16	Portal vein thrombosis	199837
17	Toxic noninfectious hepatitis	4052963
18	Tuberculosis	434557
19	Venous thrombosis	444247

Table 4.2: Concepts defined as adverse event during the 6-months following index date.

4.5 Deriving Patient-Level Features

For each patient in the dataset, we extracted all drug exposure data, lab results, procedures, and conditions, and patient demographic information.

4.5.1 Demographics

We incorporated patient demographic information by including age and sex in our analysis. These were constant variables derived from the index date. We also included a categorical variable for 4 age categories: <37 years, 37-53 years, 54-67 years, >67 years at index date, as was done in related work using patient claims data [34].

4.5.2 Medications

To incorporate medication information, we included the number of drug exposures and cumulative days supply for each type of conventional treatment: aminosalicylates, corticosteroids, and immunomodulators. These drugs were identified using the cohort-specific concept sets.

4.5.3 Conditions

Each possible condition (hypertension, anemia etc.) during the time window of interest was extracted and represented with a binary variable. We did not differentiate between primary and secondary conditions when extracting patient features.

4.5.4 Procedures

Similar to how conditions were extracted, every possible procedure (coronary artery bypass, endoscopy, etc.) during the time window of interest was included and represented with a binary variable. Across the baseline and follow-up periods, cohort patient records included 3,711 distinct procedures.

4.5.5 Lab Measurements

Since our data source was claims data, lab records for a patient did not always contain the measurement produced from the lab. To include information on the lab measurements rather than exclusively considering whether or not a lab test was performed, we only considered patient lab records that contained results. Across the patient cohort, there were 1,154 different labs that satisfied this criterion.

The majority of the 1,154 labs containing measurements were performed on less than 3% of the patient cohort, as shown in figure 4-2. To exclude labs that were rare among the cohort, we limited the number of labs included in our analyses. Of the labs with measurements for UC patients during the observation period, we considered lab measurements from the top 100 labs, ranking by the proportion of patients with a lab measurement record. For the top 100 labs, the proportion of patients with measurement data ranged from 0.01 for the least common lab (measuring Mononuclear cells/100 leukocytes in Blood by Manual count) to 0.36 for the most common lab (measuring creatinine in serum or plasma).

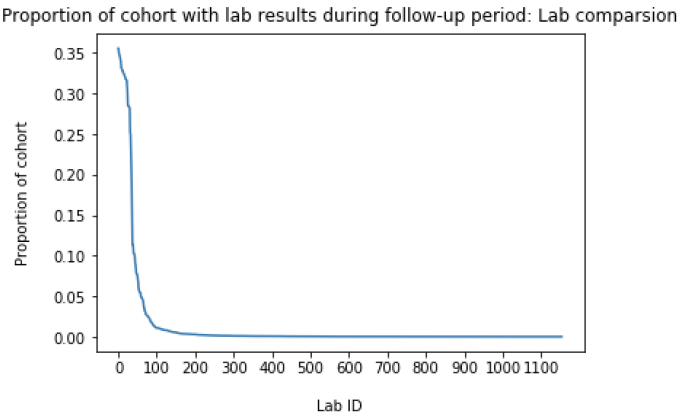

Figure 4-2: Proportion of the UC cohort with a lab result value for each of the labs containing results during the follow-up period.

Each lab with a result available was stored as either 'low', 'normal', or 'high'. The lab result category was determined using the lab result value from the claims dataset along with the low/high lab reference values.

4.5.6 Time Windows

We extracted features over multiple time periods to compare temporal trends across the cohort. We used a window of 6-months in length, which was arbitrarily chosen and consistent with window lengths used in related work [34]. When using windowing, all features except demographics appeared multiple times — once for each time window considered. We use four consecutive 6-month windows spanning from the beginning of the baseline period (12-months before index date) to the end of the follow-up period (12-months following the index date).

4.6 Dimensionality Reduction

To visualize the latent patient structure, we used PCA [44] as well as t-SNE[40], which learns a low-dimensional embedding that preserves higher-dimensional distances between data points [40]. Similar to [35], we removed features that were rare (defined as features available for $<5\%$ of the cohort) to retain the most clinically relevant information. We used features from the 12-month baseline period and 12-month follow-up period for our analysis. From the original set of over 39,000 features, we retained 377 for use in downstream analyses. The data was normalized to have a mean of 0 and standard deviation of 1 before performing PCA. A subset of the top principal components was used to construct a t-SNE output as recommended in the literature [40]. We used the first 50 principal components to project the extracted patient features into a two-dimensional space by t-SNE.

4.7 Patient Clustering

We performed clustering with the goal of identifying patients that were similar to each other, as this information could ultimately provide reference and insight for a patient's treatment plan. We performed K-Means clustering. The optimal number of clusters used in K-Means was chosen by measuring inertia with the Elbow-method [36] and silhouette evaluation methods [13].

Chapter 5

Analysis

5.1 Patients

Of 10,599 patients included in the cohort, 53.2% (n=5,642) were female, and 56.4% (n=5,982) were covered by private insurance. 69.3% (n=7,341) of patients were White, 9.8% (n=1,044) were Hispanic or Latino, 8.6% (n=911) were Black or African American, and 4.3% (n=454) were Asian. The remaining 8% (n=849) of patients did not have a race or ethnicity value recorded in the dataset.

The year of Conventional Therapy (CT) initiation ranged from 2011 to 2018, with an average year of 2014.7 (SD=2.4 years). On the index date, which is the date of CT initiation, the average patient age was 57.7 years (SD=17.7 years), which is higher than the average patient age observed in the literature.

To differentiate between monotherapy and combination therapy routes, four types of initial conventional therapy prescriptions were considered: an index date prescription for 1) an aminosalicylate, 2) a corticosteroid, 3) an immunomodulator, and 4) some combination of an aminosalicylate, corticosteroid, and immunomodulator. 70.5% of patients (n=7,472) initially received exclusively a corticosteroid drug. 24.5% (n=2,601) of patients were initially prescribed exclusively an aminosalicylate drug and 1.7% (n=185) were prescribed exclusively an immunomodulator. 3.2% (n=341) of patients initially received combination therapy.

Across all patients in the cohort, the most common baseline comorbid conditions

included essential hypertension (40.2%), hyperlipidemia (33.6%), and abdominal pain (23.5%). The most common baseline procedures included blood draw (75.7%), biopsy (43.8%), and colonoscopy (27.2%).

Index Date Characteristic	**Total Patients (n=10,599)**
Age	57.7 (17.7) years
Female	5642 (53.2%)
Private Health Insurance	5982 (56.4%)
Race	White, 7341 (69.3%) Hispanic or Latino, 1044 (9.8%) Black or African American, 911 (8.6%) Asian, 454 (4.3%) Other, 849 (8.0%)
Year of Initial CT Prescription	2014.7 (2.4)
Initial CT Prescription	Corticosteroid, 7472 (70.5%) Aminosalicylate, 2601 (24.5%) Immunomodulator, 185 (1.7%) Combination Therapy, 341 (3.2%)

Table 5.1: Patient Demographics and type of Conventional Therapy (CT) prescribed on index date, the date of CT initiation.
Data presented as mean (standard deviation) for continuous variables, or n (%) for categorical variables.

Throughout the two years spanning the baseline and follow-up period, patients received care from care sites across every state in the U.S.; Florida, Texas, and California had the highest numbers of documented care site visits from patients in the cohort during the two year period saddling the index date, as shown in figure B-4.

Baseline Characteristic	Total Patients (n=10,599)
Comorbid Condition	1. Essential hypertension (40.2%) 2. Hyperlipidemia (33.6%) 3. Abdominal pain (23.5%) 4. Diarrhea (23.2%) 5. Hemorrhage of rectum and anus (16.6%)
Drug Exposure	1. Ciprofloxacin 500 MG Oral Tablet (12.8%) 2. {6 (Azithromycin 250 MG Oral Tablet) } Pack (11.3%) 3. Metronidazole 500 MG Oral Tablet (10.0%) 4. Ondansetron Injection (9.0%) 5. Midazolam Injectable Solution (9.0%)
Procedure	1. Collection of venous blood by venipuncture (75.7%) 2. Level IV - Surgical pathology, gross and microscopic examination Abortion - spontaneous/missed Artery, biopsy Bone marrow, biopsy Bone exostosis Brain/meninges, other than for tumor resection Breast, biopsy, not requiring microscopic evaluation of surgica (43.8%) 3. Colonoscopy, flexible; with biopsy, single or multiple (27.2%) 4. Emergency department visit for the evaluation and management of a patient, which requires these 3 key components within the constraints imposed by the urgency of the patient's clinical condition and/or mental status: A comprehensive history; A comprehensi (22.7%) 5. Anesthesia for lower intestinal endoscopic procedures, endoscope introduced distal to duodenum (19.5%)
Lab Measurement	1. Comprehensive metabolic panel This panel must include the following: Albumin, Bilirubin, Calcium, Carbon dioxide (bicarbonate), Chloride, Creatinine, Glucose, Phosphatase, alkaline Potassium (63.9%) 2. Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (63.3%) 3. Lipid panel This panel must include the following: Cholesterol, serum, total Lipoprotein, direct measurement, high density cholesterol (HDL cholesterol),Triglycerides (56.7%) 4. Thyroid stimulating hormone (TSH) (42.0%) 5. Creatinine [Mass/volume] in Serum or Plasma (35.1%)

Table 5.2: Patient Medical History, Spanning Baseline Period.
Data are presented as the clinical characteristic (% of patients). For each medical history category (conditions, drugs, procedures, labs), the five most prevalent clinical characteristics are shown in descending order. Prevalence data were gathered from the baseline period, the 1-year period preceding CT initiation.

5.2 Comparing Patients by Initial Type of Conventional Therapy

While majority of patients initially received corticosteroids, compared to aminosalicylates, the average days supply for a corticosteroid prescription among patients in the cohort was much lower, with a median days supply of 15 days on the index date, as shown in figure 5-1.

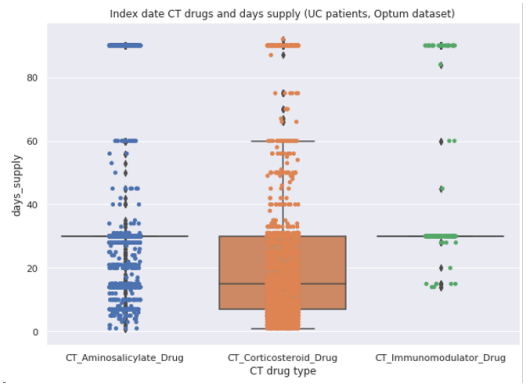

Figure 5-1: Medication days supply on patient index date by treatment type in the cohort.

We examined pair plots between the different treatment types during the 6-month period after index date (T1) and the 6-12 month period after index date (T2) (figure 5-2). Note that a patient may be initially prescribed multiple types of the same drug. For example one patient was initially prescribed multiple aminosalicyte drugs, resulting in a much higher days supply of aminosalicylates than other patients.

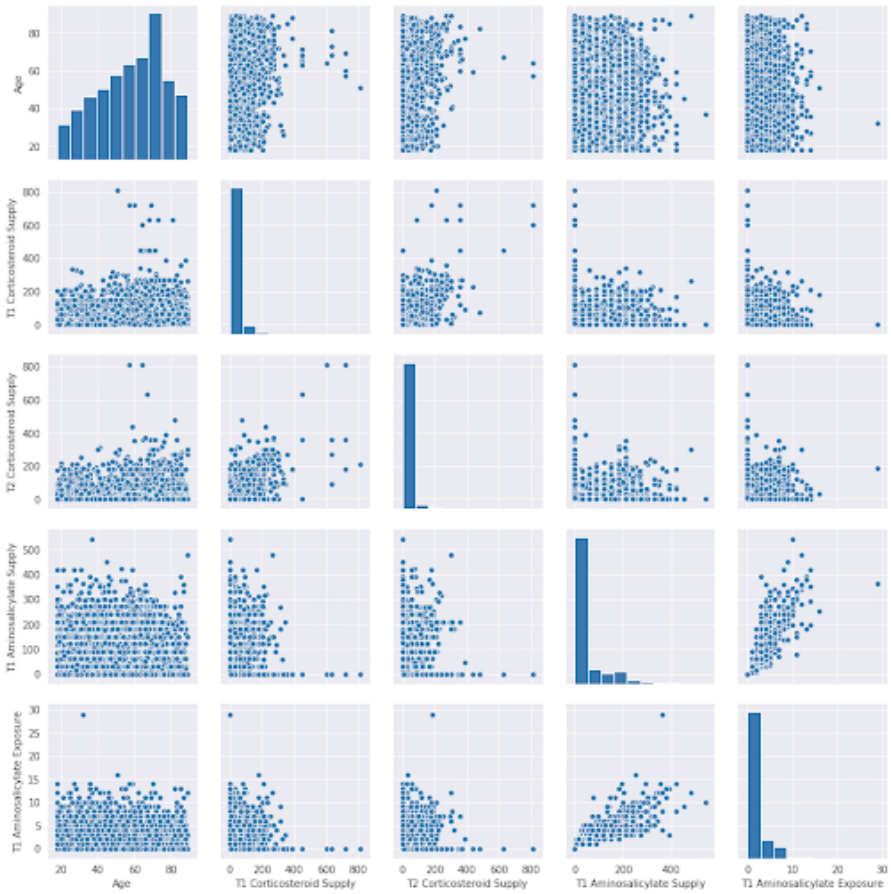

Figure 5-2: Pairplots on Conventional Treatment use among UC Patients. T1 represents the 6 month period after index date, T2 represents the 6-12 month period after index date.

5.2.1 Adverse Events of Interest

Of the 10,599 patients in the cohort, 2.4% (n=259) experienced an adverse event within 6 months of initiating conventional therapy and had no prior history of adverse events. Adverse events were determined as events that are types of the overarching events listed in table 4.4. Patients with a history of adverse events were omitted to increase the possibility of the adverse event truly occurring due to the conventional therapy rather than other factors such as preexisting conditions. Note that 59.7% (n=384) of patients who had an adverse event occurrence within 6 months of CT initiation had a history of adverse events. When omitting patients who experienced an adverse event prior to CT initiation, a total of 2.4% (n=259) of patients in the cohort experienced an adverse event of interest.

Adverse Event	**Patients with Adverse Event within 6mo of CT initiation (n=259, 2.4% of cohort)**
Congestive heart failure	53 (20.5%)
Type 2 diabetes mellitus	41 (15.8%)
Acute deep vein thrombosis of lower limb	24 (9.3%)
Osteoporosis	16 (6.2%)
Disorder of bone	11 (4.2%)
Closed fracture of distal end of radius	10 (3.9%)
Disorder of bone and articular cartilage	7 (2.7%)
Acquired spondylolisthesis	7 (2.7%)
Acute thrombosis of superficial vein of upper extremity	7 (2.7%)
Acute deep venous thrombosis of upper extremity	6 (2.3%)
Chronic congestive heart failure	6 (2.3%)
Closed fracture of neck of femur	6 (2.3%)
Acute deep venous thrombosis of femoral vein	5 (1.9%)
Closed fracture of one rib	5 (1.9%)
Drug-induced diabetes mellitus	4 (1.5%)

Table 5.3: Patient counts for the 15 most common adverse events among the patients within 6 months of initiating treatment, excluding patients with a history of adverse events. Data are shown as n patients (% of the 259 patients who experienced an adverse event).

One aim of this work was to explore whether specific types of initial treatment are associated with adverse events. Out of the patients who experience an adverse event

within 6 months of initiating CT, 79.5% of them (n=206) initially start monotherapy on a corticosteroid drug. However, most patients in the cohort initially received corticosteroids, so class imbalance along with very few observations of adverse events among patients who weren't prescribed corticosteroids made it difficult to draw any causal conclusions between the initial drug prescription and adverse event occurrence. Figure 5-3 shows the adverse event counts for patients by the type of treatment on index date. The full patient counts for each type of initial CT and adverse event outcome are listed in Table 5.4.

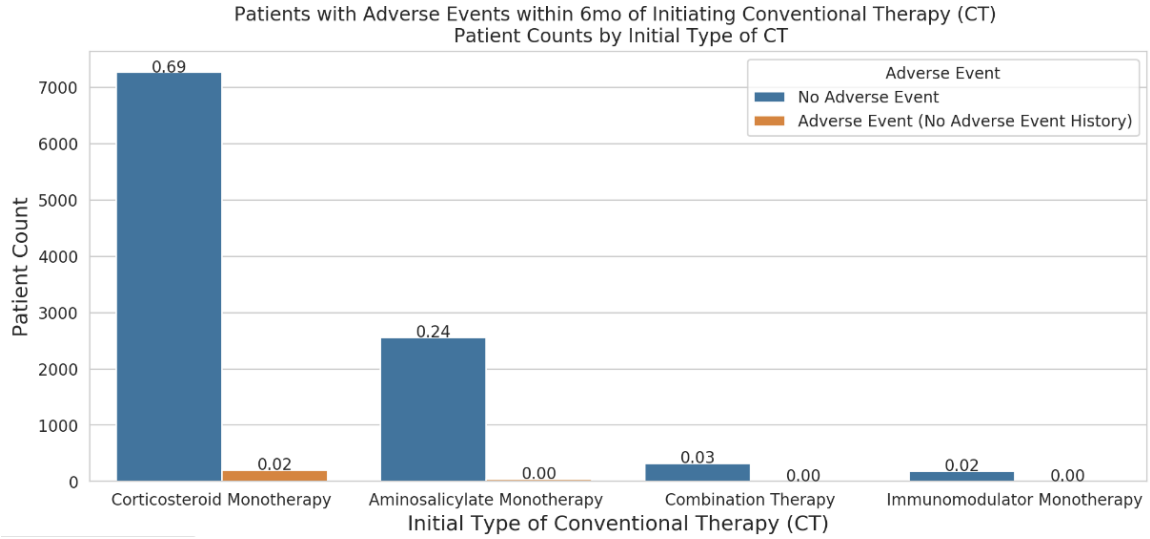

Figure 5-3: Adverse event patient counts by type of initial treatment. The number on top of each bar denotes proportion of the entire cohort that experiences an adverse event within 6-months of CT initiation.

Type of Initial CT	No Adverse Event	Adverse Event	*Total Patients*
Aminosalicylate Monotherapy	2559	42	2601
Corticosteroid Monotherapy	7266	206	7472
Immunomodulator Monotherapy	183	2	185
Combination Therapy	332	9	341
Total Patients	10340	259	10599

Table 5.4: Patient counts for patients who have adverse events within 6 months of CT initiation and no history of adverse events of interest.

5.3 Visualizing Patients with Dimensionality Reduction

To visualize the latent patient structure, we used PCA followed by t-SNE. From the original set of over 39,000 patient-level features, we retained 377 for use in downstream analyses. We used the first 50 principal components for t-SNE and used the t-SNE results to visualize patient data in 2-dimensions, as shown in Figure 5-4. The t-SNE results used in subsequent analyses were created using a perplexity of 500.

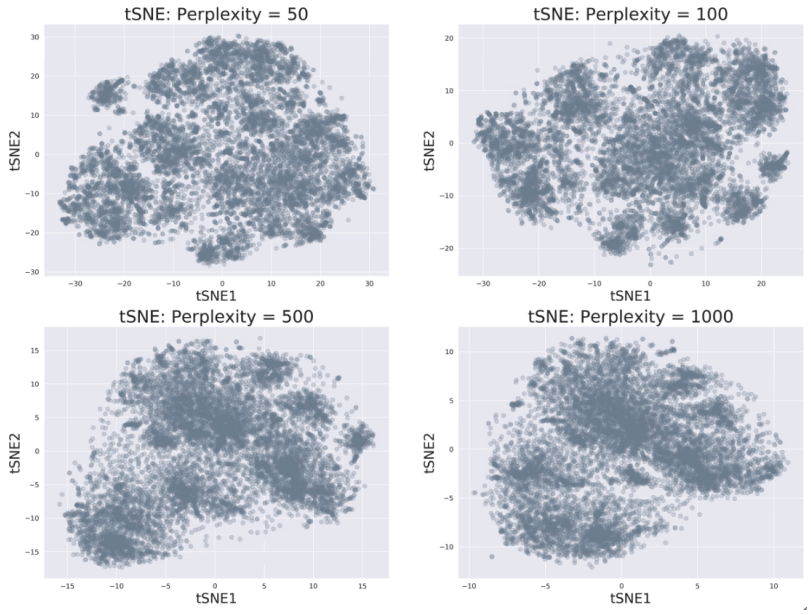

Figure 5-4: t-SNE clusters of UC patient features from baseline and follow-up period, created with varying levels of perplexity. Perplexity defines the number of neighbors to consider.

Looking at the t-SNE results labeled by age in figure 5.3 and sex in figure 5.3, we see that some patient embeddings appear close to many other patients of the same sex and age category. The plot showing initial type of conventional therapy in figure 5-7 and plot of patients with a history of adverse events in figure 5.3 show that patients who initially received corticosteroids and had a history of adverse events appear close to each other.

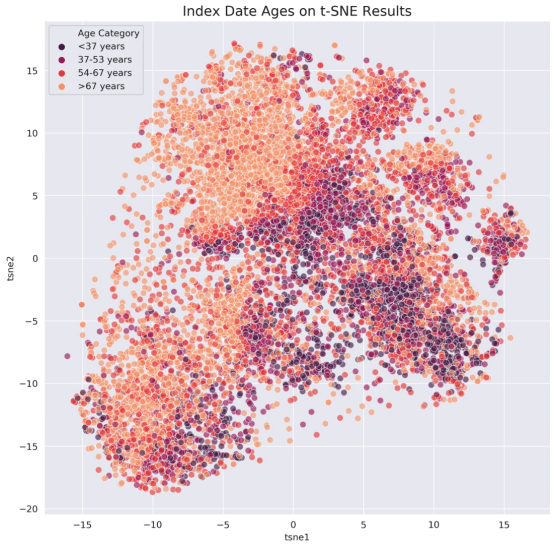

Figure 5-5: t-SNE results of UC patients labeled by patient age categories. Lighter colors represent older ages.

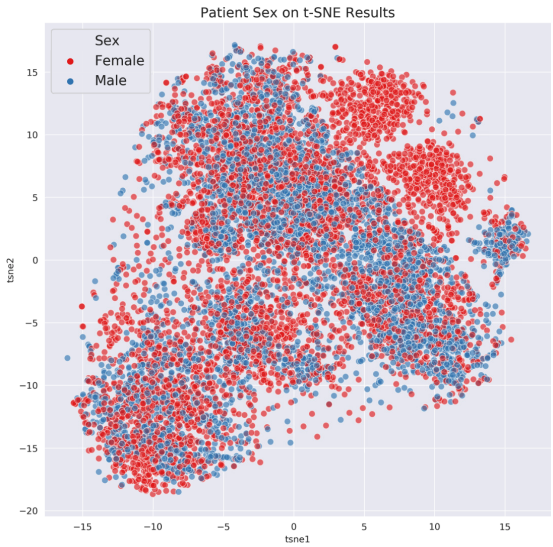

Figure 5-6: t-SNE results labeled by patient sex.

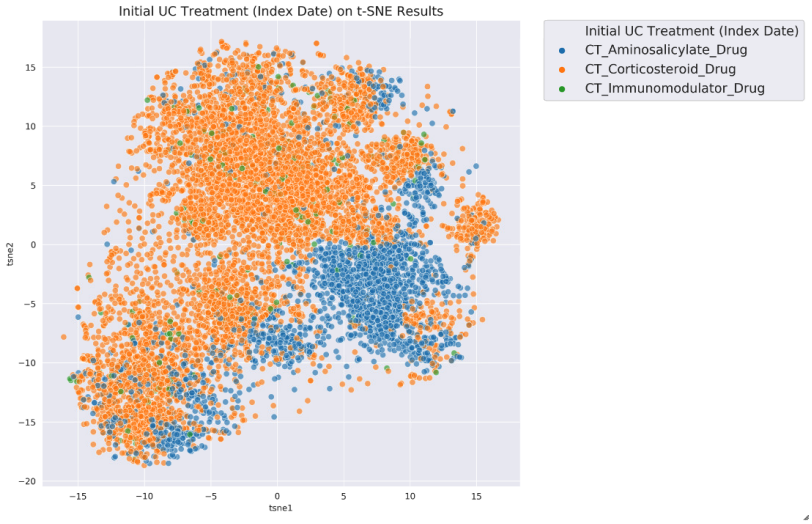

Figure 5-7: t-SNE results of UC patients labeled by patient's initial treatment type.

Figure 5-8: t-SNE results of UC patients with an adverse event occurrence prior to initiating CT.

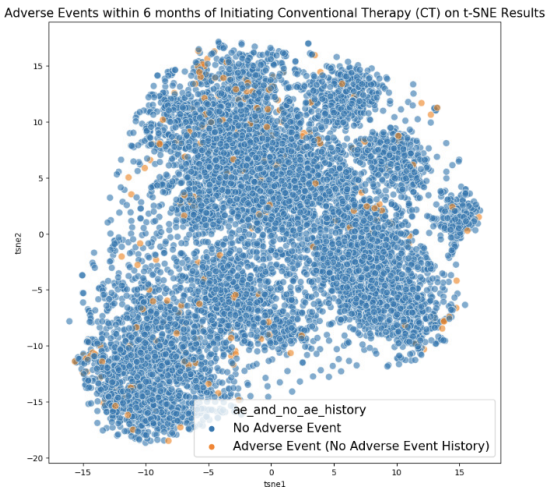

Figure 5-9: t-SNE results of UC patients labeled by adverse event occurrences and no history of adverse events prior to initiating CT.

5.4 Patient Clustering

We clustered patients and their features derived from the baseline and follow-up period using K-Means on the t-SNE output with the goal of identifying distinct patient subgroups. The resulting 4 clusters, which are shown in figure 5-10, had distinct patient characteristics.

Figure 5-10: K-Means Clusters on the t-SNE Results

Cluster 1 had the youngest patients and the highest proportion of male patients, the majority of whom were covered by private insurance and were prescribed an aminosalicylate on index date. Cluster 1 was composed of patients with an average age of 47 years at the time of initiating conventional therapy. Of the patients, 71.3% were white, 59.7% were male, and 80.5% were covered by private insurance. With 61.7% of patients exclusively prescribed an aminosalicylate on the index date, the most common initial conventional therapy was an aminosalicylate drug. Common baseline conditions included diarrhea (34.6%) and hemorrhage of the rectum and

Index Date Characteristic	Cluster 1 (n=2,726)	Cluster 2 (n=2,922)	Cluster 3 (n=3,328)	Cluster 4 (n=1,623)
Age	47 (17.2) years	61 (16.3) years	63.5 (17) years	58 (13.7) years
Female	1098 (40.3%)	1600 (54.8%)	1608 (48.3%)	1336 (82.3%)
Private Health Insurance	2195 (80.5%)	1379 (47.2%)	1396 (41.9%)	1012 (62.4%)
Race	White, 1945 (71.3%) Hispanic or Latino, 271 (9.9%) Black or African American, 194 (7.1%) Asian, 124 (4.5%) Other, 192 (7.0%)	White, 1815 (62.1%) Hispanic or Latino, 386 (13.2%) Black or African American, 264 (9.0%) Asian, 162 (5.5%) Other, 295 (10.1%)	White, 2367 (71.1%) Hispanic or Latino, 277 (8.3%) Black or African American, 311 (9.3%) Asian, 119 (3.6%) Other, 254 (7.6%)	White, 1214 (74.8%) Hispanic or Latino, 110 (6.8%) Black or African American, 142 (8.7%) Asian, 49 (3.0%) Other, 108 (6.7%)
Year	2014.4 (2.3)	2015 (2.3)	2015 (2.4)	2013.9 (2.3)
Initial CT Prescription	Corticosteroid, 786 (28.8%) Aminosalicylate, 1682 (61.7%) Immunomodulator, 22 (0.8%) Combination Therapy, 236 (8.7%)	Corticosteroid, 2225 (76.1%) Aminosalicylate, 580 (19.8%) Immunomodulator, 55 (1.9%) Combination Therapy, 62 (2.1%)	Corticosteroid, 3166 (95.1%) Aminosalicylate, 82 (2.5%) Immunomodulator, 67 (2.0%) Combination Therapy, 13 (0.4%)	Corticosteroid, 1295 (79.8%) Aminosalicylate, 257 (15.8%) Immunomodulator, 41 (2.5%) Combination Therapy, 30 (1.8%)

Table 5.5: Patient Demographics and Initial CT, by Cluster
Data are presented as mean (standard deviation) for continuous variables, or n (%) for categorical variables.

Clinical Characteristic	Cluster 1 (n=2,726)	Cluster 2 (n=2,922)	Cluster 3 (n=3,328)	Cluster 4 (n=1,623)
Comorbid Condition	1. Diarrhea (34.6%) 2. Hemorrhage of rectum and anus (33.5%) 3. Abdominal pain (24.2%) 4. Essential hypertension (20.9%) 5. Hyperlipidemia (18.3%)	1. Essential hypertension (47.4%) 2. Hyperlipidemia (41.1%) 3. Abdominal pain (24.2%) 4. Diarrhea (21.2%) 5. Pure hypercholesterolemia (19.1%)	1. Essential hypertension (52.5%) 2. Hyperlipidemia (40.1%) 3. Abdominal pain (21.2%) 4. Chest pain (19.9%) 5. Pure hypercholesterolemia (19.0%)	1. Essential hypertension (34.6%) 2. Hyperlipidemia (32.3%) 3. Abdominal pain (25.8%) 4. Diarrhea (20.9%) 5. Malaise (16.5%)
Drug Exposure	1. Suprep Bowel Prep Kit (11.9%) 2. Ciprofloxacin 500 MG Oral Tablet (11.4%) 3. Metronidazole 500 MG Oral Tablet (11.2%) 4. Midazolam Injectable Solution (9.9%) 5. Fentanyl 0.1 MG (8.8%)	1. Ciprofloxacin 500 MG Oral Tablet (13.5%) 2. {6 (Azithromycin 250 MG Oral Tablet) } Pack (12.1%) 3. Metronidazole 500 MG Oral Tablet (10.3%) 4. Omeprazole 20 MG Delayed Release Oral Capsule (8.6%) 5. Amoxicillin 500 MG Oral Capsule (8.6%)	1. Ciprofloxacin 500 MG Oral Tablet (12.4%) 2. {6 (Azithromycin 250 MG Oral Tablet) } Pack (12.2%) 3. Fentanyl 0.1 MG (9.8%) 4. Midazolam Injectable Solution (9.7%) 5. Ondansetron Injection (9.1%)	1. Ciprofloxacin 500 MG Oral Tablet (15.0%) 2. {6 (Azithromycin 250 MG Oral Tablet) } Pack (14.0%) 3. Ondansetron Injection (12.1%) 4. Midazolam Injectable Solution (11.3%) 5. Fentanyl 0.1 MG (11.0%)

Table 5.6: Baseline Conditions and Drugs by Cluster
Data are presented as the clinical characteristic (% of patients). Prevalence data were gathered from the baseline period, the 1-year period preceding CT initiation.

Clinical Characteristic	Cluster 1 (n=2,726)	Cluster 2 (n=2,922)	Cluster 3 (n=3,328)	Cluster 4 (n=1,623)
Procedure	1. Collection of venous blood by venipuncture (69.8%) 2. Level IV - Surgical pathology, gross and microscopic examination Abortion - spontaneous/missed Artery, biopsy Bone marrow, biopsy Bone exostosis Brain/meninges, other than for tumor resection Breast, biopsy, not requiring microscopic evaluation of surgica (54.3%) 3. Colonoscopy, flexible; with biopsy, single or multiple (43.6%) 4. Anesthesia for lower intestinal endoscopic procedures, endoscope introduced distal to duodenum (25.5%) 5. General examination of patient (19.6%)	1. Collection of venous blood by venipuncture (82.6%) 2. Level IV - Surgical pathology, gross and microscopic examination Abortion - spontaneous/missed Artery, biopsy Bone marrow, biopsy Bone exostosis Brain/meninges, other than for tumor resection Breast, biopsy, not requiring microscopic evaluation of surgica (43.5%) 3. Colonoscopy, flexible; with biopsy, single or multiple (24.6%) 4. Emergency department visit for the evaluation and management of a patient, which requires these 3 key components within the constraints imposed by the urgency of the patient's clinical condition and/or mental status: A comprehensive history; A comprehensi (23.3%) 5. Screening mammography, bilateral (2-view study of each breast), including computer-aided detection (CAD) when performed (22.5%)	1. Collection of venous blood by venipuncture (73.6%) 2. Level IV - Surgical pathology, gross and microscopic examination Abortion - spontaneous/missed Artery, biopsy Bone marrow, biopsy Bone exostosis Brain/meninges, other than for tumor resection Breast, biopsy, not requiring microscopic evaluation of surgica (35.5%) 3. Emergency department visit for the evaluation and management of a patient, which requires these 3 key components within the constraints imposed by the urgency of the patient's clinical condition and/or mental status: A comprehensive history; A comprehensi (27.8%) 4. Radiologic examination, chest, 2 views, frontal and lateral (22.9%) 5. Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: An expanded problem focused interval history; An expanded problem focused examination; Medical decision making of moder (21.1%)	1. Collection of venous blood by venipuncture (77.7%) 2. Screening mammography, bilateral (2-view study of each breast), including computer-aided detection (CAD) when performed (51.4%) 3. Computer-aided detection (computer algorithm analysis of digital image data for lesion detection) with further review for interpretation, with or without digitization of film radiographic images; screening mammography (List separately in addition to code (48.3%) 4. Screening mammography (44.4%) 5. Level IV - Surgical pathology, gross and microscopic examination Abortion - spontaneous/missed Artery, biopsy Bone marrow, biopsy Bone exostosis Brain/meninges, other than for tumor resection Breast, biopsy, not requiring microscopic evaluation of surgica (43.6%)

Table 5.7: Baseline Procedures by Cluster
Data are presented as the clinical characteristic (% of patients). Prevalence data were gathered from the baseline period, the 1-year period preceding CT initiation.

anus (33.5%). Common baseline procedures included blood draw (69.8%), biopsy (54.3%), and colonoscopy (43.6%).

Cluster 2 had the most racially diverse composition of patients, about half of whom were covered by private health insurance. Patients had an average age of 61 years at the time of initiating conventional therapy. Of the patients, 62.1% were white, 54.8% were female, and 52.8% were not covered by private insurance. With 76.1% of patients exclusively prescribed a corticosteroid on the index date, the most common initial conventional therapy was a corticosteroid drug. Common baseline conditions included essential hypertension (47.4%) and hyperlipidemia (41.1%). Common baseline procedures included blood draw (82.6%), biopsy (43.5%), and colonoscopy (24.6%).

Cluster 3 had the oldest group of patients, the least private insurance coverage, and the highest rate of initial CT prescriptions for corticosteroid drugs, with just over 95% of patients prescribed a corticosteroid on index date. The average age was 63.5 years at the time of initiating conventional therapy. Of the patients, 71.1% were white, 51.7% were male, and 58.1% were not covered by private insurance. With 95.1% of patients exclusively prescribed a corticosteroid on the index date, the most common initial conventional therapy was a corticosteroid drug. Common baseline conditions included essential hypertension (52.5%) and hyperlipidemia (40.1%). Common baseline procedures included blood draw (73.6%), biopsy (35.5%), and chest radiological exam (22.9%).

Cluster 4 had the highest proportion of female patients (82.3%), as well as the highest proportion of white patients (74.8%), majority of whom were covered by private insurance. Patients had an average age of 58 years at the time of initiating conventional therapy. Of the patients, 74.8% were white, and 62.4% were covered by private insurance. With 79.8% of patients exclusively prescribed a corticosteroid on the index date, the most common initial conventional therapy was a corticosteroid drug. Common baseline conditions included essential hypertension (34.6%) and hyperlipidemia (32.3%). Common baseline procedures included blood draw (77.7%), bilateral screening mammography (51.4%), and biopsy (43.6%).

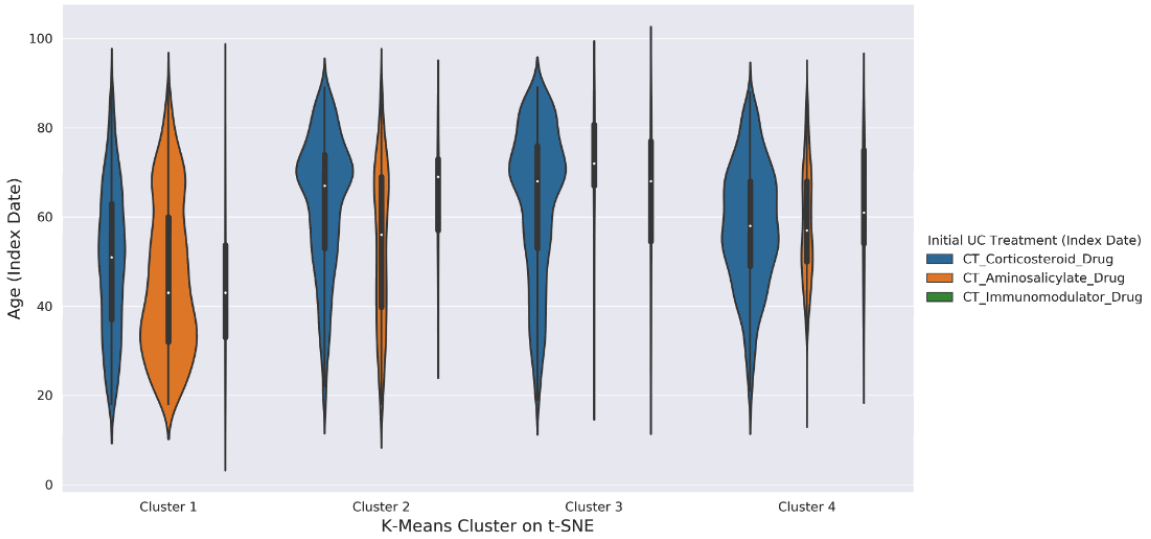

Figure 5-11: Comparison of patients by cluster assignment, age, and initial UC treatment type. Size corresponds to the number of individuals in each group.

Patient age and initial type of conventional therapy prescribed differed across the clusters, as shown in the violin plot in Figure 5-11. Cluster 1 contained younger patients initially prescribed aminosalicylates, whereas high proportions of older patients in clusters 2 and 3 were initially prescribed corticosteroids.

5.4.1 Other Clustering Analyses Performed

In addition to clustering patients using the full patient feature matrix across the baseline and follow-up period in 6-month windows, we considered clustering patients using features from just the follow-up period. We examined clusters produced when only considering lab results (low, normal, high categorical values) for patients across the baseline and follow-up period and just the follow-up period but did not find distinct clusters. We ran separate analyses considering the top 100 labs across the period of interest, the labs which at least 5% of patients had measurements, and the labs for which at least 10% of patients had measurements recorded. When only using the lab data, clusters were not visible and thus additional data sources (e.g., procedures, conditions) were used for the final analysis.

Chapter 6

Discussion and Conclusion

In this thesis, we outlined a cohort generation tool to process data in the OMOP CDM format, presented a patient timeline visualization tool compatible with datasets in the OMOP CDM format, analyzed a cohort of newly diagnosed UC patients with a focus on the initial type of conventional therapy, and demonstrated the potential to use clustering methods on claims data to uncover distinct patient subgroups.

To compare patient outcomes by initial prescription type, we investigated the relationship between initial type of conventional therapy and the occurrence of an adverse event shortly after a patient newly diagnosed with Ulcerative colitis initiated conventional therapy. As observed in related work, the most common adverse events among Ulcerative colitis patients within 6 months of initiating treatment (see Table 5.3) included Congestive heart failure, Type 2 diabetes mellitus, and Osteoporosis [43].

Our results suggest that sparse features from data sources like claims databases can be used to discover underlying patient subgroups and associations between patient treatment plans and outcomes. In particular, Cluster 1 (blue cluster in figure 5-10) had the youngest patients and the highest proportion of male patients, the majority of whom were covered by private insurance and were prescribed an aminosalicylate on index date. Interestingly, Cluster 1 had the highest rates of colonoscopy procedures prior to initiating conventional therapy. Cluster 2 (orange cluster in figure 5-10) had the most racially diverse composition of patients, about half of whom were covered by

private health insurance. Cluster 3 (upper left, green cluster in figure 5-10) contained older patients initially treated with corticosteroids, the majority of whom were not covered by private insurance. Cluster 3 also contained many patients with a history of adverse events of interest, which is interesting since adverse event labels were not explicitly included in the patient-level features used as input, but this could also just be due to the patients being older. Cluster 4 (upper right, red cluster in figure 5-10) contained mostly female patients.

In the patients analyzed, corticosteroids were the most commonly prescribed initial type of conventional therapy, followed by aminosalicylates, combination therapy (at least two prescriptions for a corticosteroid, aminosalicylate, and/or immunomodulator), and immunomodulators. While longitudinal adult studies have demonstrated that aminosalicylates were the most common conventional therapy drug type[22], the high rates of corticosteroid prescriptions observed in our analysis may be explained by the fact that we only reported prescriptions for a specific point in time, which was the time a patient initiated conventional therapy for Ulcerative colitis; under these conditions, a shorter prescription to treat a flare-up using a drug such as a corticosteroid may be merited.

In our results, not only were corticosteroids commonly prescribed, but the patients were on average older than the average age of a patient newly diagnosed with Ulcerative colitis [33]. Since Ulcerative colitis is a chronic condition and corticosteroids are commonly used to treat flare ups, one possible reasons for patients initially receiving corticosteroids might be that these patients are not truly newly diagnosed UC patients; older patients may have an unknown history of UC due to the diagnoses occurring before the earliest date included in the dataset.

6.1 Limitations and Future Work

Using claims data to identify patient subgroups has many limitations since we are using the data for a purpose other than intended, as claims data is used for the purposes of billing. Specifically, limitations include: the reliance of claims data accuracy, re-

quirement of the patient to be insured to be represented in the dataset, requirement of continuous insurance plan enrollment to obtain continuous data, and limited patient demographic and mortality information. Since claims data is derived from insurance claims, we collaborated with pharmaceutical experts to carefully define the cohort definition and ensure our cohort contains Ulcerative Colitis patients as intended [14].

While we included patient demographics, conditions, lab results, procedures, and drugs in our analysis, future work can be done to include additional patient information such as geographic location, family history, and cost data. In our analysis, we only included lab tests containing results and ignored labs that were performed but missing results, which may introduce a strong bias into the data. It would be useful to perform these analyses incorporating data on whether or not a lab was performed, regardless of the lab result. Our findings may not generalize to populations outside of the U.S., since the data used in our analysis is restricted to data for care sites within the United States. Use of more comprehensive data and more patients may produce different results. As we only consider claims data, we do not consider any services or medicines that are not prescribed or offered by healthcare providers and subsequently entered into the billing system. Additionally, we rely on billing codes from multiple care sites to identify and label patients under the assumption that each healthcare provider logs information consistent with other providers; we do not capture information on how or why healthcare providers bill.

With respect to the patient features used, there are many other ways of representing a patient's longitudinal health data to explore. For example, we could include all labs that were performed instead of only the labs that contain measurements. Instead of using one window length, we could consider multiple window lengths and see how the results compare. It would be interesting to see whether patients are classified in the same cluster when varying window lengths, and if so to compare these patients to patients who are classified in different clusters when varying the window length.

It would be interesting to evaluate the robustness of our clustering strategy by performing it on an external cohort of patients newly diagnoses with UC and examine its utility in identifying clinical distinct subsets of patients. It would be interesting to

see if patients with similar demographics such as age, sex, and insurance type cluster together when some or all demographic features are omitted. To better understand characteristics of patients initially prescribed each type of conventional therapy and investigate potentially distinct characteristics across these patient groups, it would be interesting to examine patient demographics and prevalent events for patients based on the type of initial CT prescribed, similar to what was done to compare the patient composition of the different clusters with Table 5.5 and Table 5.6.

This was a retrospective study and not a randomized control trial, which makes it difficult to attribute causal links between sets of observations and outcomes. Additionally, a medication a patient was previously taking might interact with an existing medication, making it difficult to attribute a causal link between a treatment pattern and an adverse event. We assume that medications were taken as prescribed, so medication non-adherence is not accounted for in our analysis. Additionally, since there are very few patients with adverse events who receive initial treatment other than corticosteroids, future analyses must be done to examine the role of initial treatment and adverse events by accounting for class imbalance and using causal inference techniques like propensity scoring.

Appendix A

Tables

Clinical Characteristic	Cluster 1 (n=2,726)	Cluster 2 (n=2,922)	Cluster 3 (n=3,328)	Cluster 4 (n=1,623)
Lab Measurement	1. Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (58.5%) 2. Comprehensive metabolic panel This panel must include the following: Albumin, Bilirubin, Calcium, Carbon dioxide (bicarbonate), Chloride, Creatinine, Glucose, Phosphatase, alkaline Potassium (54.8%) 3. Lipid panel This panel must include the following: Cholesterol, serum, total Lipoprotein, direct measurement, high density cholesterol (HDL cholesterol),Triglycerides (42.6%) 4. Thyroid stimulating hormone (TSH) (28.4%) 5. Potassium [Moles/volume] in Serum or Plasma (23.1%)	1. Comprehensive metabolic panel This panel must include the following: Albumin, Bilirubin, Calcium, Carbon dioxide (bicarbonate), Chloride, Creatinine, Glucose, Phosphatase, alkaline Potassium (77.8%) 2. Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (73.6%) 3. Lipid panel This panel must include the following: Cholesterol, serum, total Lipoprotein, direct measurement, high density cholesterol (HDL cholesterol),Triglycerides (71.2%) 4. Creatinine [Mass/volume] in Serum or Plasma (68.5%) 5. Potassium [Moles/volume] in Serum or Plasma (68.2%)	1. Comprehensive metabolic panel This panel must include the following: Albumin, Bilirubin, Calcium, Carbon dioxide (bicarbonate), Chloride, Creatinine, Glucose, Phosphatase, alkaline Potassium (61.0%) 2. Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (59.5%) 3. Lipid panel This panel must include the following: Cholesterol, serum, total Lipoprotein, direct measurement, high density cholesterol (HDL cholesterol),Triglycerides (54.7%) 4. Thyroid stimulating hormone (TSH) (41.2%) 5. Creatinine [Mass/volume] in Serum or Plasma (23.6%)	1. Blood count; complete (CBC), automated (Hgb, Hct, RBC, WBC and platelet count) and automated differential WBC count (60.5%) 2. Comprehensive metabolic panel This panel must include the following: Albumin, Bilirubin, Calcium, Carbon dioxide (bicarbonate), Chloride, Creatinine, Glucose, Phosphatase, alkaline Potassium (60.4%) 3.Lipid panel This panel must include the following: Cholesterol, serum, total Lipoprotein, direct measurement, high density cholesterol (HDL cholesterol),Triglycerides (58.5%) 4. Thyroid stimulating hormone (TSH) (42.9%) 5. Creatinine [Mass/volume] in Serum or Plasma (18.7%)

Table A.1: Baseline Lab Measurements, by Cluster
Data are presented as the clinical characteristic (% of patients). Prevalence data are gathered from the baseline period, the 1-year period preceding CT initiation.

Appendix B

Figures

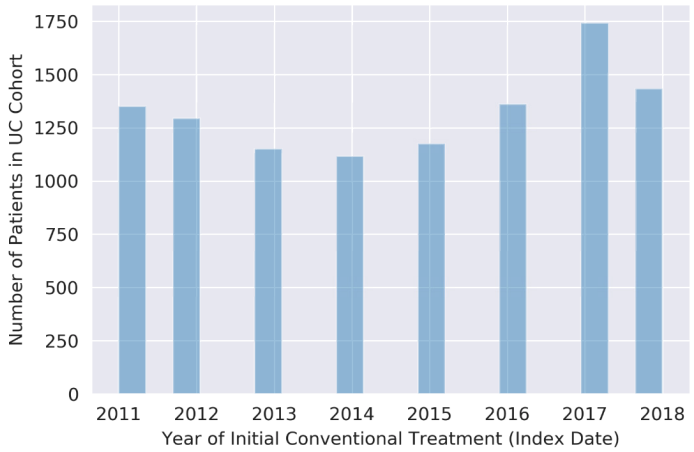

Figure B-1: Patient counts by year of initial treatment (index date) in the cohort.

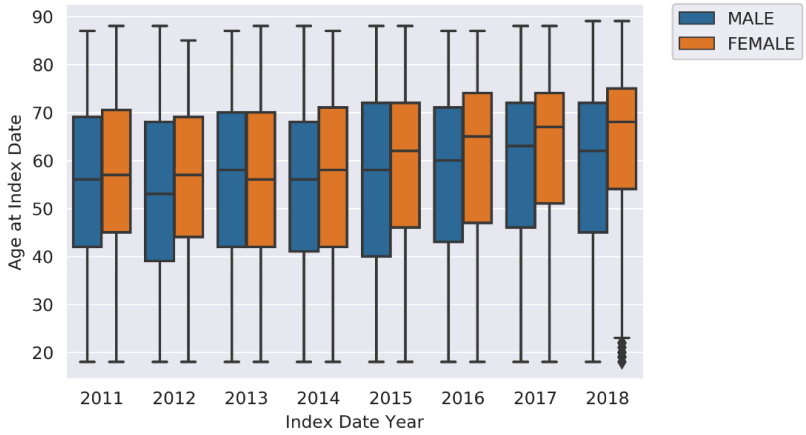

Figure B-2: Boxplot of patient age and sex across the range of patient index dates in the cohort.

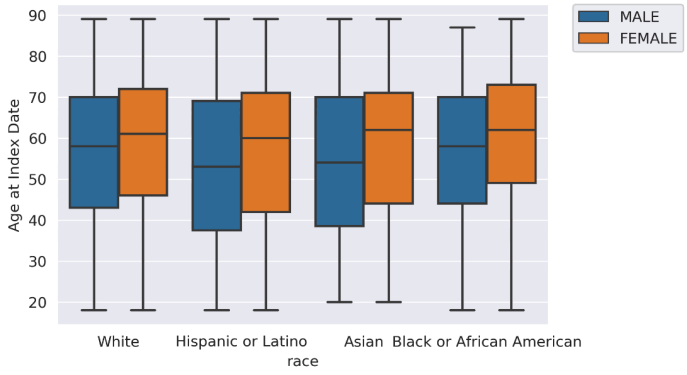

Figure B-3: Boxplot of patient age and sex for each race/ethnicity represented in the cohort.

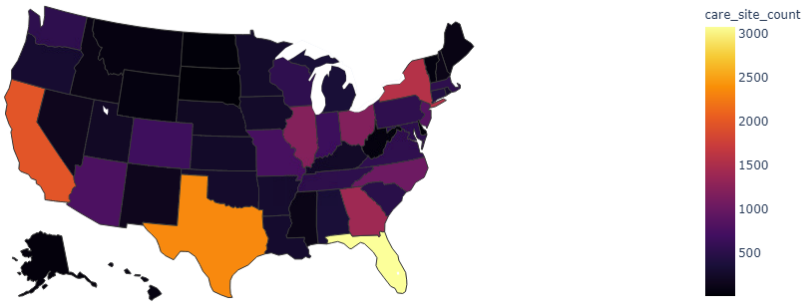

Figure B-4: Map highlighting states by total number of care site visits by patients in the cohort during the observation period.

Bibliography

[1] Hany Alashwal, Mohamed El Halaby, Jacob J Crouse, Areeg Abdalla, and Ahmed A Moustafa. The application of unsupervised clustering methods to alzheimer's disease. *Front. Comput. Neurosci.*, 13:31, May 2019.

[2] Brigida Barberio, Jonathan P Segal, M Nabil Quraishi, Christopher J Black, Edoardo Vincenzo Savarino, and Alexander C Ford. Efficacy of oral, topical, or combined oral and topical 5-aminosalicylates, in ulcerative colitis: Systematic review and network Meta-Analysis.

[3] Brett K Beaulieu-Jones, Patryk Orzechowski, and Jason H Moore. Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database. In *Biocomputing 2018*, pages 123–132. WORLD SCIENTIFIC, October 2017.

[4] Nick Bowman, Dong Liu, Patrick Paczkowski, Jon Chen, John Rossi, Sean Mackay, Adrian Bot, and Jing Zhou. Advanced cell mapping visualizations for single cell functional proteomics enabling patient stratification. *Proteomics*, 20(13):e1900270, July 2020.

[5] Robert H Dolin, Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M Behlen, Paul V Biron, and Amnon Shabo Shvo. HL7 clinical document architecture, release 2. *J. Am. Med. Inform. Assoc.*, 13(1):30–39, January 2006.

[6] Hossein Estiri, Jeffrey G Klann, and Shawn N Murphy. A clustering approach for detecting implausible observation values in electronic health records data. *BMC Med. Inform. Decis. Mak.*, 19(1):142, July 2019.

[7] Alexander C Ford, Khurram J Khan, Jean-Paul Achkar, and Paul Moayyedi. Efficacy of oral vs. topical, or combined oral and topical 5-aminosalicylates, in ulcerative colitis: Systematic review and Meta-Analysis, 2012.

[8] Jocelyn Gal, Caroline Bailleux, David Chardin, Thierry Pourcher, Julia Gilhodes, Lun Jing, Jean-Marie Guigonis, Jean-Marc Ferrero, Gerard Milano, Baharia Mograbi, Patrick Brest, Yann Chateau, Olivier Humbert, and Emmanuel Chamorey. Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. *Comput. Struct. Biotechnol. J.*, 18:1509–1524, June 2020.

[9] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, Johan van der Lei, Nicole Pratt, G Niklas Norén, Yu-Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.*, 216:574–578, 2015.

[10] Hyerim Ji, Seok Kim, Soyoung Yi, Hee Hwang, Jeong-Whun Kim, and Sooyoung Yoo. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. *J. Biomed. Inform.*, 107:103459, July 2020.

[11] Dipak Kalra, Thomas Beale, and Sam Heard. The openEHR foundation. *Stud. Health Technol. Inform.*, 115:153–173, 2005.

[12] Rohan S Kodialam, Rebecca Boiarsky, Justin Lim, Neil Dixit, Aditya Sai, and David Sontag. Deep contextual clinical prediction with reverse distillation. *AAAI*, July 2020.

[13] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in K-Means clustering. *Aquat. Microb. Ecol.*, 1(6):90–95, 2013.

[14] Isaac S Kohane, Bruce J Aronow, Paul Avillach, Brett K Beaulieu-Jones, Riccardo Bellazzi, Robert L Bradford, Gabriel A Brat, Mario Cannataro, James J Cimino, Noelia García-Barrio, Nils Gehlenborg, Marzyeh Ghassemi, Alba Gutiérrez-Sacristán, David A Hanauer, John H Holmes, Chuan Hong, Jeffrey G Klann, Ne Hooi Will Loh, Yuan Luo, Kenneth D Mandl, Daniar Mohamad, Jason H Moore, Shawn N Murphy, Antoine Neuraz, Kee Yuan Ngiam, Gilbert S Omenn, Nathan Palmer, Lav P Patel, Miguel Pedrera-Jiménez, Piotr Sliz, Andrew M South, Amelia Li Min Tan, Deanne M Taylor, Bradley W Taylor, Carlo Torti, Andrew K Vallejos, Kavishwar B Wagholikar, Griffin M Weber, and Tianxi Cai. What every reader should know about studies using electronic health record data but may be afraid to ask. *J. Med. Internet Res.*, January 2021.

[15] Asher Kornbluth, David B Sachar, and Practice Parameters Committee of the American College of Gastroenterology. Ulcerative colitis practice guidelines in adults: American college of gastroenterology, practice parameters committee. *Am. J. Gastroenterol.*, 105(3):501–23; quiz 524, March 2010.

[16] Bum Chul Kwon, Ben Eysenbach, Janu Verma, Kenney Ng, Christopher De Filippi, Walter F Stewart, and Adam Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE Trans. Vis. Comput. Graph.*, 24(1):142–151, January 2018.

[17] Peter Laszlo Lakatos. Prediction of disease course in inflammatory bowel diseases. *World J. Gastroenterol.*, 16(21):2589–2590, June 2010.

[18] Bei Li and Rich Tsui. How to improve the reuse of clinical data– openEHR and OMOP CDM. *J. Phys. Conf. Ser.*, 1624(3):032041, October 2020.

[19] Daniel M Lima, Jose F Rodrigues-Jr, Agma J M Traina, Fabio A Pires, and Marco A Gutierrez. Transforming two decades of ePR data to OMOP CDM for clinical research. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, volume 264, pages 233–237. IOS Press, 2019.

[20] Carolina Lucas, Patrick Wong, Jon Klein, Tiago B R Castro, Julio Silva, Maria Sundaram, Mallory K Ellingson, Tianyang Mao, Ji Eun Oh, Benjamin Israelow, Takehiro Takahashi, Maria Tokuyama, Peiwen Lu, Arvind Venkataraman, Annsea Park, Subhasis Mohanty, Haowei Wang, Anne L Wyllie, Chantal B F Vogels, Rebecca Earnest, Sarah Lapidus, Isabel M Ott, Adam J Moore, M Catherine Muenker, John B Fournier, Melissa Campbell, Camila D Odio, Arnau Casanovas-Massana, Yale IMPACT Team, Roy Herbst, Albert C Shaw, Ruslan Medzhitov, Wade L Schulz, Nathan D Grubaugh, Charles Dela Cruz, Shelli Farhadian, Albert I Ko, Saad B Omer, and Akiko Iwasaki. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature*, 584(7821):463–469, August 2020.

[21] Qinli Ma, Michael Mack, Sonali Shambhu, Kathleen McTigue, and Kevin Haynes. Characterization of bariatric surgery and outcomes using administrative claims data in the research network of a nationwide commercial health plan. *BMC Health Serv. Res.*, 21(1):116, February 2021.

[22] Fernando Magro, Andreia Rodrigues, Ana Isabel Vieira, Francisco Portela, Isabelle Cremers, José Cotter, Luis Correia, Maria Antónia Duarte, Maria Lourdes Tavares, Paula Lago, Paula Ministro, Paula Peixe, Susana Lopes, and Elizabeth Benito Garcia. Review of the disease course among adult ulcerative colitis population-based longitudinal cohorts. *Inflamm. Bowel Dis.*, 18(3):573–583, March 2012.

[23] Michal Markovich Gordon, Asher M Moser, and Eitan Rubin. Unsupervised analysis of classical biomedical markers: robustness and medical relevance of patient clustering using bioinformatics tools. *PLoS One*, 7(3):e29578, March 2012.

[24] Amy Matcho, Patrick Ryan, Daniel Fife, and Christian Reich. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. *Drug Saf.*, 37(11):945–959, November 2014.

[25] Megan E McAuliffe, Stephan Lanes, Timothy Leach, Asit Parikh, Gerald Faich, Jane Porter, Crystal Holick, Daina Esposito, Yueqin Zhao, and Irving Fox. Occurrence of adverse events among patients with inflammatory bowel disease in the HealthCore integrated research database. *Curr. Med. Res. Opin.*, 31(9):1655–1664, August 2015.

[26] Johannes Meier and Andreas Sturm. Current treatment of ulcerative colitis. *World J. Gastroenterol.*, 17(27):3204–3212, July 2011.

[27] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.*, 6:26094, May 2016.

[28] Debanjali Mitra, Paul Hodgkins, Linnette Yen, Keith L Davis, and Russell D Cohen. Association between oral 5-ASA adherence and health care utilization and costs among patients with active ulcerative colitis. *BMC Gastroenterol.*, 12:132, September 2012.

[29] Observational Health Data Sciences and Informatics. Chapter 4 the common data model, January 2021. Accessed: 2021-3-21.

[30] Observational Health Data Sciences and Informatics. Chapter 5 standardized vocabularies, January 2021. Accessed: 2021-3-21.

[31] U.S. National Library of Medicine. Health data sources. *U.S. National Library of Medicine*, February 2019.

[32] J Marc Overhage, Patrick B Ryan, Christian G Reich, Abraham G Hartzema, and Paul E Stang. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.*, 19(1):54–60, January 2012.

[33] Sandra M Quezada and Raymond K Cross. Association of age at diagnosis and ulcerative colitis phenotype. *Dig. Dis. Sci.*, 57(9):2402–2407, September 2012.

[34] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-Level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, December 2015.

[35] Matthew W Segar, Kershaw V Patel, Colby Ayers, Mujeeb Basit, W H Wilson Tang, Duwayne Willett, Jarett Berry, Justin L Grodin, and Ambarish Pandey. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. *Eur. J. Heart Fail.*, 22(1):148–158, January 2020.

[36] M A Syakur, B K Khotimah, E M S Rochman, and B D Satoto. Integration K-Means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conf. Ser.: Mater. Sci. Eng.*, 336(1):012017, April 2018.

[37] Ming Tang, Chao Gao, Stephen A Goutman, Alexandr Kalinin, Bhramar Mukherjee, Yuanfang Guan, and Ivo D Dinov. Model-Based and Model-Free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering. *Neuroinformatics*, 17(3):407–421, July 2019.

[38] Kartikeya Tripathi and Joseph D Feuerstein. New developments in ulcerative colitis: latest evidence on management, treatment, and maintenance. *Drugs Context*, 8:212572, April 2019.

[39] Ryan Ungaro, Saurabh Mehandru, Patrick B Allen, Laurent Peyrin-Biroulet, and Jean-Frédéric Colombel. Ulcerative colitis. *Lancet*, 389(10080):1756–1770, April 2017.

[40] Laurens van der Maaten. Visualizing data using t-SNE, 2008. Accessed: 2021-3-25.

[41] Yanshan Wang, Yiqing Zhao, Terry M Therneau, Elizabeth J Atkinson, Ahmad P Tafti, Nan Zhang, Shreyasee Amin, Andrew H Limper, Sundeep Khosla, and Hongfang Liu. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *J. Biomed. Inform.*, 102:103364, February 2020.

[42] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

[43] Thomas Wilke, Antje Groth, Gráinne H Long, Amanda R Tatro, and Diana Sun. Rate of adverse events and associated health care costs for the management of inflammatory bowel disease in germany. *Clin. Ther.*, 42(1):130–143.e3, January 2020.

[44] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics Intellig. Lab. Syst.*, 2(1):37–52, August 1987.

[45] Seng Chan You, Seongwon Lee, Soo-Yeon Cho, Hojun Park, Sungjae Jung, Jaehyeong Cho, Dukyong Yoon, and Rae Woong Park. Conversion of national health insurance Service-National sample cohort (NHIS-NSC) database into observational medical outcomes Partnership-Common data model (OMOP-CDM). *Stud. Health Technol. Inform.*, 245:467–470, 2017.

[46] Xiantong Zou, Xianghai Zhou, Zhanxing Zhu, and Linong Ji. Novel subgroups of patients with adult-onset diabetes in chinese and US populations. *Lancet Diabetes Endocrinol*, 7(1):9–11, January 2019.